Math Tools: Stochastic Processes

Revised: November 4, 2014

All of modern macroeconomics, and most of modern finance, is concerned with how randomness unfolds through time. Such combinations of randomness and dynamics go by the name of *stochastic processes*. Thus we might say, as Lars Hansen and Tom Sargent often do, that a macroeconomic model — or an asset pricing model — is a stochastic process.

Our goal here is to review the theory of stochastic processes in a relatively low-tech way. We start where we left off, with an event tree, which we extend to more than two periods. This isn't difficult, but we need new notation to track how uncertainty unfolds with time.

We then proceed to impose more structure. Typically we'd like uncertainty to have some regularity to it, so that behavior at one date is similar to the behavior at another. Without something like this, it's hard to know how we'd compare models with data. Think, for example, about estimating the mean from time series data. If every period has its own distribution, what does this even mean? We focus on linear normal processes, where the analysis if relatively transparent, but the same ideas extend to other environments.

There are some optional sections that use linear algebra. Keep in mind: they're optional. But also keep in mind: linear algebra is one of the great things in life, something you should think about learning if you want to have more fun in life. The optional applications to Markov chains are a good example.

1 Event trees

We can expand the concept of an event tree to more than the two dates we've used so far. An example is Figure 1, but we can imagine extending it further.

An event tree starts with a single event that we'll label z_0: the event z that occurred at date $t = 0$. At date $t = 1$ another event z_1 occurs. In the figure, there are two possible events, $z_1 = 1$ and $z_1 = 2$, each represented by a branch stemming from the initial event z_0. The probability of each branch, conditional on starting at z_0, is $\text{Prob}(z_1|z_0)$. These probabilities are all greater than or equal to zero, and the sum across events z_1 is one.

We can extend the idea to as many periods as we like, with the node at any date t giving rise to further branches at $t + 1$. It's a little cumbersome, but nonetheless helpful, to have notation for the various nodes we generate this way. We refer to the *history* of events through date t as the sequence $(z_0, z_1, \ldots, z_t)$, which we denote by z^t. Each such history defines a specific node in the tree. For example, the node labeled (A) in the figure comes from the history $z^1 = (z_0, z_1 = 1)$. The node in the upper right labeled (C) comes from the history $z^2 = (z_0, z_1 = 1, z_2 = 1)$. We're defining nodes by listing every step we followed to get there, which is sufficient but seems like overkill.

We construct probabilities of nodes by stringing together probabilities of the branches that give rise to them. Suppose for example, that the probability of taking branch $z_1 = 1$ from z_0 is $\text{Prob}(z_1 = 1|z_0)$, and the probability of continuing with $z_2 = 1$ is $\text{Prob}[z_2 = 1|z^1 = (z_0, z_1 = 1)]$, then the probability at z_0 of getting to the upper right node is the product:

$$\text{Prob}[z_2 = 1|z^1 = (z_0, z_1 = 1)] \quad = \quad \text{Prob}(z_1 = 1|z_0)\ \text{Prob}[z_2 = 1|z^1 = (z_0, z_1 = 1)].$$

In words: the probability of getting from z_0 to (C) is the probability of getting from z_0 to (A) times the probability of getting from (A) to (C).

Given a tree, we can then specify things like cash flows at each node and compute their prices at earlier nodes. Prices and cash flows are random here because the nodes are random: we don't know which ones will occur. It's analogous to what we did earlier with random variables. In our two-period settings, cash flows (and other things) were random because they were functions of the state z. Here the same idea applies but with state z replaced by history z^t. In this respect, the appropriate definition of a "state" at date t is the complete history of events up to then, namely z^t.

All of this works reasonably well, but it's way too cumbersome. And without some regularity in how uncertainty works at different dates, it's hard to see how we'd be able to connect properties of the tree to analogous properties of data. What kind of tree, for example, corresponds to the behavior of the short-term interest rate? Or the growth rate of consumption?

2 Well-behaved stochastic processes

We will use stochastic processes with a lot more structure on them than this. Why? Lots of reasons, but mainly because they're simpler, a give us a clear link between models and evidence.

One kind of structure is what is called the *Markov property*: the probabilities of possible outcomes next period depend on this period's state, but not the complete history. Formally, we would write $\text{Prob}(z_{t+1}|z^t) = \text{Prob}(z_{t+1}|z_t)$. as a result, the event z_t tells us everything we need to know about the situation at date t. We will refer to it as the state for that reason. In a sense, we're limiting the memory to one period, but there's a standard trick to extend it without going back to the complete history. If we want a two-period memory, for example, we might define the state by $y_t = (z_t, z_{t-1})$ and apply the definition to y_t rather than z_t. This works for any finite history, although it undercuts some of the simplification we were looking for.

Another kind of structure is to use the same conditional probabilities $\text{Prob}(z_{t+1}|z_t)$ at all dates. We would refer to such a process as *stationary* (it stays the same) or *time-homogeneous* (the same at all times). The date, for example, is irrelevant. It doesn't mean that the distribution over future states is always the same. It does mean that any variation in the distribution works through the current state z_t.

The Markov property refers to one-period conditional distributions $\text{Prob}(z_{t+1}|z_t)$, but we can extend these one-period distributions to multiple periods. The logic is similar to what

we did with event trees: By stringing together one-period conditional distributions we can construct multi-period conditional distributions $\text{Prob}(z_{t+k}|z_t)$ for any $k \geq 1$. We'll save the details until we have some specific candidates to work with, but we can probably imagine doing something of this sort.

What happens as we extend k to longer time horizons? The structure we've imposed so far allows a lot of different kinds of behavior, but in all the examples we'll use, the conditional distribution will "settle down." By that we mean that the distribution $\text{Prob}(z_{t+k} = z|z_t)$ converges as we increase k to a unique distribution $\text{Prob}(z)$ that doesn't depend on the current state z_t. If it does, we say the process is *stable*. There are two conditions here: that it converges, and that it converges to the same thing for every z_t. Terminology varies, but we will refer to the resulting $\text{Prob}(z)$ as the *equilibrium distribution*.

Despite the brevity of this section, it has a lot of content. We'll see how it all works when we apply it.

3 Digression: matrix multiplication (skip)

Here's a quick summary of some linear algebra we'll find helpful. A matrix is a table of numbers. The matrix A for example has (say) m rows and n columns, so we would say it is "m by n." The element in the ith row and j column is usually written a_{ij}. A (column) vector is a matrix with one column ($n = 1$).

If the dimensions line up, we can multiply one matrix times another. Suppose we multiply A times B and call the result C (a new matrix). The rule for doing this is

$$c_{ij} = \sum_k a_{ik} b_{kj},$$

where c_{ij} is the ijth element of C. We apply the rule for all possible i and j. This works only if the second dimension of A equals the first dimension of B: if, for example, A is m by n and B is n by p. Then the subscript k in the sum runs from 1 to n.

Matlab is set up to work directly with vectors and matrices. When we define a matrix in Matlab, we separate rows with semi-colons. The command `A = [1 2 5; 6 4 2]`, for example, defines the 2 (rows) by 3 (columns) matrix

$$A = \begin{bmatrix} 1 & 2 & 5 \\ 6 & 4 & 2 \end{bmatrix}.$$

Similarly, `B = [1 4; 3 2]` defines the matrix

$$B = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}.$$

You might verify by hand that

$$BA = \begin{bmatrix} 25 & 18 & 13 \\ 15 & 14 & 19 \end{bmatrix}.$$

You can do the same calculation in Matlab with the command `C = B*A`.

As a test of your knowledge, what is AB? What happens if you type $D = A*B$ in Matlab?

4 Markov chains (skip)

Markov chains are wonderful examples of Markov processes for illustrating the concepts of stationarity and stability. Both are relatively simple if you're comfortable with linear algebra and matrix multiplication.

Here's the structure. The state z_t takes on a finite set of values, the same at all dates, which we label with the integers i and j. The probability of state j next period depends only on today's state i (the Markov property) and is the the same at all dates (stationary):

$$p_{ij} \quad = \quad \text{Prob}(z_{t+1} = j | z_t = i).$$

We can collect them in a matrix P, called the *transition matrix*, where p_{ij} is the element in the ith row and jth column. Thus row i gives us the probabilities of moving to any state j next period from state i now.

Here's an example. Suppose

$$P \quad = \quad \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}.$$

Row i is the probability distribution over states next period if the current state is i. If we're in state 1 now, there's an 80% chance we'll stay there and a 20% chance we'll move to state 2. These probabilities are nonnegative and sum to one. Ditto for row 2. Put another way: any matrix P is a legitimate transition matrix if (i) $p_{ij} \geq 0$ and (ii) $\sum_j p_{ij} = 1$. By choosing different numbers, we can generate a wide range of behavior.

The matrix P describes probabilities of one-period changes in the state. But what are the probabilities of moving from state 1 to state 2 in 2 periods? In 10 periods? As it happens, the probability of moving from state i at date t to state j at date $t + k$ is the ijth element of P^k. So we can compute the whole set of probabilities by computing powers of P. In our example, we get

$$P^2 \quad = \quad \begin{bmatrix} 0.70 & 0.30 \\ 0.45 & 0.55 \end{bmatrix}, \quad P^3 \quad = \quad \begin{bmatrix} 0.650 & 0.350 \\ 0.525 & 0.475 \end{bmatrix}.$$

Try it yourself in Matlab.

We're typically interested in examples in which the probabilities are stable. If we increase the time horizon k, what happens to the probabilities? With the right conditions, P^k converges to a matrix whose rows are all the same. This says that the probability distribution across states doesn't depend on the current state if we get far enough in the future. Each row represents the *equilibrium distribution*.

Sometimes we converge to a unique equilibrium distribution, and sometimes we don't. Here are some examples to help us think about how this works. (i) In the example above, show that the equilibrium distribution is $(0.6, 0.4)$. [How? Keep taking higher powers of P. You

should see that every row is $(0.6, 0.4)$.] (ii) For each of these choices of P, describe what happens as we compute P^k for larger values of k.

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

In each case, verify that P is a legitimate transition matrix, then compute powers of P and see how they behave.

We're interested, for the most part, in P's that have unique equilibrium distributions. One way to guarantee that is the condition: $p_{ij} > 0$. It's sufficient, but stronger than we need. The main point is that we need some structure to make it work. We'll see the same thing in other contexts.

5 Linear models 1: autoregressions

Next up are linear stochastic processes. Linear processes are incredibly useful because they illustrate the concepts in relatively simple form and are capable of approximating a lot of what we observe about the economy and asset returns. It's possible nonlinearities are important in the real world, but there's enough noise in most real-world situations that it's typically hard to say whether that's true or not. In any case, linear processes are the predominant tool of modern macroeconomics and finance and a good starting point even if we're ultimately interested in something else.

One of the simplest linear processes is a first-order autoregression or AR(1). We'll use it over and over again. A random variable x_t evolves through time according to according to

$$x_t = \varphi x_{t-1} + \sigma w_t. \tag{1}$$

Here $\{w_t\}$ is a sequence of random variables, all of them standard normal (normal with mean zero and variance one) and mutually independent. The w's are commonly referred to as "errors," "disturbances," or "innovations." Equation (1) is referred to as an autoregression because x_t is regressed on a lag of itself ("auto" = "same") and first-order because one lag of x_t appears on the right.

Note that we've switched from z_t to x_t here. In this case we use the letter x to refer to both the state and a specific random variable that depends on the state. It's similar to our treatment of random variables earlier, where the random variable is a function of the state, but in most cases we could ignore the latter. Here it's more complicated. We'll see examples shortly in which x_t remains the random variable of interest but the state z_t is a more complicated object.

It's important to be clear about what we know and when we know it. If we go back to the event tree, we usually assume that we know where we are. If we've experienced a history z^t, then we know that, and therefore know our current location or node in the tree. Here we make a similar assumption: that we know all the x's and w's up through date t, but do not know their future values.

Since each w_t is independent of every other one, their covariances are

$$\text{Cov}(w_t, w_{t-k}) \quad = \quad E(w_t w_{t-k}) \quad = \quad \begin{cases} 1 \text{ if } k = 0 \\ 0 \text{ otherwise.} \end{cases}$$

That is, the expectation of "cross terms" is zero. These conditions define the w_t's as having no dynamics: what happens at one date is independent of all other dates. Therefore any dynamics must come from the rest of the equation — in this case the x_t terms.

Let's go back to the autoregression, equation (1) and ask ourselves: (i) Is it Markov? (ii) Is it stationary? (iii) Does it have a unique equilibrium distribution? We'll consider each in turn.

On question (i): Yes, it's Markov. If we take equation (1) updated one period, we see that the conditional distribution of x_{t+1} depends only in the state x_t. What is the distribution? Well, since it comes from w_{t+1}, which is normal, it's also normal. So once we know the mean and variance we're done. The conditional mean is

$$E(x_{t+1}|x_t) \quad = \quad E(\varphi x_t + \sigma w_{t+1}|x_t) \quad = \quad \varphi x_t.$$

Why? Because we know x_t (that's what we meant by conditional on x_t) but we don't know w_{t+1} (it hasn't happened yet). So we take the mean of w_{t+1}, which is zero. What about the variance of the conditional distribution — what we might call the conditional variance? The variance, as we know, is the expected squared difference between the random variable and its mean. The conditional variance is therefore

$$\text{Var}(x_{t+1}|x_t) \quad = \quad E\left[(x_{t+1} - \varphi x_t)^2|x_t\right] \quad = \quad E\left[(\sigma w_{t+1})^2|x_t\right] \quad = \quad \sigma^2.$$

Thus we have: $x_{t+1}|x_t \sim \mathcal{N}(\varphi x_t, \sigma^2)$. This conditional distribution is the analog of a row of the transition matrix in a Markov chain.

On question (ii): Yes, it's stationary. Nothing in (1) depends on the date, including the distribution of the disturbance w_t.

Now about question (iii): Is the process stable? What is the equilibrium distribution? Is it unique? The answer is yes if $|\varphi| < 1$, but let's show that for ourselves. One approach is to compute the distribution directly from (1). The logic here is to presume an equilibrium distribution exists and compute its properties. It's the same distribution for x_t and x_{t-1}, so it has the same mean at both dates:

$$E(x) \quad = \quad \varphi E(x) \quad \Rightarrow \quad E(x) \quad = \quad 0/(1-\varphi) \quad = \quad 0.$$

If we added an intercept to (1) we would get something else. [Try it. The simplest approach is to define $y_t = x_t + \mu$ and see how μ changes our calculations.] What about the variance? Again, the distributions of x_t and x_{t+1} are the same, so we have

$$\text{Var}(x) \quad = \quad E(x_t^2) \quad = \quad E\left[(\varphi x_{t-1} + \sigma w_t)^2\right] \quad = \quad \varphi^2 \text{Var}(x) + \sigma^2.$$

That gives us $\text{Var}(x) = \sigma^2/(1-\varphi^2)$. This works if $\varphi^2 < 1$, but if not the variance is negative, which we know can't be. So if $\varphi^2 < 1$, the equilibrium distribution is normal with

mean zero and variance $\sigma^2/(1-\varphi^2)$. If $\varphi^2 \geq 1$, we don't converge: there is no equilibrium distribution.

The stability property is more obvious if we go through the effort of deriving the equilibrium distribution as a limit. Suppose we think of x_{t+k} from the perspective of date t. If the current state is x_t, what is the distribution of x_{t+k}? This is the conditional distribution of x_{t+k}: conditional on the current state x_t. Repeated application of (1) gives us

$$x_{t+k} = \varphi^k x_t + \sigma\left[w_{t+k} + \varphi w_{t+k-1} + \cdots + \varphi^{k-1} w_{t+1}\right].$$

The conditional mean and variance are

$$\begin{aligned} E(x_{t+k}|x_t) &= \varphi^k x_t \\ \text{Var}(x_{t+k}|x_t) &= \sigma^2\left[1 + \varphi^2 + \cdots + \varphi^{2(k-1)}\right] = \sigma^2(1-\varphi^{2k})/(1-\varphi^2). \end{aligned}$$

The second one follows because $E(w_t^2) = 1$ and $E(w_t w_{t+j}) = 0$ for $j \neq 0$. Does this settle down as we increase k? Yes, as long as $\varphi^2 < 1$. The conditional mean converges to zero and the conditional variance converges to $\sigma^2/(1-\varphi^2)$, the same answers we had before.

*** Law of iterated expecations...
http://en.wikipedia.org/wiki/Law_of_total_expectation

*** Var is var of cond mean plus forecast error
http://en.wikipedia.org/wiki/Law_of_total_variance

*** Multistep forecasting...

We can do the same for high-order autoregressions: versions of (1) with additional lags of x_{t+1} on the right. Examples include

$$\begin{aligned} \text{AR}(1): \quad x_t &= \varphi x_{t-1} + \sigma w_t \\ \text{AR}(2): \quad x_t &= \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \sigma w_t \\ \text{AR}(p): \quad x_t &= \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + \sigma w_t. \end{aligned}$$

These models are stationary by construction. They're Markov if we define the state at $t-1$ by $y_{t-1} = (x_{t-1}, \ldots, x_{t-p})$. And they're stable under some conditions on the φ's that we'll leave alone for now. We mention it briefly in Section 9.

6 Digression: components of variance

The AR(1) is a good example for illustrating the behavior of the variance over different time horizons. Consider the behavior of x_{t+k} from the perspective of date t. We can write

$$x_{t+k} = E(x_{t+k}|x_t) + \left[x_{t+k} - E(x_{t+k}|x_t)\right],$$

the sum of two components. Since we're lazy, our first step is to use more compact notation:

$$x_{t+k} = E_t(x_{t+k}) + \left[x_{t+k} - E_t(x_{t+k})\right],$$

where $E_t(x_{t+k}) = E(x_{t+k}|x_t)$ is shorthand for "the mean of x_{t+k} conditional on the state at date t." In both versions, the first component is the conditional mean, which we might think of as a forecast. The second component is the forecast error. The components are independent, because the first one depends on w's that happened at t or before and the second one depends on w' that happened after. [Use the expressions in the previous section to remind yourself of this.]

The variance of x_{t+k} can be broken into similar components. Since the two terms are independent, we have

$$\begin{aligned} \text{Var}(x_{t+k}) &= E\left\{E_t(x_{t+k}) + \left[x_{t+k} - E_t(x_{t+k})\right]\right\}^2 \\ &= E\left[E_t(x_{t+k})\right]^2 + E\left[x_{t+k} - E_t(x_{t+k})\right]^2. \end{aligned}$$

This seems a little mysterious, so let's substitute the expressions for the AR(1). That gives us

$$\begin{aligned} \text{Var}(x_{t+k}) &= E\left(\varphi^k x_t\right)^2 + E\left[(\sigma^2(1-\varphi^{2k})/(1-\varphi^2)\right] \\ &= \varphi^{2k}\sigma^2/(1-\varphi^2) + \sigma^2(1-\varphi^{2k})/(1-\varphi^2). \end{aligned}$$

The first term is the part of the variance of x_{t+k} that's predictable — it comes from the forecast after all. The second is the part that's unpredictable. You might verify that they sum to the variance.

Now think about how the components change as we increase the forecast horizon. As we increase k, our ability to forecast declines, and the forecast $E(x_{t+k}|x_t) = \varphi x_t$ approaches zero, a constant. The variance of this term evidently also approaches zero. The other term grows to make up the difference: as we increase the time horizon k, the variance of the forecast error approaches the equilibrium variance of x. This takes a simple form in the AR(1) case, but the idea is more general.

7 Linear models 2: moving averages

Our next example is one we'll put to work repeatedly: what we call a *moving average*.

The simplest example is a first-order moving average or MA(1):

$$x_t = \sigma\left(\theta_0 w_t + \theta_1 w_{t-1}\right). \tag{2}$$

Here, as in the previous section, w_t is a sequence of independent standard normal random variables. The state z_{t-1} at date $t-1$ is the lagged disturbance w_{t-1}.

Does equation (2) satisfy our conditions? Well, it's stationary (it holds for all t) and Markov (the state variable w_t summarizes the situation at t). This is essential, so let's go through it slowly. How do we know what the state is? The state is whatever we need to say what the distribution of x_{t+1} is as of date $t1$. In this example, we need to know w_t, but w_{t+1} is simply part of the distribution of x_{t+1}. So the state at date t is w_t.

Is equation (2) stable? Let's see. The conditional mean and variance of x_{t+1} are

$$\begin{aligned} E_t(x_{t+1}) &= \sigma\theta_1 w_t \\ \mathrm{Var}_t(x_{t+1}) &= \sigma^2\theta_0^2. \end{aligned}$$

As above, the subscript t on E and Var means "conditional on the state z at date t": $E_t(x_{t+1}) = E_t(x_{t+1}|z_t)$ and $\mathrm{Var}_t(x_{t+1}) = \mathrm{Var}(x_{t+1}|z_t)$.

What about x_{t+2}? From (2) we have

$$x_{t+2} = \sigma\left(\theta_0 w_{t+2} + \theta_1 w_{t+1}\right).$$

Viewed from the perspective of the state w_t at t, none of this is known. The conditional mean and variance are therefore

$$\begin{aligned} E_t(x_{t+2}) &= 0 \\ \mathrm{Var}_t(x_{t+2}) &= \sigma^2\left(\theta_0^2 + \theta_1^2\right). \end{aligned}$$

Therefore, conditional on the state at t, x_{t+2} is normal with mean zero and variance $\sigma^2(\theta_0^2 + \theta_1^2)$. What about x_{t+3}? It should be clear that it has the same conditional distribution as x_{t+2}. An MA(1) has a one-period memory, so once we get beyond one period the conditional distribution is the same. Evidently we've converged to the equilibrium distribution: normal with mean zero and variance $\sigma^2(\theta_0^2 + \theta_1^2)$.

How about an MA(2)? It has the form

$$x_t = \sigma\left(\theta_0 w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}\right).$$

It's Markov with state $z_t = (w_t, w_{t-1})$. If we computed conditional distributions, we'd find that the effect of the state disappears, in this case, after two periods. We'll skip that for now, but consider a more general case shortly.

Moving averages in general have the form

$$\begin{aligned} \mathrm{MA}(1): \quad x_t &= \sigma\left(\theta_0 w_t + \theta_1 w_{t-1}\right) \\ \mathrm{MA}(2): \quad x_t &= \sigma\left(\theta_0 w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}\right) \\ \mathrm{MA}(q): \quad x_t &= \sigma\left(\theta_0 w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}\right) \\ \mathrm{MA}(\infty): \quad x_t &= \sigma\left(\theta_0 w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \theta_3 w_{t-3} + \cdots\right). \end{aligned}$$

Since the infinite MA includes the others as special cases, let's consider its properties. Is it stationary? Yes, the equation holds at all dates t. Is it Markov? Yes, if we allow ourselves an infinite dimensional state $z_t = (w_t, w_{t-1}, \ldots)$. This goes against our goal of simplicity, but we'll see that it's convenient for other reasons.

Is the infinite MA stable? Let's see how the conditional distribution of x_{t+k} changes as we increase k. The random variable x_{t+k} is, for any positive k,

$$x_{t+k} = \sigma\left(\theta_0 w_{t+k} + \theta_1 w_{t+k-1} + \cdots + \theta_{k-1} w_{t+1}\right) + \sigma\left(\theta_k w_t + \theta_{k+1} w_{t-1} + \cdots\right).$$

Conditional on $z_t = (w_t, w_{t-1}, \ldots)$, the second collection of w_t's is known, but the first is not. The conditional mean and variance are therefore

$$\begin{aligned} E_t\left(x_{t+k}\right) &= \sigma\left(\theta_k w_t + \theta_{k+1} w_{t-1} + \cdots\right) \\ \mathrm{Var}_t\left(x_{t+k}\right) &= \sigma^2\big(\theta_0^2 + \theta_1^2 + \cdots + \theta_{k-1}^2\big). \end{aligned}$$

Do they converge as we increase k? Evidently we need

$$\lim_{k\to\infty} \theta_k = 0$$

for the mean to converge. [This isn't a tight argument, but it'll do for us.] And we need

$$\lim_{k\to\infty} (\theta_0^2 + \theta_1^2 + \cdots + \theta_{k-1}^2) = \sum_{k=0}^{\infty} \theta_k^2 < \infty$$

for the variance to converge. (The notation "$< \infty$" means here that the sum converges.) The distribution of x_{t+k} for k sufficiently far in the future is then normal with mean zero and variance $\sigma^2 \sum_{k=0}^{\infty} \theta_k^2$. The second condition is sometimes phrased as "the moving average coefficients are square summable." The second implies the first, so we'll stick with that. Roughly speaking, we need the squared moving average coefficients to go to zero at a fast enough rate.

One of the things that makes moving averages so useful is that even autoregressions can be written this way. Consider the AR(1), equation (1). If we substitute backwards, we find

$$\begin{aligned} x_t &= \varphi x_{t-1} + \sigma w_t \\ &= \sigma w_t + \varphi\left(\varphi x_{t-2} + \sigma w_{t-1}\right) \\ &= \sigma\big(w_t + \varphi w_{t-1} + \varphi^2 w_{t-2} + \cdots\big). \end{aligned}$$

We refer to this as the "moving average representation" of the AR(1). The AR(1) simply applies some structure to the moving average coefficients, namely that they decline geometrically.

This representation is sometimes useful in finding the equilibrium distribution. For example, the mean here is zero: take expectations of both sides. The variance is

$$\mathrm{Var}(x_t) = \sigma^2 E\big(w_t + \varphi w_{t-1} + \varphi^2 w_{t-2} + \cdots\big)^2,$$

which converges, as we've seen, if $\varphi^2 < 1$.

8 Linear models 3: ARMA(1,1)'s

We now have two linear models at our disposal: autoregressions and moving averages. Combining them gives us a wide range of behavior with a small number of parameters. One of my favorites is the ARMA(1,1):

$$x_t = \varphi_1 x_{t-1} + \sigma\left(\theta_0 w_t + \theta_1 w_{t-1}\right).$$

This is Markov with date t state $z_t = (x_t, w_t)$.

Its properties are evident from its moving average representation. Repeated substitution gives us

$$x_t = \sigma\theta_0 w_t + \sigma(\varphi_1\theta_0 + \theta_1)w_{t-1} + \sigma(\varphi_1\theta_0 + \theta_1)\varphi_1 w_{t-2} + \sigma(\varphi_1\theta_0 + \theta_1)\varphi_1^2 w_{t-3} + \cdots .$$

That is: the first two moving average coefficients are arbitrary, then they decline at rate φ_1. Variants of this model are the basis of the most popular models of bond pricing. One useful feature is that its conditional mean is autoregressive:

$$E_t(x_{t+1}) = \sigma(\varphi_1\theta_0 + \theta_1)w_t + \sigma(\varphi_1\theta_0 + \theta_1)\varphi_1 w_{t-1} + \sigma(\varphi_1\theta_0 + \theta_1)\varphi_1^2 w_{t-2} + \cdots .$$

How do we know it's autoregressive? Because the moving average coefficients decline at a constant rate.

Comment: This is simpler if we set $\theta_0 = 1$, a convenient normalization as long as $\theta_0 \neq 0$.

9 Matrix representations (skip)

We can extend all of this to vector processes and, moreover, express univariate ARMA models in compact matrix form. The standard form is an AR(1) in vector form:

$$x_t = Ax_{t-1} + Bw_t, \tag{3}$$

where x_t and w_t are both vectors. This has the same structure as equation (1) with the matrix A playing the role of φ and B the role of σ. It's a stationary Markov process in the state x_t. It's stable if the eigenvalues of A are all less than one in absolute value: if A^k approaches zero as we increase k. (This last statement will make sense if you have taken a linear algebra course, but not if you haven't. See, for example, Steven Pinker's comment.)

Finite ARMA models — ARMA(p,q) for finite p and q — are conveniently expressed in this form with the appropriate definition of the state. An ARMA(1,1), for example, can be expressed

$$\begin{bmatrix} x_t \\ w_t \end{bmatrix} = \begin{bmatrix} \varphi_1 & \theta_1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ w_{t-1} \end{bmatrix} + \begin{bmatrix} \theta_0 \\ 1 \end{bmatrix} [w_t].$$

In practice, then, if we can handle the vector process (3) we can handle finite ARMA models.

10 Autocorrelation functions

We don't observe moving average coefficients, but we do observe one consequence of them: the autocovariance and autocorrelation functions.

We can compute both from the moving average representation. Suppose we have a time series with moving average representation

$$x_t = \mu + \sum_{j=0}^{\infty} a_j w_{t-j}.$$

(The reason for the change in notation should be apparent soon.) The mean is evidently μ. What is the covariance between x_t and x_{t-k} for $k \geq 0$? We've seen that we can easily compute sample analogs. The covariance is

$$\begin{aligned}\gamma_x(k) &= \mathrm{Cov}(x_t, x_{t-k}) = E\left[(x_t - \mu)(x_{t-k} - \mu)\right] \\ &= E\left(\sum_{j=0}^{\infty} a_j w_{t-j}\right)\left(\sum_{j=0}^{\infty} a_j w_{t-k-j}\right) \\ &= \sum_{j=k}^{\infty} a_j a_{j-k} = \sum_{j=0}^{\infty} a_j a_{j+k}.\end{aligned}$$

We eliminated the cross terms because their expectation is zero [$E(w_t w_{t-j}) = 0$ for $j \neq 0$]. The notation $\gamma_x(k)$ is standard. The variance is the value for $k = 0$,

$$\gamma_x(0) = \mathrm{Cov}(x_t, x_t) = \sum_{j=0}^{\infty} a_j^2, \tag{4}$$

which we've seen before.

We refer to $\gamma_x(k)$, plotted as a function of k, as the *autocovariance function*. The *autocorrelation function* or acf is the same thing scaled by the variance:

$$\rho_x(k) = \gamma_x(k)/\gamma_x(0).$$

The scaling insures that $|\rho(k)| \leq 1$. By construction $\rho_x(0) = 1$. Both are symmetric: we compute them for $k \geq 0$, but you get the same for positive and negative values of k. (Try it, you'll see.)

Example 1. An MA(1) has autocovariances

$$\gamma_x(k) = \begin{cases} \theta_0^2 + \theta_1^2 & k = 0 \\ \theta_0\theta_1 & k = 1 \\ 0 & k > 1. \end{cases}$$

The autocorrelations are

$$\rho_x(k) = \gamma_x(k)/\gamma_x(0) = \begin{cases} 1 & k = 0 \\ \theta_0\theta_1/(\theta_0^2 + \theta_1^2) & k = 1 \\ 0 & k > 1. \end{cases}$$

That's the defining property of an MA(1): it has only a one-period memory. You might verify for yourself that an MA(q) has a q-period memory.

Example 2. We've seen that an AR(1) has an infinite moving average representation with MA coefficients $a_j = \varphi^j$. Therefore the autocovariances (4) are

$$\gamma_x(k) = \sum_{j=0}^{\infty} \varphi^j \varphi^{j+k} = \varphi^k/(1 - \varphi^2).$$

The autocorrelations are therefore

$$\rho_x(k) = \gamma_x(k)/\gamma_x(0) = \varphi^k.$$

In words: the acf declines at the geometric rate φ. In contrast to an MA, it approaches zero gradually.

11 Sample autocorrelations

We can do the same thing with data. You may recall that the sample mean is

$$\bar{x} = T^{-1}\sum_{t=1}^{T} x_t$$

and the sample variance is

$$\gamma_x(0) = T^{-1}\sum_{t=1}^{T}(x_t - \bar{x})^2.$$

The notation follows that of the previous section.

We're typically interested in some aspects of the dynamics of x, summarized in the sample autocovariance and autocorrelation functions. The sample covariance is

$$\gamma_x(k) = T^{-1}\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x}).$$

Since we only have the observations x_t for $t = 1, ..., T$, we need to start the sum at $t = k+1$. By longstanding convention, we nevertheless divide the sum by T rather than $T - k$. We could also consider negative values of k and adjust the range in the sum appropriately.

The shape of $\gamma_x(k)$ is useful in telling us about the dynamics of x, but it's more common to scale it by $\gamma_x(0)$ and convert it to a correlation. The autocorrelation function $\rho_x(k)$ is defined by

$$\rho_x(k) = \gamma_x(k)/\gamma_x(0).$$

Obviously $\rho_x(0) = 1$: x_t is perfectly correlated with x_t. But for other values of k it can take a variety of forms.

When we compute autocorrelations with financial data we see, for example, that autocorrelations for equity returns are very small: returns are virtually uncorrelated over time. Interest rates, however, are very persistent: the autocorrelations decline slowly with k.

Bottom line

There's a lot here, most of it essential to modeling economies and asset returns over time. We'll use stochastic processes with these three properties:

- Markov. The distribution over next period outcomes depends only on this period's state.
- Stationary. The conditional distribution is that same at all dates.
- Stable. The distribution over future outcomes settles down as we increase the time horizon.

We'll put all of them to work.

Practice problems

1. *Probabilities in an event tree.* Consider the event tree in Figure 1. Suppose that at each node the probability of taking the up branch is ω and the probability of the down branch is $1-\omega$. What are the probabilities of the four nodes at $t=2$?

   Answer. We simply multiply the probabilities of the branches taken to get to each node. Starting at the top and working our way down, they are ω^2, $\omega(1-\omega)$, $(1-\omega)\omega$, and $(1-\omega)^2$.

2. *Odd and even days.* Consider a process that differs on odd and even days. Specifically, let $x_t = w_t + \theta w_{t-1}$ with

$$w_t \sim \begin{cases} \mathcal{N}(a,b) & \text{if } t \text{ is even} \\ \mathcal{N}(c,d) & \text{if } t \text{ is odd.} \end{cases}$$

   Is x_t Markov? Stationary? Stable?

   Answer. It's Markov, but not stationary (odd and even days are different) or stable (the conditional distribution of x_{t+1} depends on whether k is even or odd. This kind of thing comes up all the time. We either adjust the data (seasonal adjustment) or somehow include it in our model.

3. *Two-state Markov chain (skip).* We can get a sense of how Markov chains work with a two-state example. A two-state chain is characterized by a 2 by 2 transition matrix P. Because the rows sum to one, P has (essentially) two parameters. A convenient parameterization is

$$P = (1-\varphi)\begin{bmatrix} \omega & 1-\omega \\ \omega & 1-\omega \end{bmatrix} + \varphi \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5}$$

   where the two parameters are ω and φ.

   (a) Under what conditions on (ω, φ) is P a legitimate transition matrix?

   (b) What are the two-period transitions P^2? You can either do this by hand or get Matlab to do it. Either way, the key is to arrange the terms into a form similar to (5).

   (c) What about the k-period transitions?

   (d) What happens as we continue to increase k? What is the equilibrium distribution?

   (e) (extra credit) What are the eigenvalues of P?

   Answer. This question is useful if you're comfortable with linear algebra, but skip it if you're not. This is the simplest possible Markov chain, but it illustrates a number of possibilities.

   (a) The probabilities have to be between zero and one, which gives us these inequalities:

$$\begin{aligned} 0 &\leq (1-\varphi)\omega + \varphi \leq 1 \\ 0 &\leq (1-\varphi)(1-\omega) \leq 1 \\ 0 &\leq (1-\varphi)\omega \leq 1 \\ 0 &\leq (1-\varphi)(1-\omega) + \varphi \leq 1. \end{aligned}$$

That's sufficient for your answer. If you'd like to go further, here's how it works. The second and third inequalities imply (add them together) $-1 \leq \varphi \leq 1$. The first and fourth imply $0 \leq \omega \leq 1$. That's not quite sufficient though. The second and third imply (directly, divide by $1-\varphi$)

$$-\varphi/(1-\varphi) \;\leq\; \omega \;\leq\; 1/(1-\varphi),$$

a joint restriction on ω and φ. If $\varphi \geq 0$ this is irrelevant, it's less restrictive than our earlier condition on ω. But if $\varphi < 0$, it limits the range of ω. For example, if $\varphi = -1/2$, then $1/3 \leq \omega \leq 2/3$.

(b,c) The k-period transitions have the form

$$P \;=\; (1-\varphi^k)\begin{bmatrix} \omega & 1-\omega \\ \omega & 1-\omega \end{bmatrix} + \varphi^k \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

(d) If $|\varphi| < 1$, then as we increase k, $\varphi^k \to 0$ and P converges to the first matrix. The equilibrium distribution is evidently $(\omega, 1-\omega)$.

(e) If you're comfortable with linear algebra, you might notice that P has eigenvalues of 1 and φ. The first is a feature of all Markov chains: since the rows sum to one, there's an eigenvalue of one. The second tells us how fast we converge to the equilibrium distribution.

4. *Properties of an MA(2).* An MA(2) can be written

$$x_t \;=\; \delta + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$$

with $\{w_t\} \sim \text{NID}(0,1)$ (the w's are independent normals with mean zero and variance one).

(a) What is the equilibrium distribution of x?

(b) What are the conditional means, $E_t(x_{t+1})$, $E_t(x_{t+2})$, and $E_t(x_{t+3})$?

(c) What are the conditional variances, $\text{Var}_t(x_{t+1})$, $\text{Var}_t(x_{t+2})$, and $\text{Var}_t(x_{t+3})$?

(d) What is the autocovariance function,

$$\gamma(k) \;=\; \text{Cov}(x_t, x_{t-k}),$$

for $k = 0, 1, 2, 3$?

(e) What is the autocorrelation function? Under what conditions are $\rho(1)$ and $\rho(2)$ positive?

Answer.

(a) Since x is a linear combination of normals, it's normal as well. It's therefore sufficient to say what its mean and variance are. Its mean is

$$E(x_t) \;=\; E(\delta + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}) \;=\; \delta.$$

Its variance is

$$\text{Var}(x_t) \;=\; E(x_t - \delta)^2 \;=\; 1 + \theta_1^2 + \theta_2^2.$$

(b) The conditional means are

$$\begin{aligned} E_t(x_{t+1}) &= E_t(\delta + w_{t+1} + \theta_1 w_t + \theta_2 w_{t-1}) &= \delta + \theta_1 w_t + \theta_2 w_{t-1} \\ E_t(x_{t+2}) &= E_t(\delta + w_{t+2} + \theta_1 w_{t+1} + \theta_2 w_t) &= \delta + \theta_2 w_t \\ E_t(x_{t+3}) &= E_t(\delta + w_{t+3} + \theta_1 w_{t+2} + \theta_2 w_{t+1}) &= \delta. \end{aligned}$$

You can see that as we increase the forecast horizon, the conditional mean approaches the mean.

(c) The conditional variances are

$$\begin{aligned} \mathrm{Var}_t(x_{t+1}) &= E_t[(w_{t+1})^2] &= 1 \\ \mathrm{Var}_t(x_{t+2}) &= E_t[(w_{t+2} + \theta_1 w_{t+1})^2] &= 1 + \theta_1^2 \\ \mathrm{Var}_t(x_{t+3}) &= E_t[(w_{t+3} + \theta_1 w_{t+2} + \theta_2 w_{t+1})^2] &= 1 + \theta_1^2 + \theta_2^2. \end{aligned}$$

You see here that as we increase the time horizon, the conditional variance approaches the variance.

(d) The autocovariance function is

$$\mathrm{Cov}(x_t, x_{t-k}) = \begin{cases} 1 + \theta_1^2 + \theta_2^2 & k = 0 \\ \theta_1 + \theta_1\theta_2 & k = 1 \\ \theta_2 & k = 2 \\ 0 & k \geq 3. \end{cases}$$

(e) Autocorrelations are scaled autocovariances: $\rho(k) = \gamma(k)/\gamma(0)$. $\rho(2)$ is positive if θ_2 is. $\rho(1)$ is positive if $\theta_1(1 + \theta_2)$ is. Both are therefore positive if θ_1 and θ_2 are positive.

5. *Combination models.* Consider the linear time series model

$$x_t = \varphi x_{t-1} + w_t,$$

with $\{w_t\}$ independent normal random variables with mean zero and variance one. Now consider a second random variable y_t built from x_t and the same disturbance w_t by

$$y_t = x_t + \theta w_t.$$

The question is how this combination behaves.

(a) Is there a state variable for which x_t is Markov? What is the distribution of x_{t+1} conditional on the state at date t?

(b) Express x_t as a moving average. What are its coefficients?

(c) Is there a state variable for which y_t is Markov? What is the distribution of y_{t+1} conditional on the state at date t?

(d) Express y_t as a moving average. What are its coefficients?

(e) Under what conditions is y_t stable? That is: under what conditions does the distribution of y_{t+k}, conditional on the state at t, converge as k gets large?

(f) What is the equilibrium distribution of y_t?

(g) What is the first autocorrelation of y_t?

Answer.

(a) It's Markov with state x_t. The conditional distribution of x_{t+1},

$$x_{t+1} \quad = \quad \varphi x_t + w_{t+1},$$

is normal with mean φx_t and variance one.

(b) The moving average representation is

$$x_t \quad = \quad w_t + \varphi w_{t-1} + \varphi^2 w_{t-2} + \cdots .$$

The coefficients are $(1, \varphi, \varphi^2, \ldots)$.

(c) Two answers, both work: the state can be x_t or (more commonly) the vector (y_t, w_t). The distribution of y_{t+1}:

$$y_{t+1} \quad = \quad \varphi x_t + (1+\theta) w_{t+1} \quad = \quad \varphi(y_t - \theta w_t) + (1+\theta) w_{t+1}$$

is (conditionally) normal with mean $\varphi x_t = \varphi(y_t - \theta w_t)$ and variance $(1+\theta)^2$).

(d) If we add θw_t to the expression for x_t above, we get

$$y_t \quad = \quad (1+\theta) w_t + \varphi w_{t-1} + \varphi^2 w_{t-2} + \cdots .$$

(e) It's stable if $|\varphi| < 1$: we need the moving average coefficients to approach zero.

(f) Equilibrium distribution: x_t is normal with mean zero and variance

$$\mathrm{Var}(x_t) \quad = \quad (1+\theta)^2 + \varphi^2 + \varphi^4 + \cdots \quad = \quad (1+\theta)^2 + \varphi^2/(1-\varphi^2).$$

6. *Vector autoregressions (skip).* We can write many linear models in the form

$$x_{t+1} \quad = \quad Ax_t + Bw_{t+1}. \tag{6}$$

Here x is a vector, $w \sim \mathcal{N}(0, I)$ is also a vector (of possibly different dimension), and (A, B) are matrices.

(a) Consider the ARMA(1,1):

$$y_t \quad = \quad \varphi_1 y_{t-1} + \theta_0 w_t + \theta_1 w_{t-1}.$$

Show that this can be expressed in the same form as (6).

(b) Ditto for the ARMA(2,1):

$$y_t \quad = \quad \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \theta_0 w_t + \theta_1 w_{t-1}.$$

(c) For the general model (6), what is the distribution of x_{t+2} given x_t?

(d) Ditto for x_{t+k}. Under what conditions does this converge as k gets large?

Answer. This course is designed to avoid linear algebra. Nevertheless, in this case you might want to know that seemingly complex models can often be written simply in matrix form. In that case, we're dealing with essentially higher-dimensional analogs of an AR(1), which makes programming and insight much easier. If this line of thought doesn't work for you, just ignore it.

(a) The ARMA(1,1) can be written

$$\begin{bmatrix} y_{t+1} \\ w_{t+1} \end{bmatrix} = \begin{bmatrix} \varphi_1 & \theta_1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_t \\ w_t \end{bmatrix} + \begin{bmatrix} \theta_0 \\ 1 \end{bmatrix} [w_{t+1}],$$

which you should recognize from the notes on stochastic processes.

(b) The ARMA(2,1) becomes

$$\begin{bmatrix} y_{t+1} \\ y_t \\ w_{t+1} \end{bmatrix} = \begin{bmatrix} \varphi_1 & \varphi_2 & \theta_1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} y_t \\ y_{t-1} \\ w_t \end{bmatrix} + \begin{bmatrix} \theta_0 \\ 0 \\ 1 \end{bmatrix} [w_{t+1}].$$

(c) Now we can put the AR(1) structure to work. After substituting,

$$x_{t+2} = A^2 x_t + B w_{t+2} + AB w_{t+1}.$$

The first term ($A^2 x_t$) is the conditional mean. The others generate variance. It goes beyond this course, but for a vector like w_t the variance is a matrix, with variances on the diagonal and covariances otherwise. If that's confusing, just skip it.

If not, we write the variance matrix as $E(w_t w_t^\top)$, where $^\top$ is a brute force way to write the transpose. In our case, we assumed the variance matrix for w_t is I, with ones on the diagonal and zeros off.

What's the variance matrix for x_{t+2}? It's defined by

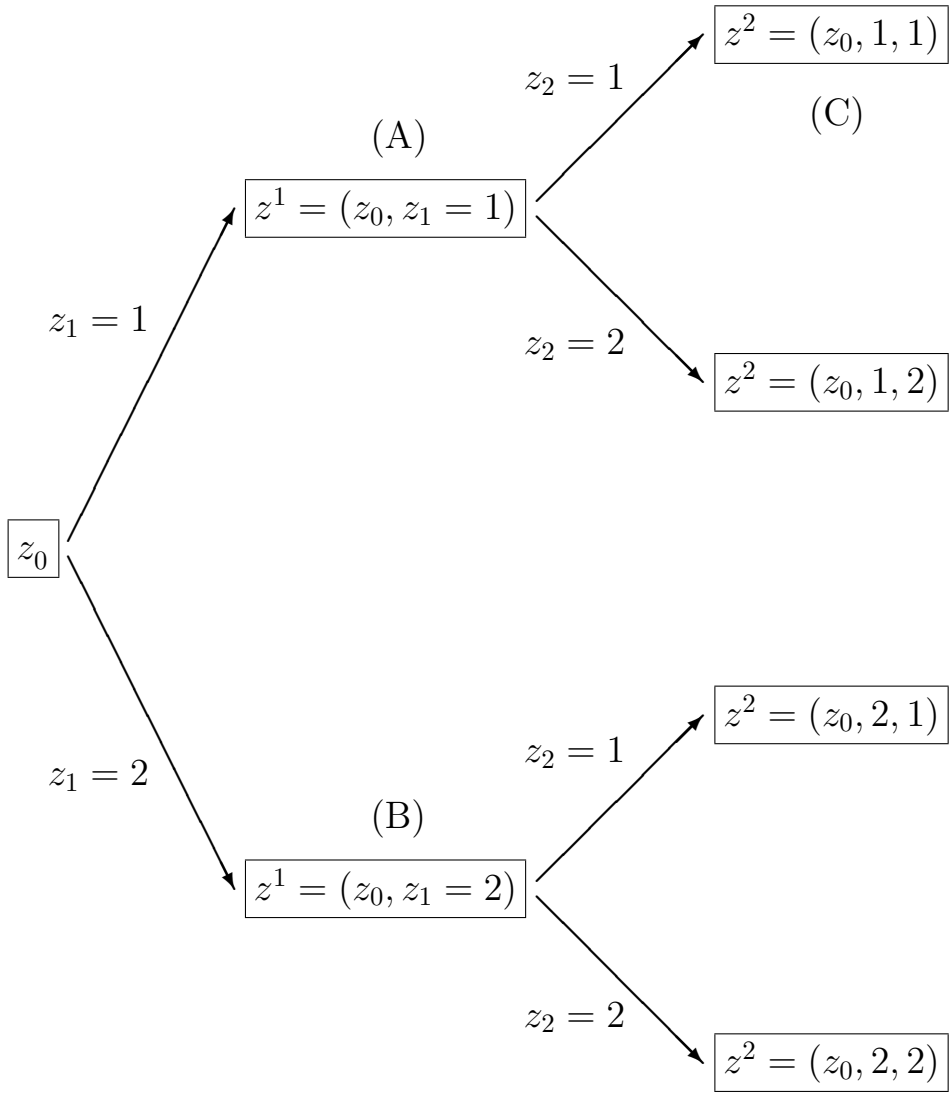
$$\begin{aligned} \text{Var}_t(x_{t+2}) &= E_t\left[(x_{t+2} - Ax_t)(x_{t+2} - Ax_t)^\top\right] \\ &= E_t\left[(Bw_{t+2} + ABw_{t+1})(Bw_{t+2} + ABw_{t+1})^\top\right] \\ &= BB^\top + ABB^\top A^\top. \end{aligned}$$

(d) Same idea repeated:

$$\text{Var}_t(x_{t+k}) = BB^\top + ABB^\top A^\top + A^2 BB^\top (A^2)^\top + \cdots + A^{k-1} BB^\top (A^{k-1})^\top.$$

As k gets large, we get more terms. It converges, though, if the terms shrink fast enough, which requires powers of A to shrink. If you're familiar with linear algebra, you need the eigenvalues of A to be less than one in absolute value.

Figure 1
Representative Event Tree



The figure illustrates how uncertainty unfolds over time. Time moves from left to right, starting at date $t = 0$. At each date t, an event z_t occurs. In this example, z_t is drawn from the set $\mathcal{Z} = \{1, 2\}$. Each node is associated with a box and can be identified from the path of events that leads to it, which we refer to as a history and denote by $z^t \equiv (z_0, ..., z_t)$, starting with an arbitrary initial node z_0. Thus the upper right node follows two up branches, $z_1 = 1$ and $z_2 = 1$, and is denoted $z^2 = (z_0, 1, 1)$. The set $\mathcal{Z}^2$ of all possible 2-period histories is therefore $\{(z_0, 1, 1), (z_0, 1, 2), (z_0, 2, 1), (z_0, 2, 2)\}$, illustrated by the "terminal nodes" on the right. Not shown are conditional probabilities of particular branches, from which we can construct probabilities for each node/history.