ECON-UB 233 Dave Backus @ NYU

# Math Tools: Stochastic Processes

(Started: December 5, 2011; Revised: April 23, 2012)

We look at some simple examples of *stochastic processes*, which combine randomness with dynamics. There's a natural link to bond pricing, our next topic.

It's common to describe randomness in dynamic settings using one-period conditional distributions: given the state at date $t$ ("today"), the probabilities of feasible states at $t+1$ ("tomorrow"). The probabilities over multiple periods then follow by stringing periods together.

We look at three examples. The first is an event tree, the same object we've used before but with more than two dates. The second is a Markov chain, which imposes some regularity on the tree. The third is a set of linear processes, where similar structure...

## 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 you can imagine extending it further.

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

We can extend the idea 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 states 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 $p(z_1 = 1|z_0)$, and the probability of continuing with $z_2 = 1$ is $p[z_2 = 1|z^1 = (z_0, z_1 = 1)]$, then the probability at $z_0$ of getting to the upper right node is the product:

$$p[z_2 = 1|z^1 = (z_0, z_1 = 1)] \quad = \quad p(z_1 = 1|z_0)p[z_2 = 1|z^1 = (z_0, z_1 = 1)].$$

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).

## Digression: matrix multiplication

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 $i$th 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

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

where $c_{ij}$ is the $ij$th 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 you define a matrix, say, you separate rows with semi-colons. The command `A = [1 2 5; 6 4 2]`, for example, defines the 2 by 3 matrix

$$A \quad = \quad \left[ \begin{array}{ccc} 1 & 2 & 5 \\ 6 & 4 & 2 \end{array} \right].$$

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

$$B \quad = \quad \left[ \begin{array}{cc} 1 & 4 \\ 3 & 2 \end{array} \right].$$

You might verify by hand that

$$BA \quad = \quad \left[ \begin{array}{ccc} 25 & 18 & 13 \\ 15 & 14 & 19 \end{array} \right].$$

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?

## Markov chains

Event trees are reasonably general, but we make some progress by imposing more structure than this. One kind of structure is to make the set of states the same at every date. That's true in the picture, but is not a requirement of an event tree.

Another kind of structure is to limit the memory of the conditional probabilities. Rather than specifying the probability of a state $z_{t+1}$ conditional on the complete history $z^t$ to that point, we specify it conditional on the current state $z_t$ alone. This is usually referred to as the *Markov property*. With a finite set of states, the probabilities might be expressed

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

which we can collect in a matrix $P$ (the *transition matrix*). As written, these conditional probabilities do not depend on the date $t$.

Here's an example. Suppose

$$P \quad = \quad \left[ \begin{array}{cc} 0.8 & 0.2 \\ 0.3 & 0.7 \end{array} \right].$$

Row $i$ is the probability distribution over next period's states 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? 10 periods? As it happens, the probability of moving from state $i$ at date $t$ to state $j$ at date $t + k$ is the $ij$th element of $P^k$. So we can compute the whole set of probabilities by computing powers of $P$. In our example, we get

$$P^2 \quad = \quad \left[ \begin{array}{cc} 0.70 & 0.30 \\ 0.45 & 0.55 \end{array} \right], \quad P^3 \quad = \quad \left[ \begin{array}{cc} 0.650 & 0.350 \\ 0.525 & 0.475 \end{array} \right].$$

Try it yourself in Matlab.

We're interested in examples in which the probabilities "settle down." 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 you get far enough in the future. We refer to this as the *stationary* or *equilibrium* distribution.

Here are some examples to think about. (i) In the example above, show that the stationary distribution is $(0.6, 0.4)$. (ii) For each of these choices of $P$, describe what happens as you compute $P^k$ for larger values of $k$.

$$P \quad = \quad \left[ \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right], \quad P \quad = \quad \left[ \begin{array}{ccc} 0.8 & 0.1 & 0.1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right], \quad P \quad = \quad \left[ \begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array} \right].$$

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 stationary distributions. One way to guarantee that is the condition: $p_{ij} > 0$. It's sufficient, but stronger than needed. The main point is that we need some conditions to make it work. We'll see the same thing in other contexts.

## Moving averages

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

The starting point is an iid (independent and identically distributed) series $w_t$. Let's say each $w_t$ is normal with mean zero and variance one. Since each $w_t$ is independent of every other one, their covariances are zero:

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

Independence also tells us that the variance of sums of $w_t$'s is the sum of the variances of the individual terms. For example,

$$\text{Var}(w_t + w_{t-1}) \;=\; E\{[(w_t + w_{t-1}) - 0]^2\} \;=\; E(w_t^2) + E(w_{t-1}^2) + 2E(w_t w_{t-1}) \;=\; 2.$$

The expectations of all the cross-product terms are zero.

The next step is to consider what we know and when we know it. We'll say that we learn $w_t$ at date $t$, so that our information set at date $t$ consists of all the $w_t$'s that have happened to date. That means that $w_t$ is a random variable until date $t$, at which point it becomes a known number. Thus we might write

$$\begin{aligned} E_t(w_{t+k}) &= \begin{cases} 0 \text{ if } k > 0 & \text{(future)} \\ w_t \text{ if } k \leq 0 & \text{(past \& present)} \end{cases} \\ \text{Var}_t(w_{t+k}) &= \begin{cases} 1 \text{ if } k > 0 & \text{(future)} \\ 0 \text{ if } k \leq 0 & \text{(past \& present).} \end{cases} \end{aligned}$$

The subscript $t$ here means conditional on the state at date $t$.

Moving averages are random variables that can be expressed as linear combinations of past $w_t$'s. Here are some examples:

$$\begin{aligned} \text{MA}(1): \quad x_t &= \mu + \theta_0 w_t + \theta_1 w_{t-1} \\ \text{MA}(q): \quad x_t &= \mu + \theta_0 w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} \\ \text{MA}(\infty): \quad x_t &= \mu + \theta_0 w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots . \end{aligned}$$

Linearity here is convenient, as we'll see.

Now let's see how the conditional distribution of $x_{t+k}$ changes as we increase $k$. It's the same thing we did with Markov chains. We'll start with the MA(1) ("first-order moving average"). At date $t$, $x_t$ is a number: we know it and all its inputs. But as of date $t$, $x_{t+1}$ is random:

$$x_{t+1} \;=\; \mu + \theta_0 w_{t+1} + \theta_1 w_t.$$

We know $w_t$ but not $w_{t+1}$. So we might say that $x_{t+1}|t \sim \mathcal{N}(\mu + \theta_1 w_t, \theta_0^2)$. The "$|t$" means conditional on the state at date $t$, the state here being the history of $w_t$'s up to $t$. What about

$$x_{t+2} \;=\; \mu + \theta_0 w_{t+2} + \theta_1 w_{t+1}?$$

By the same logic, we have $x_{t+2}|t \sim \mathcal{N}(\mu, \theta_0^2 + \theta_1^2)$. At this point, the distribution has settled down. The distribution of $x_{t+k}$ conditional on the state at $t$ is the same for all $k \geq 2$. This settles down for the same reason Markov chains settled down.

What about the infinite MA? Consider $x_{t+k}$ for some positive $k$ and group terms into those you know and those you don't:

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

The second collection of $w_t$'s is known at $t$, so the conditional mean $x_{t+k}$ at date $t$ is

$$E_t(x_{t+k}) \;=\; \mu + \theta_k w_t + \theta_{k-1} w_{t-1} + \cdots.$$

The first collection of $w_t$'s is unknown at $t$, so the conditional variance is

$$\mathrm{Var}_t(x_{t+k}) \;=\; \theta_0^2 + \theta_1^2 + \cdots + \theta_{k-1}^2.$$

Do they settle down? By that we mean: Do they converge as we increase $k$? Evidently we need

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

for the mean to converge and

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

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 $\mu$ and variance $\sum_{k=0}^{\infty} \theta_k^2$. As with Markov chains, we'll refer to this as the stationary distribution.

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.

## Autoregressive models and "mixed" models

A second class of linear models are *autoregressions*: regressions of a variable on lags of itself. Examples include:

$$\begin{aligned} \mathrm{AR}(1): \quad x_t &= (1 - \varphi_1)\mu + \varphi_1 x_{t-1} + \theta_0 w_t \\ \mathrm{AR}(p): \quad x_t &= (1 - \varphi_1 - \cdots - \varphi_p)\mu + \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + \theta_0 w_t. \end{aligned}$$

The expression next to $\mu$ is needed to make $\mu$ the mean of the stationary distribution. Alternatively, we could subtract $\mu$ from every $x_{t-k}$ term.

Do these settle down to a unique stationary distribution? Under what conditions? We can see how this works with the AR(1) ("first-order autoregression"). The AR(1) has a "moving average representation" that we construct by repeated substitution:

$$\begin{aligned} x_t &= (1-\varphi_1)\mu + \varphi_1 x_{t-1} + \theta_0 w_t \\ &= (1-\varphi_1)(1+\varphi_1)\mu + \varphi_1^2 x_{t-2} + \theta_0 w_t + \varphi_1\theta_0 w_{t-1} \\ &= \mu + \sum_{k=0}^{\infty} \varphi_1^k \theta_0 w_{t-k}. \end{aligned}$$

That is: the moving average coefficients decline at rate $\varphi_1$. They are square summable if $|\varphi_1| < 1$, in which case the mean is $\mu$ and the variance is

$$\text{Var}(x_t) \;=\; \theta_0^2 \sum_{k=0}^{\infty} \varphi_1^{2k} \;=\; \theta_0^2/(1-\varphi_1^2).$$

You can see here that if $\varphi_1 = 1$, as with a random walk, we don't converge. What's happening here is that the variance keeps increasing as we increase the time horizon, it doesn't settle down. There's nothing wrong with that, but we need to approach it with somewhat different methods. What we usually do is model changes in $x$ or growth rates instead.

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

$$x_t \;=\; (1-\varphi_1)\mu + \varphi_1 x_{t-1} + \theta_0 w_t + \theta_1 w_{t-1}.$$

We construct its moving average representation by repeated substitution:

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

That is: the first two moving average coefficients are arbitrary, then they decline at rate $\varphi_1$. Variants of this model are the basis of the most popular models of bond pricing. One useful feature is that it's conditional mean is autoregressive:

$$E_t(x_{t+1}) \;=\; \mu + (\varphi_1+\theta_1) w_{t-1} + (\varphi_1+\theta_1)\varphi_1 w_{t-2} + (\varphi_1+\theta_1)\varphi_1^2 w_{t-3} + \cdots.$$

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

## 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 representations. Suppose we have a time series with moving average representation

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

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 computed from the stationary distribution is

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

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 $k = 0$:

$$\gamma_x(0) \;=\; \text{Cov}(x_t, x_t) \;=\; \sum_{j=0}^{\infty} a_j^2,$$

which we've seen before.

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

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

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.)

Here are some examples. An MA(1) has

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

What are the autocorrelations? An AR(1) has

$$\begin{aligned}\gamma_x(k) &= \varphi_1^k \theta_0^2/(1-\varphi_1^2) \\ \rho_x(k) &= \varphi_1^k\end{aligned}$$

for $k \geq 0$.

## State space representations

We can extend all of this to vector processes. The classic example is

$$x_t \;=\; Ax_{t-1} + Bw_t. \qquad (1)$$

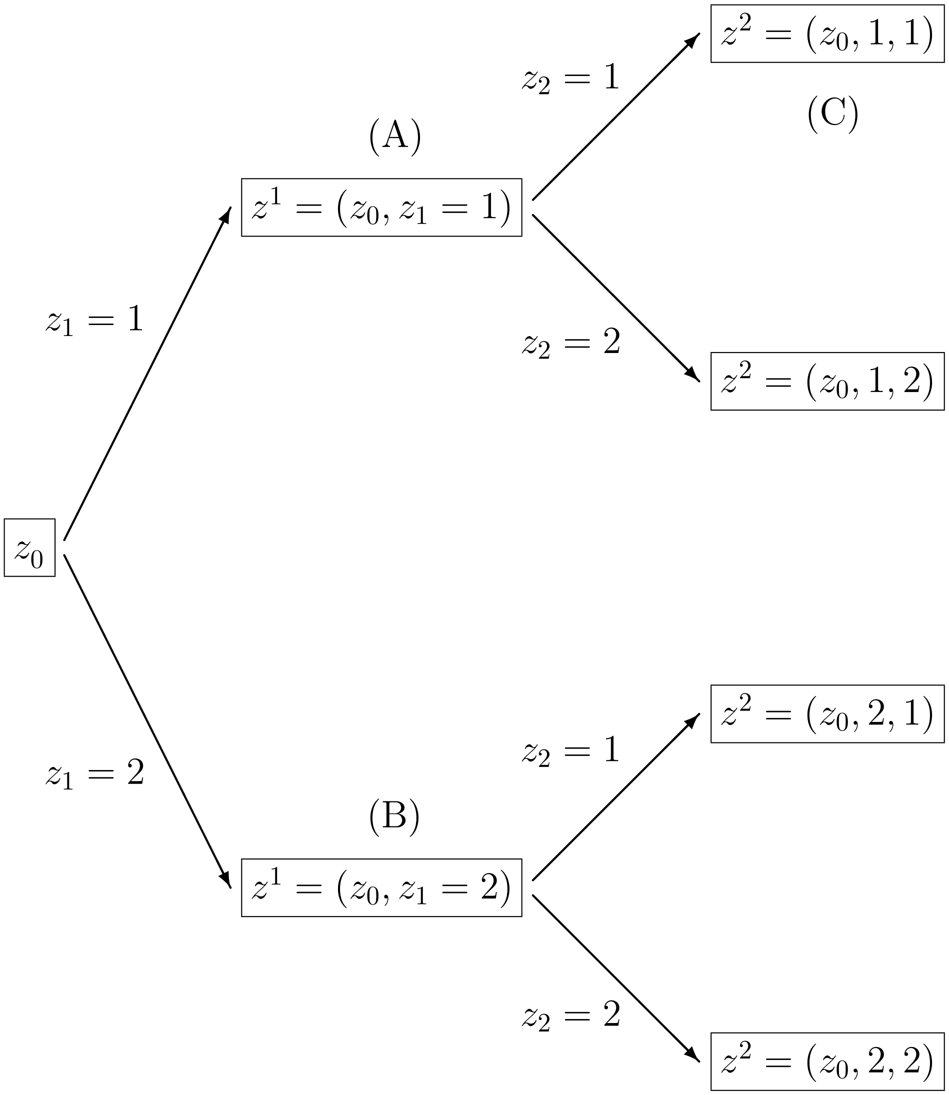
This is an AR(1) in the vector $x$, what is sometimes called a vector autoregression or VAR. We've dropped the mean for convenience, but feel free to add it back if you wish.

We won't do much of this, if any, but finite ARMA models — ARMA($p$,$q$) for finite $p$ and $q$ — are conveniently expressed in this form. An ARMA(1,1) 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 (1) we can handle finite ARMA models.

**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. Note shown are conditional probabilities of particular branches, from which we can construct probabilities for each node/history.