Random Variables Pt. 4

Markov Chains

Elliot Pickens

Jul 15, 2021

1 Intro

So far we've talked about about how random variables operate in a vacuum. We've covered the basic probabilistic concepts, while spending hardly any time elucidating just how powerful these objects can be when a little bit of creativity is applied. In this post we'll explore one way in which random variables can be reared into time dependent model of a system.

2 Markov Chains

Definition 2.1. Stochastic Processes are sequences of random variables where each random variable represents the state of a system at a given time. A sequence with a discrete time parameter will take the form $X_1, X_2, ..., X_n, ...$ with X_1 being the initial state of a the system and each $X_i \; \forall \, i > 1$ representing system at time i. Continuous stochastic processes also exist, but in this post we will focus solely on discrete ones.

Markov chains are a particular sort of discrete stochastic process where the current time state is only dependent on the previous time state. Focusing in on the nth time state in a Markov chain we can say that by definition it only depends upon the $n-1$th time state. We can restate this by saying that for any $n \geq 1$, value b, and sequence of time states $x_1, ..., x_n$

$$P(X_{n+1} \leq b | X_1 = x_1, ..., X_n = x_n) = P(X_{n+1} \leq b | X_n = x_n) \tag{1}$$

is the probability of x_{n+1} conditioned on the existing sequence of states.

Before moving on I would like to provide a little bit of clarification as to exactly what sort of Markov chains we'll be investigating in this post. Here we will only be working with a Markov chains that have been constrained to have only a finite number of possible states. While more general forms do exist, it is easier to introduce Markov chains as only having k possible states. Thus, for each time step we will refer to it's current condition by saying that at time n the chain is in state $1 \leq j \leq k$.

With basic definitions out of the way we can get onto the real mechanics of Markov chains. To begin let's work out the joint pf of the first n states of a Markov chain. We can do this by using the multiplication rule for conditional probabilities to expand the joint pf to

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) && (2)\\ &= P(X_1 = x_1)P(X_2 = x_2 | X_1 = x_1)... && (3)\\ &\ldots P(X_n = x_n | X_{n-1} = x_{n-1}) && (4) \end{aligned}$$

And we can also show the probability of $m > 0$ future states conditioned on $X_n = x_n$ in a similar way to get

$$\begin{aligned} &P(X_{n+1} = x_{n+1}, X_{n+2} = x_{n+2}, ..., X_{n+m} = x_{n+m} | X_n = x_n) && (5)\\ &= P(X_{n+1} = x_{n+1} | X_n = x_n)P(X_{n+2} = x_{n+2} | X_{n+1} = x_{n+1})... && (6)\\ &\ldots P(X_{n+m} = x_{n+m} | X_{n+m-1} = x_{n+m-1}) && (7) \end{aligned}$$

Much of our focus when working with Markov chains is understanding how exactly the process shifts from state to state. The first step towards that goal is understanding what we call transmission distributions.

Definition 2.2. Transmission Distributions are the conditional distributions that describe the movement of the system between states. An example of such a distribution is $P(X_{n+1} = j|X_n = i)$ for $i, j = 1, ..., k$ and $n \geq 1$.

A special case of these distributions are **stationary transmission distributions** where the probability of shifting from one state to another remains the same regardless of what time step we are on (for all n). This allows us to establish constant probabilities to describe a shift from state i to j that we denote p_{ij}. More directly

$$P(X_{n+1} = j|X_n = i) = p_{ij} \, \forall \, n \tag{8}$$

which can be further shortened to

$$g(j|i) = p_{ij} \, \forall \, n, i, j \tag{9}$$

If we know that the transition distributions for all possible state changes are stationary then we can create what we call a **transmission matrix**. Assuming our transmission probabilities are given by $p_{ij} = P(X_{n+1} = j|X_n = i) \, \forall \, n, j, i$ then we define the transmission matrix $\mathbf{P}$ of our Markov chain to be a $k \times k$ matrix where all entries are p_{ij} values. The overall composition of the matrix will be

$$\mathbf{P} = \begin{bmatrix} p_{11} & \cdots & p_{1k} \\ p_{21} & \cdots & p_{2k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{bmatrix} \tag{10}$$

A few notable things about these matrices are that they must be composed of exclusively non-negative elements, and each row must sum to 1. The non-negativity follows directly from the fact that each element is itself a probability value, while the assertion that each row must sum to 1 is a byproduct of the conditionality woven into the matrix through each p_{ij} value. Since each row i is built out of the conditional probabilities given by $g(\cdot|i)$ then the sum $\sum_{j=1}^{k} p_{ij}$ must equal one, because i must transition to one of the other states (or stay the same) 100% of the time.

I won't wander off into uses of these transmission matrices at this time, but to give a little intuition as to how they might be used I would ask what happens if multiply a $1 \times k$ matrix representing the current state of the system with $\mathbf{P}$? What might that tell us about how the system could evolve?

$\mathbf{P}$ and all transmission matrices fall under the broader category of matrices called **stochastic matrices**. A stochastic matrix is a square matrix composed of non-negative entries with the property that every row must sum to 1. Clearly $\mathbf{P}$ is such a matrix, and that every $k \times k$ stochastic matrix is a valid transmission matrix for a finite Markov chain with stationary transmission probabilities and k possible states.

2.1 Multi-Step Transitions

To answer the question posed a few paragraphs back we can take a diversion to spend some time thinking about multi-step transmissions. A single transmission matrix can be used to understand what the next state of a system might be, so by stacking multiple matrices together we can find the probability of landing in some state j from state i after m steps. Thus, we can say that the mth power $\mathbf{P}^m$ of $\mathbf{P}$ has elements p_{ij}^m that represent the chance of moving from i to j in m steps.

Note that the rows of $\mathbf{P}^m$ maintain the conditional property held by $\mathbf{P}$. This means that the ith row of $\mathbf{P}^m$ holds the conditional distribution of X_{n+m} given $X_n = i$ (where $n, m \geq 1$ and $i = 1, ..., k$).

2.1.1 Absorbing States

An absorbing state is one where we have a 100% chance of staying in the same state in the next step. In terms of transmission probabilities, i is an absorbing state if $p_{ii} = 1$.

2.2 Initial Distribution

Transmission matrices are a fantastic way of explaining how a system might evolve, but we currently lack the proper input to allow us to simulate a system based on starting point. The input we need for this in an initial distribution. An initial distribution is a $1 \times k$ matrix where the ith entry represents the probability of the system being in that state. If for example, we have a system where $k = 4$ and we know the system must start in the first state then the initial distribution will be $[1, 0, 0, 0]$. Alternatively if we knew that it was just as likely to be in one state as it was another our initial distribution (often denoted ν) would be $[0.25, 0.25, 0.25, 0.25]$.

Naturally this means that the output of $\nu\mathbf{P}^m$ will be another $1 \times k$ matrix that represents the chance of being in each state after m steps. We don't, however, always have to use matrix multiplication to make use of an initial distribution. By plucking out individual values from ν we can things like rewrite 2 as $P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = v_{x_1}p_{x_1x_2}\dots p_{x_{n-1}x_n}$.

It is also possible to find the marginal distributions of any step using ν and $\mathbf{P}$. As I mentioned in the previous paragraph $\nu\mathbf{P}^m$ gives the probability of being in each state after m steps, but if we just rewind back one step to $\nu\mathbf{P}^{m-1}$ we get the marginal distribution of the mth state (we subtract 1 because the ν is the first state).

We can prove this by recognizing that in order to find the probability values of the marginal distribution of the mth step we need to somehow find the probability of reaching each state by summing over the probability of each path we can take to get there. This is equivalent to the sum

$$P(X_m = x_m) = \sum_{x_{m-1}=1}^{k} \cdots \sum_{x_2=1}^{k} \sum_{x_1=1}^{k} v_{x_1}p_{x_1x_2}\dots p_{x_{m-1}x_m} \tag{11}$$

To reason through this we start from the inner most sum, which only interacts with $v_{x_1}p_{x_1x_2}$ and results in $\nu\mathbf{P}$. As we work outward each new sum is equivalent to multiplying by an additional $\mathbf{P}$. Thus, the second sum is equal to $\nu\mathbf{P}\mathbf{P} = \nu\mathbf{P}^2$, and the eventual $m - 1$th sum returns $\nu\mathbf{P}^{m-1}$.

2.3 Stationary Distributions

Earlier we defined stationary transmission distributions in 2.2 as transmission distributions that remains fixed over time. Here in this section we'll discuss stationary distributions, which are more general but share the important commonality that they are unaffected by time.

One way to define a stationary distribution is to use the transmission matrix $\mathbf{P}$ of a Markov chain. In this case we say that a probability vector ν is a stationary distribution if $\nu\mathbf{P} = \nu$ is true.

Therefore, if we are in a stationary distribution ν then the probability of being in some state i after n steps is the same as being in the ith state from the start. That is to say that where we might be is independent of time. It does not mean that we are just stuck in some state (in an absorbing state), or that the chain is not changing over time. Unless the distribution is composed completely of absorbing states (which will be stationary by definition) then the chain will jump between states at rates determined by ν. It also does not mean that if we take in additional information (like knowledge of the chain being in specific states at specific times) that we just have to default to ν and cannot do any other sort of calculations.

2.3.1 Finding Stationary Distributions

With a little linear algebra we can easily find stationary distributions of $\mathbf{P}$. Given we know that $\nu\mathbf{P} = \nu$ then we also have that $\nu\,[\mathbf{P} - \mathbf{I}[= \mathbf{0}$ since $\nu = \nu\mathbf{I}$ where $\mathbf{I}$ is a $k \times k$ identity matrix and $\mathbf{0}$ is a $1 \times k$ vector of all zeroes. Ideally we would be able to solve this system of equations directly, but sadly we cannot due to there being far too many possible solutions to identify a valid one. The reason for this is that if a valid solution $\mathbf{v}$ exists then $c\mathbf{v}$ is also a solution $\forall c \in \mathbb{R}$, because having k variables and k equations implies the existence of a redundant equation.

We can sidestep this issue by asserting that the elements of our solution $\mathbf{v}$ sum to 1. We do this by creating a new matrix $\mathbf{G}$ that is equal to $\mathbf{P} - \mathbf{I}$ with the one important change that the final column of $\mathbf{G}$ must be all ones. Then we can solve

$$\mathbf{vG} = (0, ..., 0, 1) \tag{12}$$

to find a unique stationary distribution if it exists. To solve this equation we will invoke the existence of $\mathbf{G}^{-1}$, which is the inverse of $\mathbf{G}$ with the property $\mathbf{GG}^{-1} = \mathbf{G}^{-1}\mathbf{G} = \mathbf{I}$. This inverse begets the solution

$$\mathbf{v} = (0, ..., 0, 1)\mathbf{G}^{-1} \tag{13}$$

The only drawback to this method is that if $\mathbf{G}$ is singular and has no inverse then it cannot be used. Unfortunately, this means that we cannot use it to find stationary distributions when multiple exist, because under those conditions our $\mathbf{G}$ will be singular.

2.3.2 Converging to a Stationary Distribution?

I'm going to end this section with a quick theorem. If there exists an m such that all entries of $\mathbf{P}^m$ are strictly positive we can also say that

- we have a unique stationary distribution ν for our Markov chain,
- the limit $lim_{n\to\infty}\mathbf{P}^m$ is $[\nu, \nu, ..., \nu]$ i.e. a matrix where all rows are ν, and
- the distribution of the Markov chain will converge to ν as the number of steps $n \to \infty$ regardless of the starting distribution.

I'm not going to spend much time on this theorem, but I will point out that the third statement seems to follow directly from the second.

2.3.3 Other Stationary Stochastic Processes

The concept of stationary processes exists beyond the confines of Markov chains, and is crucial to things like time series analysis. The ideas and intuition in these other cases are the same as the are for Markov chains in that they center themselves upon time independence. The definitions are, however, a little different. Instead of using a transmission matrix, joint CDFs of sequences of random variables separated by some arbitrary time parameter or the auto-covariance of the process are used to determine stationarity.

At the end of this post (5) I've included a few problem write-ups for questions about these more general stationary processes. To be completely honest I'm including them here, because I don't have anywhere better to put them and figure I might as well. Hopefully someone finds them interesting or helpful.

3 Conclusion

In this post we scratched the surface of Markov chains. The goal here was to touch on the most important concepts needed to use Markov chains, or at least read about how they are employed by researchers and practitioners. They really are a great tool to have in your tool belt even if you only ever break it out when you need to figure out what is going on in some complex model based upon them. I highly recommend reading deeper into Markov chains or stochastic processes in general if you found this post at all interesting.

With the end of this post I've reached the end of this mini series on random variables, but there is still plenty more probability to cover. Next up is expectation.

4 Acknowledgments

These notes were based on *Probability and Statistics (Fourth Edition)* by DeGroot & Schervish.

5 Worked Stationary Process Examples

Problem 1

Starting from a random walk defined as follows:

$$X_0 = 0,\ X_t = \sum_{j=1}^{t} \xi_j,\ t = 1, 2, ... \tag{14}$$

With each ξ_j taking on a value of ± 1 with probability $\frac{1}{2} - \frac{1}{2}$, and all $\xi_j s$ being i.i.d. We want to prove that the random walk $\{X_t\}$ is not weakly stationary.

Our first step is to evaluate the expectation and variance of each step in the process.

$$\mathbb{E}[\xi_j] = \frac{-1+1}{2} = 0 \tag{15}$$

$$\mathbb{E}[\sum_{j=1}^{t} \xi_j] = \sum_{j=1}^{t} \mathbb{E}[\xi_j] = t0 = 0 \tag{16}$$

$$Var(\sum_{j=1}^{t} \xi_j) = Var(\xi_1 + \xi_2 + ... + \xi_t) = tVar(\xi) = t\sigma^2 \tag{17}$$

Now that we have both variance and expectation of our walk we can check the autocovariance of the walk.

$$\begin{aligned}
Cov(X_t, X_{t+k}) &= \mathbb{E}[(X_t - \mathbb{E}[X_t])(X_{t+k} - \mathbb{E}[X_{t+k}])] = \mathbb{E}[X_t X_{t+k}] - \mathbb{E}[X_t]\,\mathbb{E}[X_{t+k}] \\
&= \mathbb{E}[(\sum_{j=1}^{t} \xi_j)(\sum_{j=1}^{t+k} \xi_j)] - \mathbb{E}[\sum_{j=1}^{t} \xi_j]\,\mathbb{E}[\sum_{j=1}^{t+k} \xi_j] \\
&= \mathbb{E}[(\sum_{j=1}^{t} \xi_j)(\sum_{j=1}^{t} \xi_j + \sum_{j=t+1}^{k} \xi_j)] - \mathbb{E}[\sum_{j=1}^{t} \xi_j]\,\mathbb{E}[\sum_{j=1}^{t} \xi_j + \sum_{j=t+1}^{k} \xi_j] \\
&= \mathbb{E}[\sum_{j=1}^{t} \xi_j (\sum_{j=1}^{t} \xi_j)] + \mathbb{E}[\sum_{j=1}^{t} \xi_j (\sum_{j=t+1}^{k} \xi_j)] - \mathbb{E}[\sum_{j=1}^{t} \xi_j]\,\mathbb{E}[\sum_{j=1}^{t} \xi_j] - \mathbb{E}[\sum_{j=1}^{t} \xi_j]\,\mathbb{E}[\sum_{j=t+1}^{k} \xi_j] \\
&= \left[\mathbb{E}[(\sum_{j=1}^{t} \xi_j)^2] - (\mathbb{E}[\sum_{j=1}^{t} \xi_j])^2\right] + \left[\mathbb{E}[\sum_{j=1}^{t} \xi_j (\sum_{j=t+1}^{k} \xi_j)] - \mathbb{E}[\sum_{j=1}^{t} \xi_j]\,\mathbb{E}[\sum_{j=t+1}^{k} \xi_j]\right] \\
&= Var(\sum_{j=1}^{t} \xi_j) + Cov(\sum_{j=1}^{t} \xi_j, \sum_{j=t+1}^{k} \xi_j) \\
&= t\sigma^2
\end{aligned} \tag{18}$$

The final result of the covariance function is $t\sigma^2$ (since $Cov(\sum_{j=1}^{t} \xi_j, \sum_{j=t+1}^{k} \xi_j) = 0$ due to the independence of the two sums). Thus, the random walk is not stationary because it's variance increases with t.

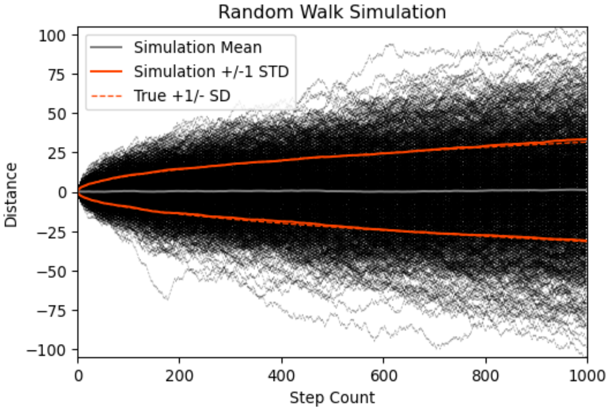
After showing that the random walk is not stationary I wanted to take a quick look at how $Var(\sum_{j=1}^{t} \xi_j)$ itself evolves over time. To do so I first evaluated the variance as follows:

$$\begin{aligned} Var(X_t) = Cov(X_t, X_t) = \mathbb{E}[X_t^2] = \mathbb{E}[(\sum_{j=1}^{t}\xi_j)(\sum_{i=1}^{t}\xi_i)] \\ = \mathbb{E}[\sum_{j=1}^{t}\sum_{i=1}^{t}\xi_j\xi_i] = \sum_{j=1}^{t}\sum_{i=1}^{t}\mathbb{E}[\xi_j\xi_i] \end{aligned} \tag{19}$$

At this point we can consider the possible values of $\mathbb{E}[\xi_j\xi_i]$. When $i \neq j$ we have four possibilities of equal probability ($(1,-1)$, $(-1,1)$, $(-1,-1)$, $(1,1)$), which gives us $\mathbb{E}[\xi_j\xi_i] = 0$ when $i \neq j$. When $i = j$ however, we get that $X_jX_i = 1$, and $\mathbb{E}[X_jX_i] = 1$. Therefore we can ignore the terms where $i \neq j$ in $\sum_{j=1}^{t}\sum_{i=1}^{t}\mathbb{E}[\xi_j\xi_i]$ and since there are n instances where $i = j$ we get $Var(X_t) = \sum_{j=1}^{t}\sum_{i=1}^{t}\mathbb{E}[\xi_j\xi_i] = n = \sigma^2$.

I wanted to check this result with a quick simulation, so I wrote a script in Python to generate random walks and then plotted them along with the simulation mean, ±1 standard deviations, and the theoretical $\sigma^2 = n \implies \sigma = \sqrt{n}$ line to see how well things line up.

Figure 1:



Things look to line up pretty well, so it appears that the experiment is matching the theoretical result well. We can also see the explosion in variance over time that we saw while proving the walk was not stationary.

Problem 2

$$X_t = Acos(\omega t) + Bsin(\omega t),\ t = 0, \pm 1, \pm 2, ... \tag{20}$$

Equation 20 is a process where A & B are uncorrelated standard random normal variables (with mean of 0 and variance of σ^2), and $\omega \in [0, 2\pi)$ is a fixed frequency. We want to show that this process is stationary.

To do so, let's evaluate the autocovariance of the process.

$$
\begin{aligned}
Cov(X_t, X_{t+k}) &= \mathbb{E}[(X_t - \cancelto{0}{\mathbb{E}}[X_t])(X_{t+k} - \cancelto{0}{\mathbb{E}}[X_{t+k}])] \\
&= \mathbb{E}[X_t X_{t+k}] \\
&= \mathbb{E}[(Acos(\omega t) + Bsin(\omega t))(Acos(\omega(t+k)) + Bsin(\omega(t+k)))] \\
&= \mathbb{E}[A^2 cos(\omega t)cos(\omega(t+k)) + B^2 sin(\omega t)sin(\omega(t+k))] \\
&\mathbb{E}[AB](cos(\omega t)Bsin(\omega(t+k)) + sin(\omega t)cos(\omega(t+k))) \\
&= \mathbb{E}[A^2 cos(\omega t)cos(\omega(t+k)) + B^2 sin(\omega t)sin(\omega(t+k))] \\
&\cancelto{0}{\mathbb{E}[A]\mathbb{E}[B]}(cos(\omega t)Bsin(\omega(t+k)) + sin(\omega t)cos(\omega(t+k))) \\
&= \mathbb{E}[A^2]cos(\omega t)cos(\omega(t+k)) + \mathbb{E}[B^2]sin(\omega t)sin(\omega(t+k)) \\
&= \sigma^2(cos(\omega t)cos(\omega(t+k)) + sin(\omega t)sin(\omega(t+k))) \\
&= \sigma^2 cos(\omega k)
\end{aligned}
\tag{21}
$$

Given that the covariance function is only dependent upon the time separation k and not t we have that X_t is weakly stationary.

Problem 3

We want to show that for every DAG there exists a topological ordering of the vertices. To do this we can take an inductive approach.

- Base case: Suppose we have a DAG with a single node v_1. This single node graph has a topological order by default.
- Now assume G is a DAG and v_k is a node with no outward edges. Then $G - \{v_k\}$ is also a DAG (we cannot create new cycles in a DAG by removing a node)
- Assume $G - \{v_k\}$ has a topological ordering
- Then we can create a topological order for G by appending v_k to the end of the topological order of $G - \{v_k\}$
    - Since $G - \{v_k\}$ has a topological ordering $v_1 ... v_{k-1}$ so $v_1 ... v_{k-1} v_k$ becomes and ordering for G, because no edge $v_i v_j$ where $i > j$ exists in the topological ordering for $G - \{v_k\}$ and i cannot be k since we chose v_k to be a node with no outward edges
- Then by induction we have that our DAG G must have a topological ordering

We can also imagine altering the approach above by removing a node v_k that has no incoming edges (rather than no outward), and creating a topological ordering for G by appending v_k to the front of the assumed order for $G - \{v_k\}$.

One question that might arise is how we can ensure the existence of nodes with no outward or inward edges within our DAG. We can show that there exists such nodes by examining any given path P with our DAG since the composition of such paths forms the graph. Let us assume that there exists a node v_k with no incoming edges. To show that such a node exists we will assume that there also exists some edge $< p, v_k >$. Then there is either some node $p \notin P$ that forms and edge $< p, v_k >$, which cannot exist as it would violate the structure of the path P. Alternatively there must be some node $v_i \in P$ that creates the edge $< v_i, v_k >$, but this would create a cycle within our path which is impossible. Therefore, there cannot be an edge $< p, v_k >$ and v_k is in fact a node with no incoming edges. Similarly we can show that there exist nodes with no outward edges by reversing the direction of the edges of our graph and then inspecting our new path P' since we must still have a no with no incoming edges and such nodes are the same nodes with no outgoing edges in the original path P.

It is also interesting to note that we can strengthen our statement that every DAG has a topological ordering by observing that a topological ordering cannot have a cycle since no ordering $v_1 ... v_i ... v_j ... v_l ... v_k$ can have a cycle $v_i < ... < v_j < ... < v_l < v_i$. Therefore it must take the form of a DAG.