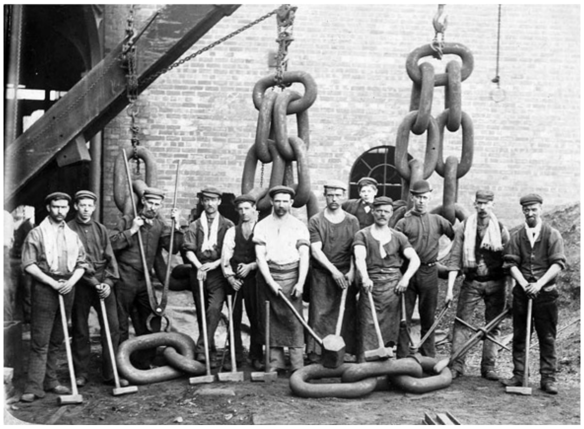
University of Illinois
Spring 2017 | Econ 574 | Department of Economics
Roger Koenker

Lecture 20
A Not Too Random Walk in Markov Chains



We will begin by considering the random walk

$$\begin{aligned} X_i &= X_{i-1} + \xi_t \\ \text{or} \qquad & \\ X_i &= X_0 + \xi_1 + \xi_2 + \ldots + \xi_t \end{aligned}$$

where ξ_t are iid random variables with probability of $\xi_t = 1$ of p and probability of $\xi_t = -1$ of $q = 1 - p$.

Gambler's Ruin: Suppose we have R and G flipping $(p-q)$ coins. R has \$$a$ initially, G has \$$b$. They flip until one is bankrupt. What is the probability that R is ruined?

Let $u_j =$ Prob(X_n hits 0, before it hits $c = a + b$ when it starts from j)
we really need just u_a, but all the other u_j's are necessary intermediate products. Now we have the difference equation,

$$(*) \qquad u_j = pu_{j+1} + qu_{j-1} \quad 1 \leq j \leq c-1$$

with boundary conditions $u_0 = 1$ and $u_c = 0$.

To see (*), note that if R is at j and the probability of getting to $j+1$ is p, and at that point the probability of ruin is u_{j+1}, while with probability q he gets to $j-1$ and then has ruin probability u_{j-1}. The result now follows by application of the rule,

$$P(B) = \sum_i P(A_i)P(B|A_i)$$

since the only way to get to state j is via state $j-1$ or $j+1$.

Now, using $p + q = 1$, write $(*)$ as,

$$q(u_j - u_{j-1}) = p(u_{j+1} - u_j)$$

or

$$d_j = rd_{j-1}$$

where $r = q/p, d_j = u_j - u_{j+1}$. Thus

$$d_j = r^j d_0$$

and summing by parts, (remember summing by parts?),

$$(+) \qquad 1 = u_0 - u_c = \sum_{j=0}^{c-1}(u_j - u_{j+1}) = \sum d_j = \sum r^j d_0 = \frac{1-r^c}{1-r}d_0$$

and similarly, we have

$$u_j = u_j - u_c = \sum_{i=j}^{c-1}(u_i - u_{i+1}) = \sum_{i=j}^{c-1} r^i d_0 = \frac{r^j - r^c}{1-r}d_0$$

so, using (+)

$$(\times) \qquad u_j = \frac{r^j - r^c}{1-r^c} \quad 0 \le j \le c$$

Now, reversing the roles of R and G, flip p and q and j to $c-j$, then for v_j we have

$$v_j = \frac{1/r^{c-j} - 1/r^c}{1 - 1/r^c} = \frac{(1/r)^c(r^j - 1)}{1-(1/r)^c} = \frac{1-r^j}{1-r^c}$$

where we see that $u_j + v_j = 1$. This solves our problem, but also shows that there is no probability that X bounces forever between 0 and c, however big c is! This is an important lesson, since it shows that everything that is possible will (eventually) happen. For $r = 1$, i.e. $p = q$, we get by L'Hôpital and (+)

$$\lim_{r\to 1}\frac{1-r^c}{1-r} = \left.\frac{-cr^{c-1}}{-1}\right|_{r=1} = c \quad \Rightarrow \quad 1 = cd_0$$

and from $(\times)$, we have the intuitive result that,

$$u_j = \frac{j-c}{-c} = \frac{c-j}{c} \Rightarrow u_a = \frac{b}{c}$$

This property of the random walk can be extended from its discrete setting to continuous time. Let δ be the new unit of time, now say $P(\xi_k = \sqrt{\delta}) = P(\xi_k = -\sqrt{\delta}) = \frac{1}{2}$ so

$$E\xi_k = 0, \ \sigma^2(\xi_k) = \frac{1}{2}(\sqrt{\delta})^2 + \frac{1}{2}(-\sqrt{\delta})^2 = \delta$$

suppose $X_0 = 0$ so $X_t = \sum_{k=1}^{t/\delta}\xi_k$. When t is fixed and $\delta \to 0$ we have by the DeMoivre-Laplace CLT, $X_t \sim \mathcal{N}(0,t)$. This yields a (very naive) construction of Brownian motion.

Definition (Brownian Motion): A family of random variables indexed by a continuous variable t over $[0.\infty)$ is a Brownian Motion iff

(i) $X(0) = 0$

(ii) $\{X(s_i + t_i) - X(s_i)\}$ over an arbitrary collection of disjoint intervals, $(s_i, s_i + t_i)$ are $\perp\!\!\!\perp$ r.v.s

(iii) for each $s \ge 0, t \ge 0, X(s+t) - X(s) \sim \mathcal{N}(0,t)$

Establishing the existence and properties of such processes is one of the major accomplishments of 20th century mathematics. The foregoing sketch of random walks and their continuous analogue yielding Brownian motion can be generalized dramatically.

Markov Chains

Consider a sequence of random variables $\{X_t\}$ taking values in a set $\mathcal{S}$ to be called the state space. (We will assume provisionally that $\mathcal{S}$ is finite.) Our fundamental assumption is that if X_t is in state i at any time, *regardless of where it has been before*, the probability that it will be in state j, in period $t+1$ is given by p_{ij}, that is,

$$(\diamond) \qquad P(X_{t+1} = j | X_t = i, A) = P(X_{t+1} = j | X_t = i) = p_{ij}$$

for any conditioning event A determined by prior history of X, $\{X_0, \ldots, X_{t-1}\}$. This is the Markov property, the second equality in $(\diamond)$ insists that the process has stationary, or temporally homogeneous, transition probabilities.

It is often convenient to think of a random initial state, so $P(X_0 = i) = p_i$ Where $\{p_i, i \in \mathcal{S}\}$ is the initial distribution. For any p_{ij} satisfying, for all $i, j \in \mathcal{S}$: (i) $p_{ij} \geq 0$ and (ii) $\sum_{j \in \mathcal{S}} p_{ij} = 1$, we have a homogeneous Markov chain.

It proves to be convenient to organize our transition probabilities into matrix form $\Pi = (p_{ij})$. Note that if we would like to compute the probability of a transition from state i to state j in exactly n periods, i.e., $P(X_n = j | X_0 = i) \equiv p_{ij}^{(n)}$, then we have

$$p_{ik}^{(n)} = \sum_j p_{ij} p_{jk}^{(n-1)} = \sum p_{ij}^{(n-1)} p_{jk}$$

e.g. $\Pi^2 = \Pi \ \Pi = (p_{ij}^{(2)})$ so we are just doing matrix multiplication and thus $\Pi^n = (p_{ij}^{(n)})$ and $\Pi^{n+m} = \Pi^n \cdot \Pi^m$.

Ex 1 Random Walk without boundary has

$$\Pi = \begin{bmatrix} 0 & \ddots & & & & & 0 \\ \ddots & \ddots & 0 & p & & & \\ \ddots & q & & & p & & \\ & & q & & & & \ddots \\ & & & \ddots & \ddots & & p \\ 0 & & & & & q & 0 \end{bmatrix}$$

at each stage there is a transition to only the adjacent state.

Ex 2

$$\Pi = \begin{bmatrix} 1 & 0 & \cdots & & & 0 \\ q & 0 & p & \cdots & & \\ 0 & q & 0 & p & \cdots & 0 \\ & \ddots & 0 & q & 0 & p \\ 0 & \cdots & & & 0 & 1 \end{bmatrix}$$

Here there are absorbing states at the ruin points. The matrix is $(c+1)$ by $c+1$.

Ex 3

$$\Pi = \begin{bmatrix} 0 & 1 & & & 0 \\ q & 0 & p & \cdots & \\ & & \ddots & & \\ & & q & 0 & p \\ 0 & \cdots & & 1 & 0 \end{bmatrix}$$

Here the ruined guy gets one dollar to restart his luck from his opponent. This is the "reflecting barrier" as opposed to the absorbing barrier case in Ex 2.

Ex 4 Extending the integer valued model further, suppose

$$X_t = X_{t-1} + \xi_t$$

Where

$$P(\xi_t = j) = a_j \quad \text{with } \xi_t \perp\!\!\!\perp X_{t-1}$$

So Ex 1 is a special case of Ex 4. "Baby" Chung offers further, more sophisticated, examples.

Structure of Markov Chains

Now we introduce some more formal terminology,

Definition: $i \rightsquigarrow j$(i leads to j) iff $\exists\ n \geq 1\ p_{ij}^{(n)} > 0$

Definition: $i \leftrightsquigarrow j$(i communicates with j) iff $i \rightsquigarrow j$ and $j \rightsquigarrow i$

The first occurrence time, or first passage time,

$$T_j = \min\{t \geq 1 | X_t = j\}$$

is an important quantity. We would like to characterize two probabilities

$$f_{ij}^{(n)} = P(T_j = n | X_0 = i) \quad \text{and} \quad f_{ij}^{*} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$

Thus, the probability that X_t gets to state j only *after* hell freezes over is

$$f_{ij}^{\infty} = 1 - f_{ij}^{*}$$

We can write more explicitly, $f_{ij}^{(1)} = p_{ij}$, and we would like to be more explicit about

$$f_{ij}^{(n)} = P(X_{t+s} \neq j,\ 1 \leq s \leq n-1,\ X_{t+n} = j | X_t = i)$$

by homogeneity we don't need to worry about when we start, the initial time t is irrelevant.

Theorem: For any i and j and $1 \leq n < \infty$,

$$p_{ij}^{(n)} = \sum_{s=1}^{n} f_{ij}^{(s)} p_{jj}^{(n-s)}$$

Proof:

$$\begin{aligned}
p_{ij}^{(n)} &= P_i(X_n = j) \\
&= P_i(T_j \leq n, X_n = j) \qquad [X_n = j \Rightarrow T_j \leq n] \\
&= \sum_{s=1}^{n} P_i(T_j = s, X_n = j) \quad [X_n = j \text{ events disjoint}] \\
&= \sum_{s=1}^{n} P_i(T_j = s) P_i(X_n = j | X_1 \neq j, \ldots, X_{s-1} \neq 1, X_s = j) \quad [\text{Conditional} P_i] \\
&= \sum_{s=1}^{n} P_i(T_j = s) P(X_n = j | X_s = j) \quad [\text{Markov Prop}] \\
&= \sum_{s=1}^{n} P_i(T_j = s) P_i(X_{n-s} = j) \quad [\text{Temporal Homogeneity}] \\
&= \sum_{s=1}^{n} f_{ij}^{(s)} p_{jj}^{(n-s)} \qquad \blacksquare
\end{aligned}$$

Computation is facilitated with help of generating functions as suggested in L1.

Let

$$\begin{aligned} P_{ij}(z) &= \sum_{k=0}^{\infty} p_{ij}^{(k)} z^k \\ F_{ij}(z) &= \sum_{k=1}^{\infty} f_{ij}^{(k)} z^k \end{aligned}$$

for $|z| < 1$. Then, using the prior Theorem, with δ_{ij} as Kronecker's δ,

$$\begin{aligned} P_{ij}(z) &= \delta_{ij} + \sum_{n=0}^{\infty} \sum_{s=1}^{n} f_{ij}^{(s)} p_{jj}^{(n-s)} z^s z^{n-s} \\ &= \delta_{ij} + \sum_{s=1}^{\infty} f_{ij}^{(s)} z^s \sum_{n=0}^{\infty} p_{jj}^{(n-s)} z^{n-s} \\ &= \delta_{ij} + F_{ij}(z) P_{jj}(z) \end{aligned}$$

The inversion of the order of summation is justified because both series are absolutely convergent for $|z| < 1$.

Theorem: For any state i we have $f_{ii}^* = 1$, iff

$$\sum_{k=0}^{\infty} p_{ii}^k = \infty$$

if $f_{ii}^* < 1$, $\sum_{k=0}^{\infty} p_{ii}^k = 1/(1 - f_{ii}^*)$.

Proof: Set $i = j$ in the foregoing argument and solve for $P_{ii}(z)$,

$$P_{ii}(z) = 1/(1 - F_{ii}(z))$$

Then put $z = 1$ and use $P_{ii}(1) = \sum_{k=0}^{\infty} p_{ii}^k$, $F_{ii}(1) = f_{ii}^*$ ■

Remark: This is an Abelian theorem in the terminology of Feller, and it motivates the following definition:

Definition: A state i is recurrent iff $f_{ii}^* = 1$, and non-recurrent iff $f_{ii}^* < 1$.

Corollary: If j is non-recurrent, then $\sum_{k=0}^{\infty} p_{ij}^{(k)} < \infty$ for every i, and thus $\lim_{k\to\infty} p_{ij}^{(k)} = 0$ for all i.

Proof: For $i = j$ this is the prior theorem, if $i \neq j$, then

$$P_{ij}(1) = F_{ij}(1) P_{jj}(1) \leq P_{jj}(1) < \infty$$

Remark: Any pair of states must either be both recurrent or both non-recurrent. But one can't go from recurrent to non-recurrent states.

For recurrent Markov chains it is important to understand the limiting behavior of averages:

$$n^{-1} \sum_{k} p_{ij}^{(k)}$$

Focusing first on $i = j$, let

$$m_{jj} = E_j T_j = \sum_{k=1}^{\infty} k f_{jj}^{(k)}$$

the mean time required to return to j from a start at j. We would like to relate m_{jj} to the former average, which can be interpreted as average expected occupation time in state j.

Since on average it requires m_{jj} time units to return to i from j, there should be, on average, about n/m_{jj} time units spent in state j during n periods, i.e.,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} p_{ij}^{k} = \frac{1}{m_{jj}}$$

This can be formalized by considering iid sequences, but this is tough. A simple approach is provided by the following discussion.

But the stationary distribution of a recurrent Markov chain is easily found given the matrix $\Pi = (p_{ij})$ assuming that we have a single recurrent class. If one has a vector of probabilities p_0 describing the distribution of states in an initial period, then of course,

$$p_1 = \Pi p_0$$

but in a stationary situation we would have

$$(\star) \qquad p(I - \Pi) = 0$$

Does this have a solution? Note that the columns of the matrix $I - \Pi$ are all summing to $1 = \sum_j p_{ij}$, so $(\star)$ has a nontrivial solution. Such a solution is free to be normalized as $\sum p_i^* = 1$. A simple way to compute the solution to $(\star)$ is to discard one of the equations, say the first one, and solve for the other p_i's in terms of p_1. Details available from Bellman (1970).

Theorem: If A is a *positive* Markov matrix and $x_t = Ax_{t-1}$, then $\lim x_t = y$, y is $\perp\!\!\!\perp$ of x_0, and y is an eigenvector of A with associated eigenvalue 1.

Proof: Consider $x_n'b = (A^n x_0, b) = x_0'(A^n)'b = x_0(A')^n b$. Let $z_n = (A')^n b$ so $z_{n+1} = A'z_n$, with $z_0 = b$. Now, let $u_n = \max_i z_n$ and $v_n = \min_i z_n$, we will show $u_n - v_n \to 0$ as $n \to \infty$. Note that

$$z_{n+1}^{i} = \sum_{j=1}^{J} A_{ji} z_n^{j}$$

and $\sum_{j=1}^{J} A_{ji} = 1$, $A_{ji} \geq 0$ so, $u_{n+1} \leq u_n$ and $v_{n+1} \geq v_n$, so we have monotone sequences bounded from below by 0 (u_n) and from above by 1 (v_n) so we have convergence to say u and v respectively. But more can be said, in fact, $u_n - v_n \to 0$ so z converges to a constant vector. To see this, consider,

$$z_{n+1}^{i} = \sum m_{ji} z_n^{j}$$

we want an upper bound so if we have $m_{ji} \geq \delta$ recall A is positive (!), then

$$u_{n+1} \leq (1-\delta)u_n + \delta v_n$$

(this puts minimal weight on v_n and *all* the rest of the weight on u_n.) And similarly,

$$v_{n+1} \geq \delta u_n + (1-\delta)v_n$$

then,

$$u_{n+1} - v_{n+1} \leq (1-2\delta)(u_n - v_n)$$

and thus,

$$u_n - v_n \leq (1-2\delta)^n(u_0 - v_0) \to 0,$$

so the min and the max converge to the same value so $z_n \to z$ and hence $x \to y$. What can we say about this y? If $z = \zeta 1_n$ has identical elements then

$$(y, b) = \lim(x_n, b) = (x_0, z) = \zeta 1' x_0 = \zeta$$

and hence is $\perp\!\!\!\perp$ of x_0. And,

$$y = \lim A^{n+1} x_0 = A \lim A^n x_0 = Ay$$

so y is eigenvector with unit eigenvalue. ■

MCMC in Action

We have seen that under some fairly general conditions, Markov chains have unique stationary distribution, with

$$\lim_{n \to \infty} P(X_n = j) = P_j \quad \forall j$$

independent of the initial state p_0. There are many further details: egodicity, speed of convergence, etc. that we won't delve into. Instead, we now turn to more practical considerations concerning how transition matrices can be constructed to simulate Markov chains. We will make a leap of faith into the realm of continuous state space chains for which we may wish to consider some sort of discrete approximation, but we will leave all details of this to other sources. A recent detailed treatment of many practical issues can be found in Jackman (2009). Lancaster (2005) provides a very readable, more econometric, treatment.

The Gibbs Sampler

The simplest schema for constructing a Markov chain sampler is to focus on sequential conditionals. To illustrate this consider the transition equation,

$$p(y) = \int K(x, y) p(x) dx$$

but now partition x and y into two pieces $x = (x_1, x_2), y = (y_1, y_2)$. To define the Markov chain, we assume that we can draw from theth two conditionals:

1) Draw Y_1 from $P_{Y_1|Y_2}(y_1|x_2)$
2) Draw Y_2 from $P_{Y_2|Y_1}(y_2|y_1)$

Repeating 1) and 2) allows us to simulate from the stationary distribution of the chain. Write

$$\begin{aligned} K(x, y) &= p_{Y_1|Y_2}(y_1|x_2) p_{Y_2|Y_1}(y_2|y_1) \\ &\equiv p_{12} p_{21}. \end{aligned}$$

We can regard successive draws from the conditionals as represented by the integral operator with kernel K, that is starting from an initial state drawn from the stationary distribution, successive draws from the conditionals reproduces this stationary distribution:

$$\begin{aligned} p(y) &= \int K(x, y) p(x) dx \\ &= \int p_{12} p_{21} p(x) dx \\ &= \int p_{12} p_{21} p_{Y_2}(x_2) dx_2 \\ &= p_{21} \int p_{Y_1, Y_2}(y_1, x_2) dx_2 \\ &= p_{Y_2|Y_1}(y_2|y_1) p_{Y_1}(y_1) \\ &= p(y) \end{aligned}$$

A nice example is latent variable normal theory models:

$$y_1 \text{ is } y \qquad \text{the latent variable}$$

$$y_2 \text{ is } \beta \qquad \text{the model parameter}$$

The conditional of $y_1|y_2$ is $y|\beta \sim \mathcal{N}(X\beta, \sigma^2 I)$ truncated to the left at zero if $y_i = 1$ and to the right at zero if $y = 0$ $y_2|y_1$ is $\beta|y \sim \mathcal{N}(\hat{\beta}, \sigma^2(X'X)^{-1})$.

This approach is unfortunately limited to cases where conditionals are convenient like this. One should be careful about this since it is crucial that the conditionals be right: arbitrary conditionals typical do not define a coherent joint distribution. We leave aside questions about how long a chain is needed and whether to remove initial realizations (burn in) – topics which are still near the research frontier. Better diagnostics for evaluating whether we have entered the stationary region of the simulation of the chain would be very worthwhile.

Metropolis Algorithm

This is more like rejection method considered earlier in the course, suppose we have the "proposal distribution", $q(y|x)$

Algorithm

```
Initialize $y = y_0, t = 0$

Draw $y$ from $q(y|y_t)$

Compute $r = p(y)/p(y_t)$

    if $r \geq 1$, set $y_{t+1} = y$

    else, $y_{t+1} = \begin{cases} y & wp \ r \\ y_t & wp \ 1-r \end{cases}$
```

Repeat. It may seem almost self contradictory that we are looking for the stationary distribution p but seem to be using it as if it were known to compute r above. A crucial feature of the above algorithm, however, is that we don't need everything about p to compute r. Consider exponential family models in which we have a sufficient statistic piece to the likelihood and then another multiplicative piece of the likelihood that depends on the parameters in some potentially very complicated way. (Logspline estimation provides a good example of this.) Since r depends only on the ratio this multiplicative factor cancels and this enables us to use only the "kernel" of p without needing the other component.

Theorem: The Metropolis sampler has stationary distribution p

Proof: Suppose, for the moment, that K satisfies the so-called "detailed balance" condition:

$$(*) \qquad K(x,y)p(x) = K(y,x)p(y) \quad \forall x, y$$

In a finite state Markov chain this would say that the probability of transit from state x to state y, is the same as the probability of transit from y to x. Then, for any set B,

$$\int K(y,B)p(y)dy \equiv \int \int_B K(y,x)p(y)dxdy$$

$$\int \int K(x,y)p(x)dxdy$$

$$\int_B p(x)dx$$

The last step uses $\int K(x,y)dy = 1$. Now to show that Metropolis kernel satisfies (*) we observe that,

$$K(x,y) = \rho(x,y)q(y|x) + (1+r(x))\delta_x(y)$$

where $\rho(x,y) = \min\left\{\frac{p(y)}{p(x)}, 1\right\}, r(x) = \int \rho(x,y)q(y|x)dx$.

Note $q(y|x)$ is Prob y is "produced", $\rho(x,y)$ is prob it is "accepted". And $r(x)$ sums these over y so $1 - r(x)$ is the probability we stay at x. Now multiplying by $p(x)$ to get

$$p(x)K(x,y) = p(x)\rho q + p(x)(1+r(x))\delta_x(y)$$

and similarly by $p(y)$,

$$p(y)K(y,x) = p(y)\rho q + p(y)(1+r(y))\delta_y(x). \qquad \blacksquare$$

Metropolis-Hastings

Often a modification of Metropolis is actually employed: rather than using

$$r(y, y_t) = p(y)/p(y_t)$$

instead we use,

$$r(y, y_t) = \frac{p(y)q(y_t|y)}{p(y_t)q(y|y_t)}$$

and the acceptance probability is no longer ρ above, but rather,

$$\rho(y, y_t) = \min(r, 1)$$

If $q(x,y)$ is symmetric, $q(x,y) = q(y,x)$, then we are back to Metropolis, but if not, not. Some examples of these methods are given by Lancaster (2005) and Koenker and Yoon (2007), and further theory is laid out in Robert and Casella (1999).

References

Bellman, R. (1970) *Introduction to Matrix Algebra*, McGraw-Hill.

Chung, K.L. (1979) *Elementary Probability Theory,* Springer.

Jackman, S. (2009) *Bayesian Analysis for the Social Sciences,* Wiley.

Koenker, R. and J. Yoon (2007) "Parametric Links for Binary Response Models," *J. of Econometrics*, forthcoming.

Lancaster, T. (2005) *An Introduction to Modern Bayesian Econometrics*," Blackwell.

Robert, C. and G. Casella (1999) *Monte Carlo Statistical Methods*, Springer.