Baseball Analytics

Readings and Research

A Markov Chain Approach to Optimizing Batting Lineups

James DeAntonis, Professor Dan Chambers

Fall 2016

Contents

1 Groundwork

The discrete nature of events within a baseball game make it a *nice* sport. It turns out that there are only a finite number of possible scenarios when any given batter steps up the the plate; a certain combination of bases occupied and outs. We outline and number the possible non-inning-ending states below.

		Runners							
		None	**1**	**2**	**3**	**1&2**	**1&3**	**2&3**	**1&2&3**
Outs	**0**	1	4	7	10	13	16	19	22
	1	2	5	8	11	14	17	20	23
	2	3	6	9	12	15	18	21	24

Here, state 1 is the scenario where there are no outs and no runners on base, state 14 is the scenario where there are runners on first and second with one out, and so forth.

In order to bookend a given inning, we add a few more states to include absorbing, inning-ending events, partitioned by number of runners left on base on the corresponding play. While it is currently unclear why this distinction of runners left on base is necessary, it will become clear in future sections.

Outs	**3**	*0 Runners Left*	25
	3	*1 Runner Left*	26
	3	*2 Runners Left*	27
	3	*3 Runners Left*	28

Given the characteristics of a batter i, the goal becomes modelling the probabilities of leaving one state for another as a result of this batter's at-bat. This can be done using a 28x28 stochastic transition matrix, where the i, jth entry of the matrix is $\rho_{i,j}$, the probability of a batter's at-bat resulting in a transition to state j, given he approached the plate with the inning in state i. We have that

$$T_i = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,28} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,28} \\ \vdots & \vdots & & \vdots \\ \rho_{28,1} & \rho_{28,2} & \cdots & \rho_{28,28} \end{bmatrix}$$

where T_i is the ith batter's transition matrix.

It is important to think intuitively when considering the behavior of this stochastic matrix. For example, $\rho_{1,7}$ in this matrix is the probability of batter i stepping up to the plate with no runners on base and no outs, and, as a result of his at-bat, there being no outs and a runner on second. In other words, $\rho_{1,7}$ is the probability of batter i hitting a double, hitting a single and reaching second as the result of an overthrow, or any other event which results in this outcome.

Now, consider the i, jth entry of matrix T_1T_2. This is the dot product of row i in T_1 with column j in T_2. An interpretation of this dot product is the probability of batter 2's at-bat resulting in a transition to state j while batter 1 stepped up to the plate in state i. For example, suppose batter 1 led off the inning, which is automatically state 1 (although a lead-off at-bat is not the only way for a batter to bat in state 1. Additionally, batter 1 is not the only batter

that may lead off any given inning). Entry $(1, 13)$ is the probability that, through these first two batters, there are runners on 1st and 2nd with no outs. This is a desirable outcome, because batter 3 has a chance to transition from state 13 to, say, state 22, loading the bases for the cleanup batter. By traditional convention, managers choose batter 4 to be a batter with a relatively high $\rho_{22,1}$, meaing this batter is capable of hitting home runs. This convention, among others, can be supported or rejected as a result of this analysis.

Similarly, notice that the i, jth entry in $T_k \cdot T_{k+1} \cdots T_\ell$ is the probability of transitioning from state i to state j as a result of the at-bats of batters k through ℓ.

Finally, we consider states 25-28 to be absorbing. Specifically, $\rho_{25,25} = \rho_{26,26} = \rho_{27,27} = \rho_{28,28} = 1$.

Baseball Conservation Law The Baseball Conservation Law states that every batter that comes to the plate either makes an out, scores, or is left on base. Specifically, for one inning,

$B = 3 + R + L$

and for a whole game,

$B = 27 + R + L \implies R = B - 27 - L$

where B = number of batters to bat in the entire game, R = number of runs scored, and L = number of runners left on base. We will model the probability distribution of runs scored based on this equation.

2 Extending a Transition Matrix to Nine Innings

Extend a transition matrix to 9 innings.

$$P_i = \begin{bmatrix} A_i & B_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_i & B_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_i & B_i & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_i & B_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_i & B_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_i & B_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_i & B_i & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_i & B_i \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_i \end{bmatrix}$$

$$\text{where } A_i = \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,28} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,28} \\ \vdots & \vdots & & \vdots \\ \rho_{24,1} & \rho_{24,2} & \cdots & \rho_{24,28} \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix} \text{ and } B_i = \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ \rho_{1,1} & \cdots & \rho_{1,28} \\ \rho_{1,1} & \cdots & \rho_{1,28} \\ \rho_{1,1} & \cdots & \rho_{1,28} \\ \rho_{1,1} & \cdots & \rho_{1,28} \end{bmatrix}$$

To illustrate why this is the desired composite matrix, we consider a sample path. P_1P_2 behaves similarly to T_1T_2, but with a slight complication. Notice that A_i is identical to T_i, but with zeroes in rows 25-28. Before the third out states, any state within any A_i behaves identically to T_i. But, upon entry into a third out state, the following transition probabilities are the rows

of the 25th-28th rows of B_i, which are each identical to the state 1 probabilities. But, this time, these transition probabilities send the system to state j in the A_i in the subsequent row of P_i. This behavior simulates a transition from one inning to the next. To see this, suppose batter i transitions into one of the three out states. Notice that, in T, batter $i+1$ remains in the three out state with probability one. In this block matrix P, batter $i+1$ has state one transition probabilities, but sending the system to the next A, as if batter $i+1$ is the first batter in the next inning. This is the reasoning for P_i having nine such A_i.

But, notice that P_i, in fact, is not a stochastic matrix. The elements of each row sum to 1, except for the final four rows in the matrix, whose rows sum to 0. A question remains: how do we track the ending of a game? Remember that it would be nice to know how many batters there were in the entire game, because, by the Baseball Conservation Law, we would be one step closer to modeling the number of runs scored in the entire game. We could choose to set the final A_i equal to T_i, where the final four states are absorbing. But, with this setup, it is inconvenient to determine the arrival time into this state, interpreted as the end of the game. Instead, it is best to add two more final states; one which is transitioned to, with probability one, by the final four states in the last A_i, and the second state being transitioned to with probability one by the aforementioned state with probability one. This final state will be absorbing. The catch here is that the second-to-last state is achieved once and only once, always by the "batter" who follows the batter who makes the final out in the game (notice that this "at-bat" has no interpretation, but is just a tool for acknowledging that the previous batter made the final out). Thus, the 40th batter being in this second-to-last state means that 39 batters batted in the entire game. This matrix, Q_i, is illustrated below.

$$Q_i = \left[\begin{array}{ccccccccc|c} A_i & B_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_i & B_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_i & B_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_i & B_i & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_i & B_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_i & B_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_i & B_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_i & B_i & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_i & D \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & E \end{array}\right]$$

where $D = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$ and $E = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$.

We now have a 9(28)+2=254-state stochastic matrix modeling an entire nine-inning baseball game. In this model, the probability that there are exactly 40 batters in a game is the (1,253) entry in the matrix $\prod_{i=1}^{41} Q_i$. Again, we have a product of 41 transition matrices because state 253 is attained one batter $after$ the batter who commits the final out of the game.

3 Constructing a Runs Scored Distribution

Recall the Baseball Conservation Law: $R = B - 27 - L$. Based on this equality, tracking the runners left on base is the remaining key to modeling and calculating probabilities of runs scored in a game. Notice that the third out states, 25-28, are defined by the runners remaining on base at the end of the inning. Hence, it becomes necessary to track, after each inning, which third out state was passed. So, we modify our B_i matrices.

$$B_i = \begin{bmatrix} 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ \rho_{1,1} & \cdots & \rho_{1,28} \\ \rho_{1,1}x & \cdots & \rho_{1,28}x \\ \rho_{1,1}x^2 & \cdots & \rho_{1,28}x^2 \\ \rho_{1,1}x^3 & \cdots & \rho_{1,28}x^3 \end{bmatrix}$$

These x's serve the purpose of creating a footprint for the number of runners left on base at the end of an inning. To illustrate the effect of this change on the final probabilities, reconsider the (1,253) entry of the matrix $\prod_{i=1}^{41} P_i$ as previously discussed. Now, instead of this entry being a single probability, this will be a polynomial whose coefficients add up to that probability. The significance of these coefficients is that the coefficient of, say, x^3 in this polynomial is the probability that, not only were there 40 total batters in the entire game, but there were also 3 runners stranded on base throughout the game.

Notice that D must also change, because it tracks the number of runs left on base in the ninth inning. Now,

$$D = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 1 & 0 \\ x & 0 \\ x^2 & 0 \\ x^3 & 0 \end{bmatrix}.$$

Next, consider the probability of scoring 3 runs in a game: this is equal to the probability that there were 30 batters in the entire game and zero runners left on base, plus the probability that there were 31 batters in the game and one total runner left on base, and so forth. These probabilities are now a summation of coefficients. In an equation, these coefficients are extracted through differentiation as follows:

$$\mathbb{P}(\text{score } z \text{ runs}) = \sum_{k=27}^{\infty} \left[\frac{d^{k-27-z-1}}{dx^{k-27-z-1}} \left(\prod_{i=1}^{k} Q_{(i-1)[\bmod 9]+1} \right)_{(1,253)} \left(\frac{1}{(k-27-z-1)!} \right) \Bigg|_{x=0} \right]$$

4 Breakdown of $Q_1 \cdots Q_n$

While this is a valid construction, 254x254 matrices are not ideal. Our next goal is to express our Q_i multiplication in terms of A's and B's, 28x28 matrices.

First, what exactly is a product of P_i's?

Proposition 1. For $n \geq 9$, $P_1 \cdots P_n$ is of the following form:

$$P_1 \cdots P_n = \begin{bmatrix} (n,0) & (n-1,1) & \cdots & & & & & \cdots & (n-8,8) \\ 0 & (n,0) & \cdots & & & & & \cdots & (n-7,7) \\ 0 & 0 & (n,0) & \cdots & & & & \cdots & (n-6,6) \\ 0 & 0 & 0 & (n,0) & \cdots & & & \cdots & (n-5,5) \\ 0 & 0 & 0 & 0 & (n,0) & \cdots & & \cdots & (n-4,4) \\ 0 & 0 & 0 & 0 & 0 & (n,0) & \cdots & \cdots & (n-3,3) \\ 0 & 0 & 0 & 0 & 0 & 0 & (n,0) & \cdots & (n-2,2) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & (n,0) & (n-1,1) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & (n,0) \end{bmatrix}$$

where $(n-a,a)$ is the sum of all combinations of products of $n-a$ A_i's and a B_i's.

Skeleton of Proof 1. We use induction.

Base case: $n = 9$.

$$P_1 \cdots P_9 = \begin{bmatrix} A_1 & B_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_1 & B_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_1 & B_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_1 & B_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_1 & B_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_1 & B_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_1 & B_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_1 & B_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_1 \end{bmatrix} \cdots \begin{bmatrix} A_9 & B_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_9 & B_9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_9 & B_9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_9 & B_9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_9 & B_9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_9 & B_9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_9 & B_9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_9 & B_9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_9 \end{bmatrix}$$

$$= \begin{bmatrix} (n,0) & (n-1,1) & \cdots & & & & & \cdots & (n-8,8) \\ 0 & (n,0) & \cdots & & & & & \cdots & (n-7,7) \\ 0 & 0 & (n,0) & \cdots & & & & \cdots & (n-6,6) \\ 0 & 0 & 0 & (n,0) & \cdots & & & \cdots & (n-5,5) \\ 0 & 0 & 0 & 0 & (n,0) & \cdots & & \cdots & (n-4,4) \\ 0 & 0 & 0 & 0 & 0 & (n,0) & \cdots & \cdots & (n-3,3) \\ 0 & 0 & 0 & 0 & 0 & 0 & (n,0) & \cdots & (n-2,2) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & (n,0) & (n-1,1) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & (n,0) \end{bmatrix} \checkmark$$

Induction step: suppose this statement is true for $n = k \geq 9$. Then,

$$P_1 \cdots P_{k+1} = (P_1 \cdots P_k) P_{k+1}$$

$$= \begin{bmatrix} (k,0) & (k-1,1) & \cdots & & & & & \cdots & (k-8,8) \\ 0 & (k,0) & \cdots & & & & & \cdots & (k-7,7) \\ 0 & 0 & (k,0) & \cdots & & & & \cdots & (k-6,6) \\ 0 & 0 & 0 & (k,0) & \cdots & & & \cdots & (k-5,5) \\ 0 & 0 & 0 & 0 & (k,0) & \cdots & & \cdots & (k-4,4) \\ 0 & 0 & 0 & 0 & 0 & (k,0) & \cdots & \cdots & (k-3,3) \\ 0 & 0 & 0 & 0 & 0 & 0 & (k,0) & \cdots & (k-2,2) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & (k,0) & (k-1,1) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & (k,0) \end{bmatrix}$$

$$\cdot \begin{bmatrix} A_1 & B_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_1 & B_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_1 & B_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_1 & B_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & A_1 & B_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & A_1 & B_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_1 & B_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_1 & B_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_1 \end{bmatrix}$$

$$= \begin{bmatrix} (k+1,0) & \cdots & & & & & & \cdots & ((k+1)-8,8) \\ 0 & (k+1,0) & \cdots & & & & & \cdots & ((k+1)-7,7) \\ 0 & 0 & (k+1,0) & \cdots & & & & \cdots & ((k+1)-6,6) \\ 0 & 0 & 0 & (k+1,0) & \cdots & & & \cdots & ((k+1)-5,5) \\ 0 & 0 & 0 & 0 & (k+1,0) & \cdots & & \cdots & ((k+1)-4,4) \\ 0 & 0 & 0 & 0 & 0 & (k+1,0) & \cdots & \cdots & ((k+1)-3,3) \\ 0 & 0 & 0 & 0 & 0 & 0 & (k+1,0) & \cdots & ((k+1)-2,2) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & (k+1,0) & ((k+1)-1,1) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & (k+1,0) \end{bmatrix}$$

Recall that (1,253) is the critical entry in $Q_1 \cdots Q_n$. But, notice that

$$(Q_1 \cdots Q_n)[1,253] = \sum_{k=28\cdot 9-3}^{28\cdot 9} (P_1 \cdots P_{n-1})[1,k].$$

Visually, this is the sum of the last four entries in the first row of the block in the top right corner of $P_1 \cdots P_{n-1}$. That is, the matrix generated as the sum of all combinations of $n-8$ A_i's and 8 B_i's. Call this resulting matrix M_n. We want

$$\sum_{k=25}^{28} M_n[1,k]$$

Then,

$$\mathbb{P}(\text{score } z \text{ runs}) = \sum_{k=27}^{\infty} \left[\frac{d^{k-27-z-1}}{dx^{k-27-z-1}} \left(\sum_{i=25}^{28} M_k[1,i] \right)_{(1,253)} \left(\frac{1}{(k-27-z-1)!} \right) \Big|_{x=0} \right]$$

5 Next Steps

This completes our construction of a runs scored probability distribution given a set of nine batters with given batting characteristics.

Pitfalls Notice that there are a few significant items not considered in our formula:

- the fielding characteristics of the opposing team and the pitching characteristics of the opposing pitcher
- in-game strategic decisions such as bunting and stealing
- base-running characteristics of each batter

The next steps in this research should be to construct a more robust model that considers these factors.