MARKOV CHAIN MODELS: THEORETICAL BACKGROUND

by Mark D. Pankin

Introduction and Notation. In this article, I will provide the fundamental mathematical relationships for Markov chain models of baseball. The advent of personal computers that are more powerful than the standard mainframe computer of 25 years ago and the availability of data in computer readable form from Project Scoresheet and other sources permit many sabermetricians to work with Markov chain models should they desire. The material presented here (without excessive notational complexities) is intended to provide those with a background in matrix algebra the requisite mathematical tools. Also, there is a list of references at the end of the article. In future articles, I plan to report on the results of my Markov analysis of 1986 Project Scoresheet data.

If you are not familiar with basic matrix arithmetic—addition, multiplication, transposing—stop reading now. If you have a vague memory of these topics, you may want to dig out an old algebra book from high school or college. Some notation: matrices will be denoted by capital letters—A,B,C… —and numbers will be denoted by lower case letters—

A^t = the transpose of the matrix A
I = the identity matrix, which consists of 1's down the main diagonal and 0's everywhere else and has the property that if A is a square matrix of the same size, then AI = IA = A.
A^{-1} = the multiplicative inverse of the matrix A: $AA^{-1} = A^{-1}A = I$.

The presentation here follows that in Cover and Keilers with some new material I have added. Howard treats the topic in a somewhat different manner and shows a method of determining optimal strategies. His presentation, however, is considerably more strenuous mathematically.

The Basic Model. A Markov chain is a mathematical model that can be thought of a being in exactly one of a number of states at any time. Moreover, the transition probabilities of moving from a one state to another, which are the basis of the model computations, are dependent only upon the starting state of any transition, rather than upon how that state was reached. I will refer to this as the Markov chain assumption. In baseball models, the states are usually the various runners and outs situations. The Markov chain assumption means that we don't care how we arrived at a particular situation. For example, if there is a runner on second and one out, the Markov chain model is not concerned about whether there was a walk and a sacrifice, a double followed by a pop fly, etc. This assumption can be relaxed, the more general model is known as a Markov process, but the mathematics are more complex and beyond the scope of this article.

There are 24 possible runners and outs states, which are shown in the table below:

	Runners:							
Outs:	None	1st	2nd	3rd	1&2	1&3	2&3	1,2,&3
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

I use the numbering system shown in the table, but any numbering of the states is permissible. Four states to account for plays on which the third out is made augment those in the table:

#25: Third out made on play, no runs scored,
#26: Third out made on play, one run scored,
#27: Third out made on play, two runs scored,
#28: Third out made on play, three runs scored.

The probability of moving from one state (state i) to another (state j) is denoted by p_{ij} or $p_{i,j}$. For example, using the numbering above, $p_{2,8}$ is the probability of going from one out and none on to a runner on second and one out, which in turn is equal to the probability of a double in this situation plus the probability of a single and one base error, plus the probability of a two base error plus the probability of such plays as a walk and wild pitch on which the batter takes second. Many of the transitions are impossible, and hence have probabilities equal to zero. The 28 by 28 matrix of all such probabilities is called a transition matrix, and is denoted by T. For completeness, $p_{25,25} = p_{26,25} = p_{27,25} = p_{28,25} = 1$, and all other transitions from the three out states have probability equal to zero. For the moment, assume that in each transition, the batter is either out or gets on base. That is, the batter changes. This is an important assumption because it affects the run scoring calculations. Later, I will discuss methods of handling plays in which the batter does not change.

Because of the properties of matrix multiplication, T^2 is the transition matrix for sequences of two plays or batters, T^3 is the transition matrix for sequences of three batters, etc., providing there is a string of identical batters. This assumption is useful for computing "average" performance, but not for specific game situations. Later, what to do when the T's are not all the same will be considered.

Computation of Expected Runs. Let R be the 28 row by 1 column matrix (column vector) containing the expected or average runs scored after each state on one play only. Denoting the elements of R by R(1), R(2), …, some example calculations are:

$$R(1) = p_{1,1} \qquad R(2) = p_{2,2} \qquad R(4) = 2p_{4,1} + p_{4,4} + p_{4,7} + p_{4,10} + p_{4,2}$$

$$R(23) = 4p_{23,2} + 3(p_{23,5} + p_{23,8} + p_{23,11} + p_{23,3}) + 2(p_{23,14} + p_{23,17} + p_{23,20} + p_{23,6} + p_{23,9} + p_{23,12} + p_{23,27}) + p_{23,23} + p_{23,15} + p_{23,18} + p_{23,21} + p_{23,26}$$

$$R(25) = R(26) = R(27) = R(28) = 0.$$

The four values of zero reflect the fact that no runs can score after any three out state. As the formula for R(23) shows, some of these expected value computations can be fairly complicated. In practice, matrix and summation notations are used to express the above in a compact manner, which also corresponds to how they are implemented on a computer. In a real sense, it is easier to setup the above computations in a spreadsheet than to write them down on paper!

The key output of the Markov chain baseball model is the computation of the expected runs in the remainder of the inning after any runners and outs state. Let E be the 28 by 1 column vector containing these values. Then,

$$E = R + TR + T^2R + T^3R + \ldots \qquad \text{(Equation 1)}$$

This equation says that the expected runs after any state is the sum of the expected runs after one play, the expected runs after two plays, and so on theoretically forever. Of course, all innings end, and for some power of T, say about 25 or 30, the probability of not being in a three out state is negligibly small. What this means in practice is that it is necessary to compute only the first 25 or so terms of Equation 1. However, there is a way to write this equation more compactly. The following algebra is not strictly correct because we are dealing with an infinite series of matrices, but it does illustrate the "trick", and the result has been proven with mathematical rigor. First, note that R = IR. Then, Equation 1 can be transformed:

$$\begin{aligned} E &= IR + TR + T^2R + T^3R + \ldots \\ &= (I + T + T^2 + T^3 + \ldots)R \\ &= (I - T)^{-1}(I - T)(I + T + T^2 + T^3 + \ldots)R \quad [\text{since } (I - T)^{-1}(I - T) = I] \\ &= (I - T)^{-1}(I - T + T - T^2 + T^2 - T^3 + T^3 - \ldots)R \\ &= (I - T)^{-1}IR \\ &= (I - T)^{-1}R \qquad \text{(Equation 2)} \end{aligned}$$

Because no runs can score after a three out state, the equations above for expected runs can be simplified in their implementation by using only the first 24 rows and columns of the transition matrix T and the column vectors E and R. In fact, that simplification is necessary to ensure that $(I - T)^{-1}$ exists, which, in general, it does for baseball transition matrices. Using a spreadsheet program that has matrix multiplication and inversion commands, Equation 2 is much easier to implement than equation 1.

Applications. Depending on the source of the probabilities in T, the expected runs matrix E has a variety of useful interpretations. For example, if T includes all events for a league, then E contains the average expected runs scored in the remainder of the inning after each runners and outs state. Similar values can be computed for teams, a team's home and road games, or for an individual player. In the last case, we obtain an estimate of run scoring if that player batted all the time. Also, by restricting the transitions to those not influenced by strategies such as stolen bases or sacrifice bunts, we obtain expected run values that can be used to analyze those strategies. One problem with most of the expected runs tables that have been published is that they are based upon actual runs scored in major league play, which are influenced by the strategies employed in the games.

Note that E(1) is the expected runs after the no runners, no out state in which all innings begin. Thus, 9E(1) is the expected number of runs per 9 innings, or per game. This computation is especially interesting when applied to an individual player's stats. It provides another way of estimating how many runs per game would score if he batted all the time.

Real batting sequences. As mentioned previously, the basic model assumes a series of identical batters, which is, of course, not what happens in real games. This has been the major criticism of the applicability of Markov models. It is possible to have a Markov model with different transition matrices for each batter. The basic idea is to modify equation 1 and use it for the expected runs calculations because the simplification to equation 2 is no longer possible. Instead of powers of one transition matrix T, use products of the matrices for each batter: T_1T_2, $T_1T_2T_3$, etc. Also, the expected runs on the next play column vector R has to be modified for each batter. With such changes, equation 1, generalizes to

$$E = R_1 + T_1R_2 + T_1T_2\,R_3 + T_1T_2T_3\,R_4 + \ldots \qquad \text{(Equation 3)}$$

An additional potential complication is that it may be necessary to repeat the calculation in equation 3 for several sequences with different first batters and then weight the results by the probability of specific batters beginning the sequence.

It is important to note that these complications are often not necessary to overcome the criticism mentioned above. The key is performing the strategy analysis assuming a uniform sequence of outstanding batters and then repeating the analysis for a uniform sequence of poor batters. If the result is the same in both cases, then it is the answer. For example, if it doesn't make sense to sacrifice with a series of Wayne Tollesons coming up, then it certainly doesn't make sense for a series of Don Mattinglys or for any actual series of Yankee batters. If the results are not so clear cut, it may be possible to determine where between the two extremes the break-even point lies, and then judge if the actual batters are above or below this point.

Plays that do not change the batter. The expected runs calculations that are used in the matrix R, examples of which appear earlier, are not valid if the batter does not change on the play. Essentially, there is one less expected run for any given transition if the batter does not change than if the batter is changed. There are at least two ways to account for plays that do not change the batter. One is to add extra states and the other is to adjust the batter changing transitions to include those that do not change the batter. Both have advantages and disadvantages.

The first method calls for adding states that are the results of plays that do not change the batter. Ignoring errors on foul pops, which are not relevant to Markov analysis, the only situations that can't be reached by a non-batter play are those with the bases full (#22-24) and the third out, three runs scored (#28). Hence, 24 additional states would be added, for a total of 52. Because these additional states occur relatively infrequently and because the fact that a play did not change the batter is not likely to have a significant effect on what happens next, the transition probabilities out of these additional states should be the same as those for the corresponding batter changing states.

The main advantage of this method is that it combines all transitions in one matrix, which enables the easiest calculation of expected runs values corresponding to run scoring in actual games. However, if a sequence of different batters is considered, this method breaks down because we need to know whether or not the batter changed in order to use the correct transition matrix in the calculation. The solution is to combine both the batting changing and non-batter plays into a single adjusted batter changing transition matrix. Not surprisingly, the Markov Chain concept can be employed to express the required adjustments in terms of matrix arithmetic.

Start with a transition matrix T for the particular batter that includes non-batter plays. Although it would include some impossible states, it is best to think of this matrix as having 56 states—the 28 states described previously plus 28 similar states for non-batter plays. Also, the only part that matters is the top half or first 28 rows because, by definition the bottom 28 rows are identical to the top 28. We extract from T two 28x28 matrices, B and N, which contain the batter changing and non-batter changing probabilities, respectively. The probabilities in B and N are the same as those in T; they are not scaled so that the rows sum to 1 (which is true in T). For example, suppose the player has batted 20 times with a runner on first and no out (state #4), and that 6 times he singled moving the runner to third (#16), 3 times he hit into a double play, resulting in two out, none on (#3), 9 times he made an out that did not advance the runner (#5), and twice the runner stole second (#7, batter does not change). Then, we have the following probabilities in B and N: $b_{4,16} = 6/20 = .3$; $b_{4,3} = 3/20 = .15$; $b_{4,5} = 9/20 = .45$; and $n_{4,7} = 2/20 = .1$. All other entries in row 4 of each matrix would be zero.

When the batter comes up, depending on the runners and outs, there will be from zero to six (another fanciful construction, left to the reader) non-batter or base running events before the batter's play. If there are no such events, the appropriate transition probabilities are contained in the matrix B. If there is exactly one non-batter play, then the matrix that describes the batter changing transition probabilities is the product NB. If there are exactly two base running events first, then the transition probabilities for these two plays by themselves are contained in the matrix N^2, and the product N^2B is correct for the entire at bat. Similar logic applies for more running events before the batting event. The numbers of non-batter plays (0 to 6) followed by a batter changing play form a mutually exclusive and exhaustive set of outcomes for the particular at bat. Hence, the transition probabilities for the entire at bat, which take into account all the non-batter possibilities, are given by the matrix sum

$$\begin{aligned} & B + NB + N^2B + N^3B + N^4B + N^5B + N^6B \\ = \; & IB + NB + N^2B + N^3B + N^4B + N^5B + N^6B \\ = \; & (I + N + N^2 + N^3 + N^4 + N^5 + N^6)B \\ = \; & \text{T, the batter changing transition matrix incorporating running plays.} \end{aligned}$$

This matrix T, which has dimensions 28x28, can be used as the transition matrix in the other equations in this article

Scoring Probabilities. Analysis of one run strategies such as sacrifice bunt and stolen base attempts depends on the chances of scoring at least one run in the remainder of the inning, not on the expectation of total runs. The Markov model can produce scoring probabilities in a relatively direct manner. The essential idea is to add a new state that corresponds to one or more

runs having scored and to add the stipulation that no runs scored to the definition of the various runners and outs states (#1 - #24).

To formalize this concept, let T be a transition matrix (which may or may not include plays that do not change the batter) consisting of transition probabilities denoted by $p_{i,j}$. Let the new state, "runs scored", be denoted by r. Then S, the transition matrix for computing scoring probabilities, consists of entries $s_{i,j}$ defined by:

$s_{i,j} = p_{i,j}$ if the transition from state i to j does NOT score runs
$s_{i,j} = 0$ if the transition from state i to j DOES score runs
$s_{i,r}$ = the sum of all "scoring" $p_{i,j}$ for each row i
$s_{r,r} = 1$.

In practice, $s_{i,r}$ can be calculated by subtracting the sum of the other $s_{i,j}$ from 1 for each row i of S. Also, the three out states in which runs score (#26, #27, #28, and corresponding non-batter play states) can be removed from S because, by the definition of S, they can not be reached. Although the removal of such states compacts S, it is not necessary.

Once S has been computed, probabilities of scoring in the remainder of the inning are found by raising S to an appropriate power (assuming the model in which all batters are identical). If all transition probabilities in S change the batter, then the appropriate power is the sixth because after six more batters either the inning will be over with no additional runs scored or at least one more run will have scored. If S includes the probabilities of non-batter plays, then 15 is the required power as the following fanciful inning illustrates: the leadoff batter uses four plays to reach first, second, and third before being out at the plate; then the second batter goes through the same sequence; the third batter takes three plays to reach third; then the fourth batter reaches second on two plays without the runner on third moving; now the fifth batter walks, so a total of 4+4+3+2+1 = 14 plays have resulted in the bases loaded, two outs, and no runs in. The 15th play must either end the inning or score a run. (Remember that foul fly errors are eliminated from the Markov model, or else we might never get out of this!) Since successive squaring is relatively easy, in practice S^{16}, which is equal to S^{15}, would be computed. In the appropriate power of S, for each state (or row) i, $s_{i,r}$ is the probability of scoring at least one run in the remainder of the inning after that state, and $1-s_{i,r} = s_{i,25}$ is the probability of not scoring (#25 is the three out, no runs scored state).

If the batters are not identical, instead of finding S^6, compute the product of matrices $S_1S_2S_3S_4S_5S_6$, using an analogous scoring transition matrix for each batter. This product assumes that each of the matrices consists solely of batter changing transition probabilities. It might be possible to find the appropriate formula if this assumption is removed, but in this case, it is easier and better to incorporate both batter and non-batter plays into a single batter changing matrix, which was discussed previously.

Further reading. You may want to read some of he following books and articles, which are arranged in chronological order. The *Analyst* referenced is *The Baseball Analyst,* which was published by Bill James and primarily contained articles submitted by its readers. Part of this article first appeared in one of the issues.

Howard, Ronald A., *Dynamic Programming and Markov Processes*, MIT Press and Wiley, 1960, pp. 49-54.

Cover, Thomas M. and Keilers, Carroll W., "An Offensive Earned Run Average for Baseball" *Operations Research*, Vol. 25, No. 5, Sept-Oct 1977.

Pavitt, Charles, "Percentage Baseball Reconsidered: Model and Method" *Analyst*, #16, Feb. 1985.

Pankin, Mark D., "A Note About 'Percentage Baseball Reconsidered' " *Analyst*, #19, Aug. 1985.

Pavitt, Charles, "A Response of Mark Pankin's 'A Note About "Percentage Baseball Reconsidered" ' " *Analyst*, #21, Dec. 1985.

Boronico, Jess "The Baseball Batting Sequence Problem: Problem Formulation and Preliminary Results" *Analyst*, #23, April 1986.

Pavitt, Charles, "Percentage Baseball Reconsidered: 2. Preliminary 1984 Finding" *Analyst*, #26, Oct. 1986.

Katz, Stanley M., "Study of 'The Count'" *1986 Baseball Research Journal* (#15), pp. 67-72. An application of Markov chains to the ball-strike count.

Pankin, Mark D., "Baseball as a Markov Chain" *The Great American Baseball Stat Book* (First Edition, 1987), pp. 520-524.

Also worth reading is one important book of broader scope: *Optimal Strategies in Sports* edited by Ladany and Machol, North-Holland Publishing Company, 1977. This book still may be in print, and it should be in major university libraries. It covers the state-of-the-art at the time of mathematical and operations research methods applied to baseball (about half the book) and lesser pastimes. It includes three articles with a Markov flavor (by Richard Trueman, Richard Bellman, and Ronald Howard). Some of the material in the book requires a fair degree of mathematical sophistication, but there is much that is understandable by and of considerable interest to all sabermetricians, including summaries of the most significant early sabermetric work (by G. R. Lindsey and Earnshaw Cook).