Modeling Baseball Outcomes as Higher-Order Markov Chains

Jun Hee Kim (junheek1@andrew.cmu.edu)
Department of Statistics & Data Science, Carnegie Mellon University

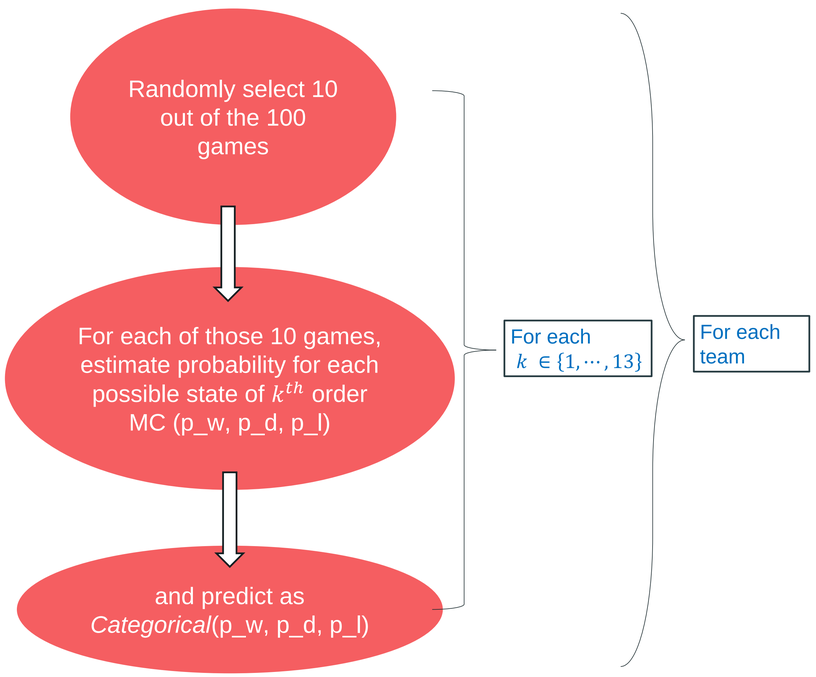
Background

- Baseball is one of the few sports in which each team plays a game nearly everyday.
- For instance, in the baseball league in South Korea, namely the KBO (Korea Baseball Organization) league, every team has a game everyday except for Mondays.
- This consecutiveness of the KBO league schedule could make a team's match outcome be associated to the results of recent games.
- Many KBO league news articles that preview and predict the games indeed mention the number of consecutive wins or losses the team currently has.

Research Objectives

- Model the match outcomes of each of the ten teams in the KBO league as a higher-order Markov chain, where the possible states are win ('W'), draw ('D'), and loss ('L').
- Identify patterns, if any, of the order of the higher-order Markov chain that best describes the match outcome sequence of a team and the team's overall performance (i.e. the rank in the league).

Method for Model Fit Assessment



For each team, compute prediction accuracy for each k.

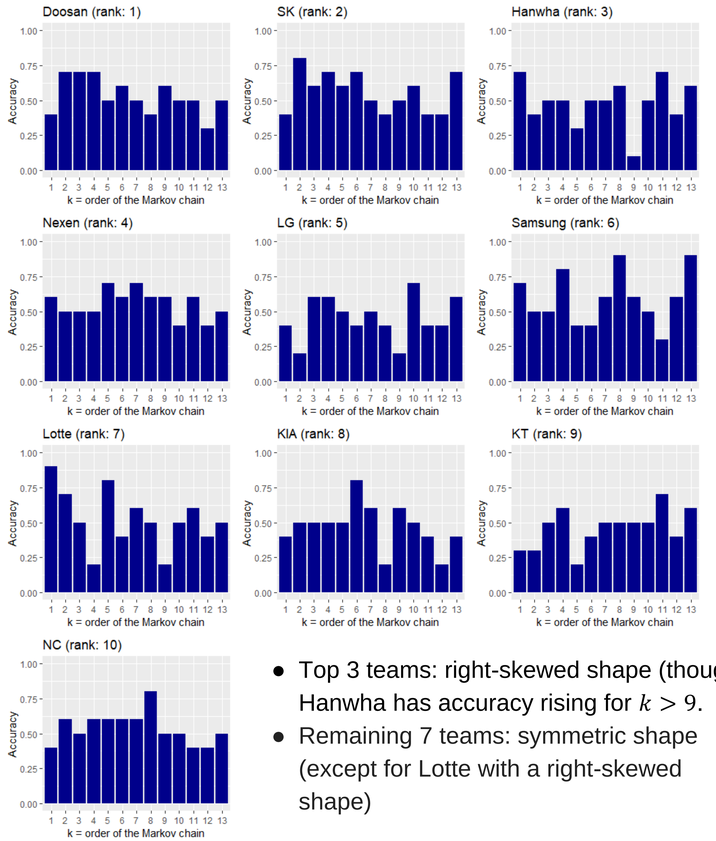
KBO League Introduction

- In each game, if the two teams have the same score after the 12th inning, the game ends as a draw.
- The 10 teams together compete in the pennant race where each faces the other 9 teams 16 times (8 home games and 8 away games): thus total 144 games per team.

Rank	Team	Games	Wins	Draws	Losses	Winning rate	Games behind
1	Doosan Bears	113	73	0	40	0.646	0.0
2	SK Wyverns	112	62	1	49	0.559	10.0
3	Hanwha Eagles	114	62	0	52	0.544	11.5
4	Nexen Heroes	118	61	0	57	0.517	14.5
5	LG Twins	116	56	1	59	0.487	18.0
6	Samsung Lions	116	54	3	59	0.478	19.0
7	Lotte Giants	110	51	2	57	0.472	19.5
8	KIA Tigers	110	51	0	59	0.464	20.5
9	KT Wiz	113	47	2	64	0.423	25.0
10	NC Dinos	116	47	1	68	0.409	27.0

Table 1: KBO League 2018 Rank (as of August 18th, 2018)

Results (1/2)



- Top 3 teams: right-skewed shape (though Hanwha has accuracy rising for $k > 9$.
- Remaining 7 teams: symmetric shape (except for Lotte with a right-skewed shape)

Higher-Order Markov Chain Model

- Extension of 1st order Markov chain

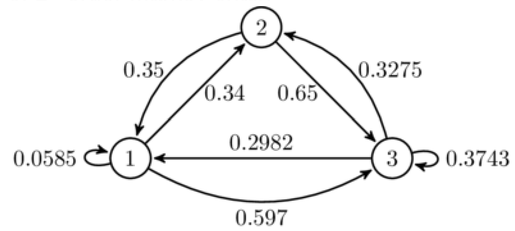


Figure 1: A 1st order Markov chain with 3 States

- k^{th} order Markov chain: the state at timestep n only depends on the states at the recent k timesteps.

$$P(X^{(n)} = x_{new} | X^{(n-1)} = x_{n-1}, X^{(n-2)} = x_{n-2}, \cdots, X^{(1)} = x_1)$$
$$= P(X^{(n)} = x_{new} | X^{(n-1)} = x_{n-1}, X^{(n-2)} = x_{n-2}, \cdots, X^{(n-k)} = x_{n-k})$$

- Each lag $l \in \{1, 2, \cdots, k\}$ has a weight λ_l its own transition probability matrix Q^l: For each pair of states i, j, the probability that the process will move to state i after l timesteps given that currently it's at state j.
- Model equation:

$$P(X^{(n)} = x_{new} | X^{(n-1)} = x_{n-1}, X^{(n-2)} = x_{n-2}, \cdots, X^{(n-k)} = x_{n-k}) = \sum_{l=1}^{k} \lambda_l q^{(l)}_{x_{new}, x_{n-l}}$$

Results (2/2)

- Top 3 teams: Once they have a good pace for a few recent games in a row, then it is likely that they will perform well again, but incorporating more earlier games makes predictions poorer.
- Remaining teams: Performance in recent games, regardless of how many we take, tend to not be influential to its performance today in the first place.

Next Steps

- For using formal statistical tests, increase size of data by collecting records from previous years of the KBO league and/or records of other leagues such as the MLB and NPB.
- Use the higher-order multivariate Markov chain model, where we are given multiple separate categorical sequences that all have the same state space. In particular, for each game, we can incorporate the two teams' sequences into this model.
- Consider other (obvious) factors including starting pitchers, whether the game is a home or away, etc thereby constructing a regression / classification model, with such factors and the recent game records as predictors.

Figure 1 from: https://en.wikipedia.org/wiki/Markov_chain