Power-ful rankings of NCAA men's basketball teams

Hugo Hinckfuss, Josh Livingston, Magnus Miller (Section #002)

2024-04-29

Abstract

Attribution

Introduction

Every spring, the fervor surrounding collegiate basketball peaks as fans engage in perpetual discourse over which team will reign supreme come the end of March. However, the process of determining the best teams is often clouded by subjective polls and rankings, such as those conducted by the AP, leaving room for confusion, disagreement, and frustration among enthusiasts, who will endlessly affirm the future success of their respective teams. While some publications offer alternative indices based on various mathematical formulas, they often remain shrouded in mystery to the general public due to the perceived complexity associated with mathematics.

Our group's interest in ranking collegiate basketball teams reached a pinnacle during this year's March Madness tournament, where multiple high seeds succumbed to the unexpected success of a poorly seeded team. These upsets prompted the group to explore a future where a mathematically-driven ranking system could not only align with the conclusions drawn by traditional voter-based polls but also aid in accurately predicting the teams' success in the tournament. In this pursuit, we discovered that numerous ranking methodologies rely heavily on the Perron-Frobenius theorem, which can offer an elegant solution to the challenge of uneven paired competition, wherein teams face disparate opponents throughout the season.

Once we have developed a comprehensive ranking system, we can then extend the system to assist with numerous systems beyond basketball, such as establishing rankings in other sports or other complex systems. Despite efforts to reduce subjectivity, ranking systems cannot completely eradicate it. Varying methodologies may produce different results for the top-ranked team, which will be largely influenced by the algorithm's emphasized factors set by us or a future user.

This paper will explore various ranking methods, showcasing the applications of the Perron-Frobenius theorem, eigenvalue computation techniques, fixed-point theorems, and probabilistic modeling. The initial method discussed in the paper presents the ranking problem as a linear eigenvalue problem, where we will utilize the Perron-Frobenius theorem directly. Next.. **have to clarify which other methods we will use then will complete this section.**

Finally, we will conclude by presenting our results from the application of the various ranking systems devised in an attempt to show a collegiate basketball ranking that is more reflective of the true capacities of the teams competing in March Madness.

Mathematical Formulation

Numerical Results

Data

We obtained team-game level data[1] from sports-reference.com[2] for the 2023-24 NCAA men's regular season. After doing basic data preparation, we bundled the data in an R package in this report. The R package, including the data and code for the computation, figures, and tables, is available on Github[3].

A sample of the data is available below. In total we have a 10,654 x 7 dataset. There are two rows per game in, one row per team. And for each team-game with have 7 columns: the team, date, home/away designation, opponent, result, team's score, and opponent's score.

Table 1: Sample of prepared NCAA men's basketball data

team	opp	result	team_score	opp_score
Texas State	Little Rock	L	66	71
Grand Canyon	Southeast Missouri State	W	88	67
Mississippi Valley State	Louisiana State	L	60	106
Southern	TCU	L	75	108
Jackson State	Memphis	L	77	94
Pacific	Sam Houston State	L	57	64
Southern Utah	Cal State Bakersfield	L	72	73
Tarleton State	Virginia	L	50	80
Bethune-Cookman	Minnesota	L	60	80
UT Arlington	Oral Roberts	W	75	71

[1] https://docs.google.com/spreadsheets/d/1-KL_Ib_YSkrnA24nWCGpmx8Y35xYIG9VmKAH71I3ObU

[2] https://stathead.com/basketball/cbb/team-game-finder.cgi?request=1&comp_type=reg&game_status=1&order_by=date&match=team_game&year_max=2024&order_by_asc=1&timeframe=seasons&comp_id=NCAAM&year_min=2024

[3] https://github.com/joshwlivingston/appm3310.final

Eigenvector computation

Win-loss matrix

We explore various methodologies to compute a_{ij}. First, we take the result of each game, and assign to the team a 0 for a loss, a 0.5 for a tie, and a 1 for a win. Then, we add the results and aggregate into a square matrix, denoted A_1 with entries $a_{1ij} : i, j = 1, 2, ..., 362$. A sample of this matrix is shown below. For simplicity, we'll let S_1 refer to the $4x4$ sample of A_1 with corresponding entries $s_{1ij} : i, j = 1, 2, 3, 4$.

Table 2: Sample of win-loss matrix

	Duke	NC State	Purdue	Tennessee
Duke	0	1	0	0
NC State	0	0	0	0
Purdue	0	0	0	1
Tennessee	0	1	0	0

In matrix notation, the matrix S is

$$S = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

The matrix entries $s_{1ij} \in S_1$ line up with team-game level results from the regular season.

Table 3: Regular season results for teams in the NCAA men's 2024 final four

team	date	home_away	opp	result	team_score	opp_score
Purdue	2023-11-21	NA	Tennessee	W	71	67
Tennessee	2023-12-16	NA	NC State	W	79	70
Duke	2024-03-04	away	NC State	W	79	64

A potential downside to this method, is that teams that have never matched have the same entry for s_{ij} as teams that have lost to another team. For both of this scenarios, $s_{ij} = 0$. You can see that $s_{21} = s_{23} = 0$, where $s_{21} = 0$ represents NC State's loss to Duke, and $s_{23} = 0$ represents NC State and Purdue never having played a game against each other in the regular season So, in this scenario, we lose information about losses. This can lead to teams with losses to have an inflated ranking.

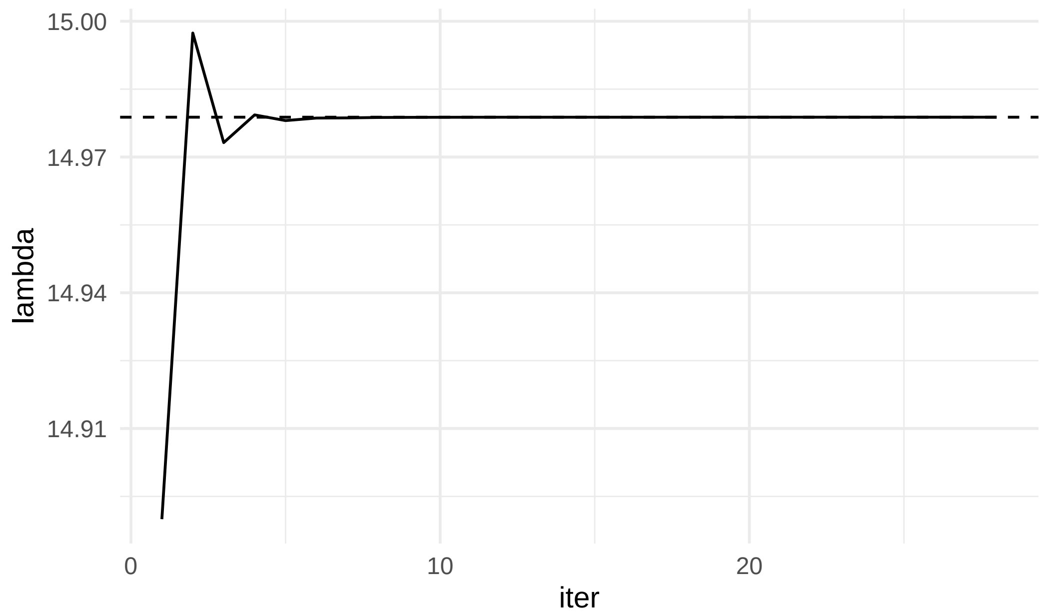
Eigenvalue approximation

We denote the eigenvalue r for matrix A_1 as r_1.

To approximate r_1, we employ the power method. At each step of the power method, we compute the approximate eigenvalue λ_{1i} where i is the iteration step of the power method approximation, and

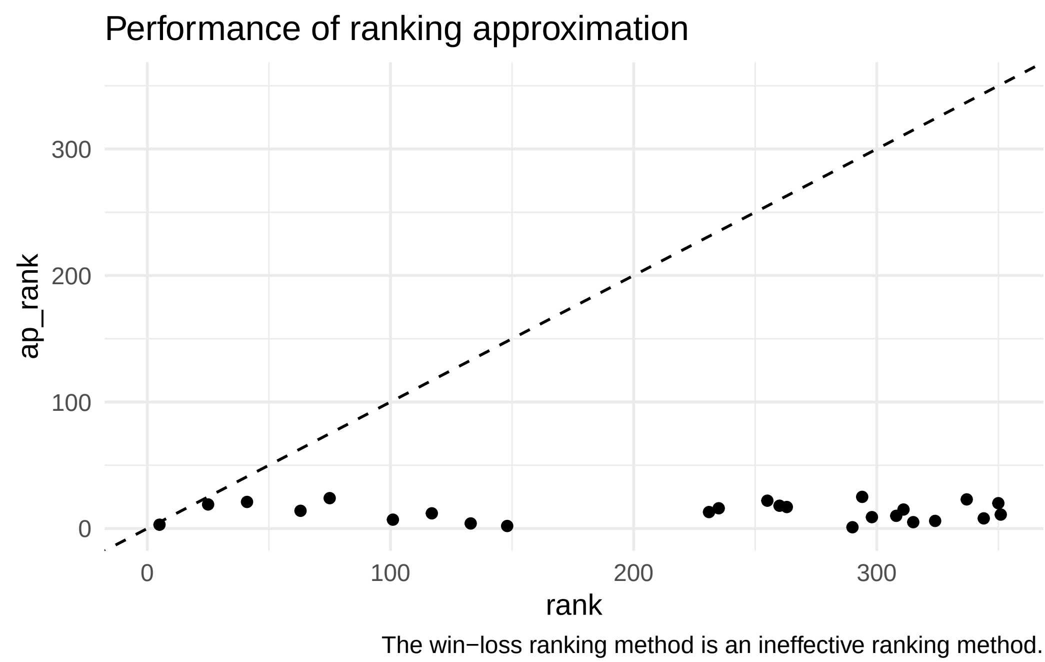
$$\lim_{i \to \inf} \lambda_{1i} = r_1$$

After approximation, we observe λ_{1i} converging to r_1, shown in the figure below.



Values of lambda converge to approximately 14.979 after 28 iterations

We use the approximated eigenvalue to compute the approximate eigenvector. We can use this eigenvector to compute rankings for NCAA men's basketball teams following the 2023-24 season. The rankings are shown below for the AP top 25 teams, showing that the win-loss method is an ineffective ranking scheme for NCAA men's basketball.



Additional Ranking Methods

Additional variations of a_{ij} did not produce better results. We used proportion of points, scored, various polynomial functions, distance between points scored, etc as values for a_{ij}.

To compare impact of values of a_{ij} on the resulting eigenvalues, we calculated the sum of squared difference between the eigenvector and the win-ratio. We denote this sum of squared difference for a matrix A_i $\vec{e}_i$ where i denotes the ranking function used to create the matrix A_i. Let $\vec{v}_i$ denote the ranking vector for A_i and $\vec{w}_i$ denote the win-ratio's for the teams in matrix A_i. Then,

$$\vec{e}_i = (\vec{v}_i - \vec{w}_i)^2$$

The best performing matrix is the A_i that minimizes $\vec{e}_i$. We denote this matrix A^*, with entries a^*_{ij}.

The table below shows a summary of each A_i with its corresponding e_i.

Table 4: Table showing the top five ranking matrices with corresponding error metrics

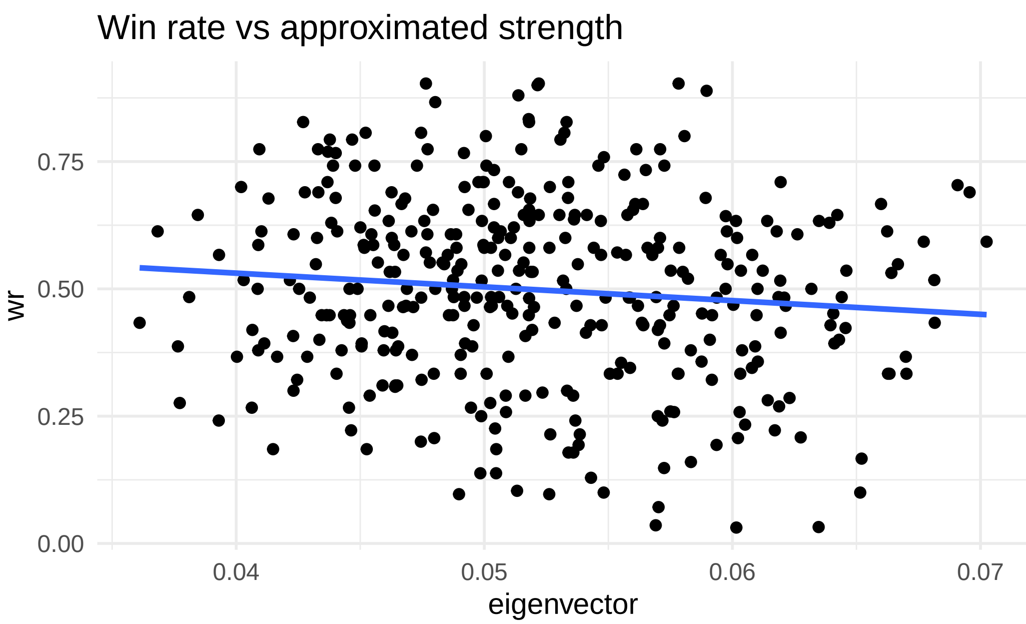
ranking_table_name	e
points_proportion	84.121
points_polynomial5	84.151
points_sqrt_points	84.166

ranking_table_name	e
points_polynomial3	84.245
points_scored_mean	84.288

The best performing table is the table with entries a_{ij} constructed with the proportion of points scored in each game. Let s_i denote a team's score in a game, where s_j denotes the opponent's score.

$$a^*_{ij} = \frac{s_i}{s_i + s_j}$$

The figure below shows the eigenvalues plotted against the win-ratio's for all 362 teams in the data using a^*_{ij} as the entries for A^*. We also included plotted the linear regression line between the two points, for reference. We can see that for the full league, the eigenvalues bear little relationship to the win-ratios, suggesting that none of the tested ranking methods are effective at ranking the entire NCAA.



re against the win−ratios for the 362 teams in the dataset, is ineffective at ranking the entire NCAA

Sparsity

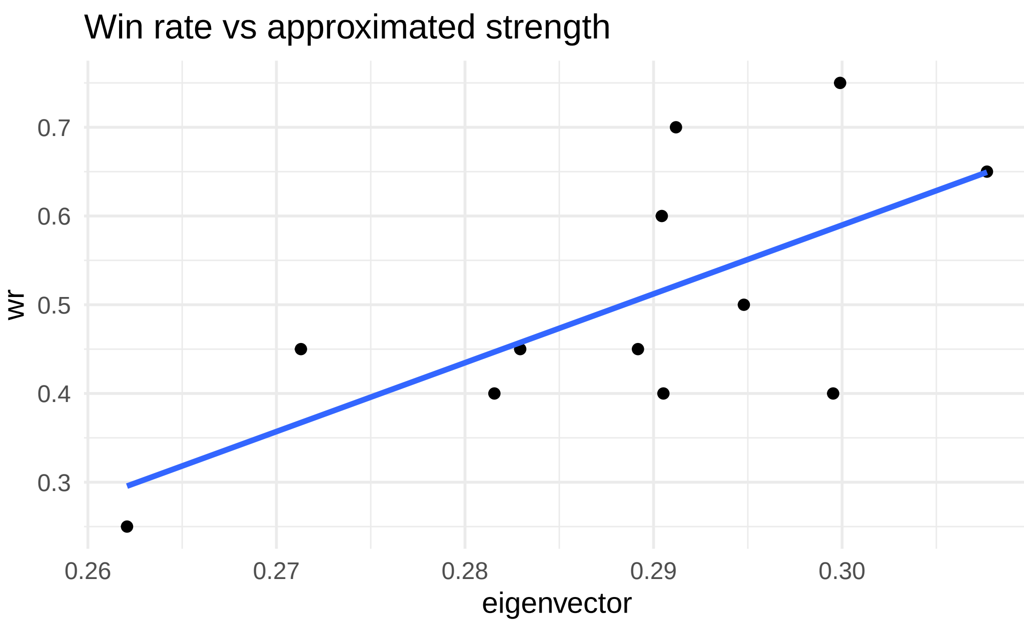
One issue effecting results in the sparsity of the ranking matrices we are using. When analyzing the matrix comprised of 362 teams x 362 teams, most of whom do not face each other in the regular season, the matrix is going to be mostly sparse.

To explore the role of sparsity, we look at a second, filtered dataset, comprising of in-conference games between PAC-12 teams. That is, we only look at games in which two PAC-12 teams played each other. The figure below shows the same 5 ranking schemes are the most effective ranking schemes.

Table 5: Table showing the top five ranking matrices with corresponding error metrics

ranking_table_name	e
points_proportion	0.747
points_scored_mean	0.763
points_sqrt_points	0.764
points_polynomial5	0.776
points_polynomial3	0.782

Looking at the relationship between the eigenvector and the win-ratio, we see a much stronger relationship, suggesting sparsity limits the effectiveness of theme ranking schemes.



When limited to conference play, the ranking method via proportion is an effective ranking method.