STAT 488 Paper 1: The NBA as an Evolving Multivariate System

Charles Hwang

2/10/2022

Introduction

The article I chose is titled "The NBA as an Evolving Multivariate System" and written by Sangit Chatterjee and Mustafa R. Yilmaz. The paper discusses its interest in how the obsolescence of the .400 batting average in Major League Baseball (with Ted Williams being the last player to accomplish the feat in 1941) was not due to a decline in overall hitting performance over time; in fact, it was quite the opposite. Rather, the league as a whole was improving, and the extinction of the .400 batting average was due to a decline in variability over time among hitters. Analyses of the league's mean, standard deviation, and maximum batting averages and the standardized score of the maximum over each of the previous nine decades (Table 1) showed no clear trend with the mean or maximum batting averages or standardized score, but the standard deviation of batting averages definitively declined after each decade. Chatterjee and Yilmaz sought to conduct a similar study with the National Basketball Association (NBA) using basic multivariate and time-series analyses.

Statistical Methods Employed

The following research questions were posed (paraphrased from the article):

1. Is there a discernible pattern of change in scoring?
2. Is there a pattern of declining variation in performance as indicated by multivariate measures of variability?
3. Is there a pattern in the best overall individual performance as measured by a multivariate standardized score? Relative to the league average, are the best performances of today comparable to the best of the past?

Statistics analyzed included points (PPM), rebounds (RPM), assists (APM), steals (SPM), and turnovers per minute (TPM), all standardized over minutes to avoid inaccuracies with different numbers of minutes, games,

or minutes per game played by each player. Basic **time-series plots** of points per game (PPG) and PPM over time (Figure 1) illustrate the differences between the two. The study used $t = 1994 - 1951 = 43$ different matrices ($\mathbf{X}_t$) of size $n \times p$, where t is the season from 1951-52 (the first season that minutes played was recorded as an official statistic) to 1993-94 (the extent of the dataset), n is the player, and p is the statistic. As steals and turnovers were not recorded until the 1973-74 season, the analysis created matrices with $p = 3$ columns for all seasons and matrices with $p = 5$ columns for the 1973-74 season onwards. For each season t, the **covariance matrix** Σ_t was computed along with its *determinant* $|\Sigma_t|$ (the generalized variance) and *trace* (the sum of the variances of p), and these are shown in Figures 2 and 3. Chatterjee and Yilmaz note their choice of the population matrix Σ_t instead of the sample matrix $\mathbf{S}_t$ since their data was from the entire population of all NBA players who played in each season.

The **squared Mahalanobis distance** $D_{it}^2 = (\mathbf{x}_{it} - \mu_{it})'\Sigma_t^{-1}(\mathbf{x}_{it} - \mu_{it})$ from the mean was calculated for the Most Valuable Player (MVP), player i, of each season, where $\mathbf{x}_{it}$ is the vector of the player's statistics for season t and μ_{it} is the vector of means of the $p = 3$ statistics. The Mahalanobis distance D is a multivariate generalization of the univariate z-score that can be found using the `mahalanobis()` function in RStudio. The D^2 values for each season's MVP from 1955-56 (the first season the award was given) to 1993-94, whether the MVP had the maximum D^2 value that season, and whether the MVP's team won their division that season is shown in Table 2 (misprinted as "Table 1"). Chatterjee and Yilmaz note they analyzed the MVPs of each season rather than the players with the maximum D^2 value because the maximum D^2 value may not necessarily have the maximum value of any individual statistic p. They write: "In 28 out of 39 seasons, there were other top players whose D^2 values were larger than the MVP[']s." An additional **time-series plot** of MVPs' D^2 over time shows no clear trend (Figure 4).

Critique of Statistical Methods

The substaintial yet basic analysis of the $\mathbf{X}_t$ matrices produced fairly robust statistics for the generalized variance and the sum of variances of p. The variables are measured consistently for each player, but Chatterjee and Yilmaz **do not sufficiently explain** how they capture the true variability between seasons, only writing: "It is noted that, with multivariate data, both measures of variability have some shortcomings that cannot be avoided. ... It is hoped that use of both measures gives a more complete picture of variability in multivariate data, especially if they exhibit similar patterns over time".

Additionally, the five variables (PPM, RPM, APM, SPM, TPM) appear to be **chosen subjectively**. The omission of blocks per minute, which is often cited before turnovers and used as a defensive metric, was not explained. Given the limitation of their dataset and the fact that this study was published in 1999—prior to

the exponential rise in usage of databases like Basketball Reference and statistical methods like data analytics and machine learning in sports—I can give some flexibility here. However, some important variables—field goal percentage, personal fouls per minute, rebounding percentage, and plus-minus, to name a few—were excluded from the matrices with no overview or explanation. Chatterjee and Yilmaz only write: "Although there are numerous other less-common measures of performance, we conducted our analyses using these five measures. ... Clearly, MVP selection involves more dimensions than the performance measures we used in this study."

Lastly, the **selection of the MVP each season** has been highly controversial and hotly debated. Since the 1980-81 season, sportswriters and broadcasters have been responsible for awarding the title (previously it was voted on by players), and arguments regarding favoritism, lack of recent basketball knowledge or involvement, and awarding based on team performance rather than individual performance have run rampant ever since. Despite the subjectivity involved, Chatterjee and Yilmaz still chose to use the D^2 values of MVP. They acknowledge in their conclusion: "... selection of the top player in the league (or more generally, a complete ranking of multivariate observations) is intrinsically difficult and ambiguous. In addition to numeric measures, reference is often made to 'intangibles' in selecting an MVP, such as motivating other players on the team to play better. That the largest D^2 does not always belong to the MVP is not an accident."

Conclusion

I believe this is a well-conducted study given the limited statistics and technology available to Chatterjee and Yilmaz at the time. They mentioned there did not appear to be any similar study previously published, and it is interesting to see the changes in variability in the NBA over time. My personal interest in sports and the basic multivariate analyses conducted made this article easier to understand and review. This is an appropriate real-world use of multivariate analysis and all statistical assumptions and practices were properly made and followed.

Works Cited

Chatterjee, Sangit, and Mustafa R. Yilmaz. "The NBA as an Evolving Multivariate System." The American Statistician, vol. 53, no. 3, Taylor & Francis Group, 1999, pp. 257–262, https://doi.org/10.1080/00031305.1999.10474469.