A Multivariate Statistical Analysis of the NBA

Lori Hoffman
University of Wisconsin River Falls
River Falls, WI

Maria Joseph
Kentucky State University
Frankfurt, KY

Abstract

Will your favorite National Basketball Association (NBA) team make it to the playoffs this year? What variables affect a team's postseason outcome? In an attempt to determine which teams will make the NBA playoffs, we will collect and analyze team data using multivariate statistical methods including Principal Components Analysis and Discriminant Analysis.

Introduction

It is midway through the NBA season. The owner of the Milwaukee Bucks takes a look at the standings; it is a close race in the Central Division. The one question weighing on his mind, will we make the playoffs this year? With so much revenue to be made by clubs as well as enjoyment to be had by the fans, making the playoffs is vital to any team. Is there a way to determine with relative certainty whether or not a given team will make it to the playoffs? Division standings are likely to change throughout the season, so are there other components that one can use to predict which teams will make it?

The goal of this paper is to analyze data from the NBA in order to gain insight into the questions asked above. We begin by constructing a list of variables that we think are important for any team to make the playoffs. Then, we use Principal Components Analysis (PCA) on these variables to find new components to describe our data. A key result of PCA is that we reduce the dimensionality of the data set. We also use Discriminant Analysis on our variables to predict the classifications of teams into one of two populations, playoffs or non-playoffs.

Data

We chose 12 variables that fall under the categories of game statistics, player demographics, team finances, coaching, and fan support. The data comes from various websites containing information on the NBA. Some variables are taken as is and others needed further calculations. The following is a table of variables and how they are calculated:

Variables	Description
Points Per Game Offense (PPG Off)	Average points scored per game
Points Per Game Defense (PPG Def)	Average points allowed per game
Field Goal Percentage (FG %)	Team percentage of field goals made
Turn Over Score (T O Score)	Defensive turnovers less offensive turnovers
Previous Team Record (Prev Rec)	Last season's percentage of games won
Average Home Crowd (Crowd)	Average attendance per game
Years in the NBA	Number of years since team's establishment
Payroll	Rank of team's total yearly payroll
Coach's Record (Coach)	Head coach's cumulative NBA record
Rebounds Per Game (Reb/game)	Average number of offensive and defensive rebounds per game
New Player Ratio (New Player)	Number of new players to the roster divided by the total number of players
Median Age	Median age of the team

Principal Components Analysis Theory

When analyzing data with multiple variables, Principal Components Analysis (PCA) can be used to reduce the size of the data set. PCA uses a linear combination of the original correlated variables to form new, uncorrelated variables called principal components. In order to use PCA on a data set, we need to verify two important conditions: normality and dependence of variables. For each of these conditions, we use hypothesis testing to see if the data meets its criterion. PCA is most effective when the vector of p variables of interest $\underline{x}$ has a multivariate normal distribution with mean vector

$$\underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$ and variance-covariance matrix $$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ & \sigma_{22} & \cdots & \sigma_{2p} \\ & & \ddots & \vdots \\ & & & \sigma_{pp} \end{bmatrix}.$$

First, let us test for normality of each individual variable x_i. We will construct a Q-Q Plot, which is a graphical representation of the relationship of the ordered statistics and the ordered normal quantiles. In order for the variables to be normal, the Q-Q Plot should show a linear relationship implying a high correlation between the statistics and normal quantiles. We test this correlation using the null hypothesis, H_o: $\rho = 1$, versus the alternative hypothesis, $H_1: \rho < 1$, at a chosen significance level α. If H_o is accepted, we conclude that our data set is normally distributed. The critical point c is the cut-off for rejecting H_o; that is, we reject H_o if $r < c$.

Next, we proceed to test for dependence of variables. Test H_o: P = I versus H_1: P ≠ I. If H_o is rejected, we conclude that the variables are dependent. The test statistic for this likelihood ratio test is

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right)\ln|R|,$$

with $\frac{p(p-1)}{2}$ degrees of freedom. If the observed χ^2 is more than the critical point χ^2_α taken from the chi-square distribution table, then H_o is rejected, concluding that the variables are dependent.

After testing the variables for normality and dependence, we can use PCA to reduce the number of variables. The principal components we derive are of the form $y_i = \underline{\ell}_i' \underline{x}$, where $\underline{\ell}_i = \begin{bmatrix} \ell_{i1} \\ \ell_{i2} \\ \vdots \\ \ell_{ip} \end{bmatrix}$. These linear combinations of the original variables maximize the variance, with ℓ_{ij} being the weight given to the jth variable on the ith component. The number of principal components we choose to retain is dependent upon the percentage of variance each component accounts to the total variance. That is, we want to account for most of the variance but still keep the number of principal components low relative to the number of original variables. Once the principal components have been found, they can be identified by carefully interpreting which variables have the greatest contribution to each component.

Application of Principal Components Analysis

As previously stated, we need to test for normality and dependence of variables. First, let us test for normality using a Q-Q Plot. With the help of Minitab to test the linearity of our Q-Q Plot, we calculated each variables' correlation with its z-score and got the following table of results:

Variable	Correlation
PPG Off	0.995
PPG Def	0.995
FG %	0.994
T O Score	0.993
Prev Rec	0.982
Crowd	0.993
Years in NBA	0.965
Payroll	0.985
Coach	0.956
Reb/ Game	0.997
New Player	0.993
Median Age	0.998

Testing at $\alpha = .05$, we obtain a critical point of c = .9838 from the table. We are able to conclude that 9 of our variables are normal since their correlations are all greater than c. Only *Previous Record*, *Years in NBA,* and *Coach's Record* fail to pass the normality test as they come from a skewed family of distributions. We tried using both natural log and square root transformations on these three variables in an attempt to normalize them but did not succeed. However, due to the robust nature of PCA, we can still include all variables in the analysis.

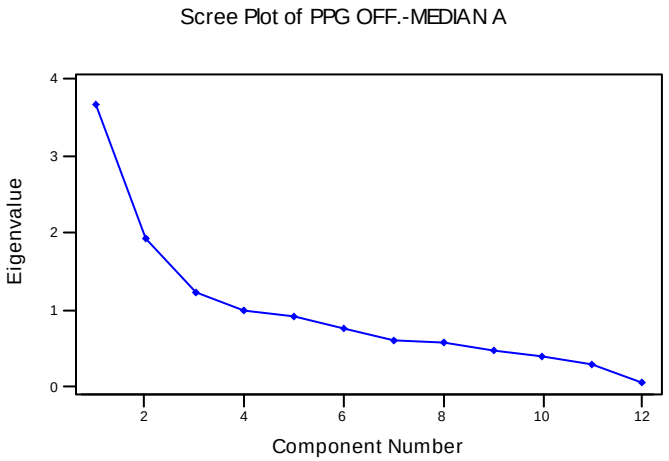
After testing for normality, let us proceed to test for dependence. Using the Chi-square test at significance level $\alpha = .05$ with 66 degrees of freedom, we test H_o: P = I versus H_1: P ≠ I. With n = 87 and p=12, we get $\chi^2 = 261.27$ with a corresponding p-value approximately equal to 0. Our calculated χ^2 is well above our cut off value of 79.082 allowing us to reject the hypothesis that the data is independent. Now, we are ready to apply PCA to the data.

Using Minitab we run PCA and compute the Eigenanalysis of the Correlation Matrix giving us the following results:

	PC 1	**PC 2**	**PC 3**	**PC 4**	**PC 5**	**PC 6**
Eigenvalue	3.6668	1.922	1.226	1.0021	0.925	0.7659
Proportion	0.306	0.16	0.102	0.084	0.077	0.064
Cumulative	0.306	0.466	0.568	0.651	0.729	0.792
	PC 7	**PC 8**	**PC 9**	**PC 10**	**PC 11**	**PC 12**
Eigenvalue	0.6257	0.584	0.49	0.4083	0.307	0.0761
Proportion	0.052	0.049	0.041	0.034	0.026	0.006
Cumulative	0.844	0.893	0.934	0.968	0.994	1

As determining the number of principal components is a subjective process, three different observations are taken into account. First, we look at the eigenvalues of each

component. Generally, we want to keep all components with eigenvalues greater than 1. In this case, we retain the first four. However, we also want to take into account how much of the total variance each component contributes. Using this second observation, we see that using the fifth component as well brings us above 70% of the total variance accounted for, which is sufficient enough for our model. Finally, we look at a Scree Plot of the results in order to have a graphical representation of the relationship between principal components and their respective eigenvalues.



Looking at the Scree Plot, we see that the graph begins to level off at component number 6. This suggests we should retain the first 5 components. Taking into account all of these observations, we decide to keep 5 principal components.

The new principal components are a linear combination of the original 12 variables. The following table shows the coefficients of each variable for the first 5 components we chose to retain.

Variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
PPG Off	0.236	0.594	0.163	-0.02	0.047
PPG Def	-0.265	0.518	0.123	0.038	0.207
FG %	0.347	0.142	0.183	-0.222	0.453
T O Score	0.189	0.194	0.404	0.46	-0.504
Prev Rec	0.432	-0.021	-0.143	-0.158	-0.024
Crowd	0.294	-0.305	-0.14	0.265	-0.107
Years in NBA	0.015	0.167	-0.578	0.624	0.219
Payroll	-0.3	0.124	0.065	-0.101	-0.475
Coach	0.325	0.088	-0.212	-0.135	-0.373
Reb / Game	0.092	0.391	-0.534	-0.304	-0.24
New Player	-0.326	-0.084	-0.209	-0.306	-0.111
Median Age	0.366	-0.129	0.098	-0.197	-0.041

Using this table we are able to interpret each principal component and label it accordingly. For PC1 we notice that *Prev Rec, Coach's Record,* and *Median Age* all

contribute significantly. Therefore, we conclude that PC1 should be labeled *Past Experience.* We use similar arguments for the PC2 and PC4 and label them as *Scoring* and *Team Establishment,* respectively. However, we couldn't find a significant relationship between the variables in the remaining components so we will label them as *PC3* and *PC5.*

Using PCA we are able to reduce the 12 original variables down to only 5 principal components. Using these new components we get a better understanding of what elements contribute to a NBA playoff team.

Theory of Discriminant Analysis

The basic goal of discriminant analysis is to classify an observation vector $\underline{x}$ into one of two different populations, Π_1 or Π_2. It is assumed that both Π_1 and Π_2 have multivariate normal distributions with respective means $\underline{\mu_1}$ and $\underline{\mu_2}$ and the same variance-covariance matrix $\Sigma_1 = \Sigma_2 = \Sigma$. In order to classify an observation vector, we first need a rule to do so. Random samples from Π_1 and Π_2 of sizes n_1 and n_2, respectively, help to develop such a rule and are referred to as training samples. For our data, we will use the *Linear Discriminant Function Rule*.

Below we construct an optimal partition of the p-dimensional space of $\underline{x}$ consisting of disjoint subsets R_1 and R_2. If $\underline{x} \in R_1$, then classify $\underline{x}$ into Π_1. Otherwise, if $\underline{x} \in R_2$ then classify $\underline{x}$ into Π_2.

Consider a linear combination of the form $\underline{a}'\underline{x} = a_1x_1 + a_2x_2 + ... + a_px_p$, where $\underline{a} = \Sigma^{-1}(\underline{\mu_1} - \underline{\mu_2})$. The function $\underline{a}'\underline{x}$ is called the *linear discriminant function* of $\underline{x}$ and it is used to classify $\underline{x}$ by comparing it with a fixed number h defined by the formula $h = \frac{1}{2}\underline{\delta}'\Sigma^{-1}(\underline{\mu_1} + \underline{\mu_2})$, where $\underline{\delta} = (\underline{\mu_1} - \underline{\mu_2})$. We proceed to define R_1 as the subset of p-space where $\underline{a}'\underline{x}$ is greater than h and R_2 as the subset where $\underline{a}'\underline{x}$ is less than or equal to h. Thus, the linear discriminant rule is to classify $\underline{x}$ into Π_1 if $\underline{a}'\underline{x} > h$. Similarly, classify $\underline{x}$ into Π_2 if $\underline{a}'\underline{x} \leq h$.

However, these classifications will not always identify $\underline{x}$ as being from the correct population. Therefore, define α_1 as the probability of misclassifying $\underline{x}$ into Π_2 and α_2 as the probability of misclassifying $\underline{x}$ into Π_1. It can be shown that h is the optimal cut-off for subsets R_1 and R_2. That is, h is chosen such that $\alpha_1 + \alpha_2$ is minimized given the constraint $\alpha_1 = \alpha_2$.

In practice we will never know the true values of $\underline{\mu_1}$, $\underline{\mu_2}$, and $\sum$. On the other hand, the training samples from each of the two populations give us $\underline{\bar{x}_1}$, $\underline{\bar{x}_2}$, S_1, and S_2 ,which are unbiased estimates of $\underline{\mu_1}$, $\underline{\mu_2}$, Σ_1, and Σ_2, respectively. The pooled estimate $S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$ is used as an unbiased estimate of $\sum$ when it is assumed that $\Sigma_1 = \Sigma_2 = \Sigma$. We use these estimates in place of the population parameters in the aforementioned discriminant rule.

After constructing the linear discriminant function for $\underline{x}$, we want to check to see how accurate the classifications are. To do this, we calculate the Apparent Error Rate (AER), which is the percentage of misclassified measurements from the training sample. We are also interested in the total probability of misclassification (TPM), which is the probability of misclassifying objects using the classification rule. In order to calculate the TPM, we need to measure the distance between the two populations, denoted Δ_p. We take the squared distance to be

$$\Delta_p^2 = (\text{Mahalanobis Distance between } \Pi_1 \text{ and } \Pi_2)^2 = (\underline{\mu_1} - \underline{\mu_2})'\Sigma^{-1}(\underline{\mu_1} - \underline{\mu_2}).$$

Using this distance, the TPM is given by

$$\text{TPM} = 2\Phi(-\frac{1}{2}\Delta_p),$$

where $\Phi(z)$ is the cumulative distribution function of $N(0,1^2)$. Furthermore, we can use the Mahalanobis distance as an alternative rule for classifying $\underline{x}$. That is, if the Mahalanobis distance of $\underline{x}$ from Π_1 is less than the Mahalanobis distance of $\underline{x}$ from Π_2, then we classify $\underline{x}$ into Π_1. A similar argument works for classifying $\underline{x}$ into Π_2. This distance rule also can be generalized to cases when we have more than two populations. Classify $\underline{x}$ into Π_i if

$$(\text{Mahalanobis Distance of } \underline{x} \text{ from } \Pi_i)^2 = (\underline{x} - \underline{\mu_i})'\Sigma^{-1}(\underline{x} - \underline{\mu_i}) = D_i^2 = \text{Min}(D_1^2, \ldots, D_k^2)$$

Application of Discriminant Analysis

For discriminant analysis, we want to classify teams into one of two populations: Π_1=Playoff and Π_2=Non-playoff. We use the NBA seasons from 1999 through 2001, to give us training sample sizes of n_1=48 and n_2=39. After developing a discriminant function based on these two samples, discriminant analysis can predict what teams make the playoffs in 2002. We use the 2002 because it is the most current data we can obtain.

We assume that $\Sigma_1 = \Sigma_2$ in order to use the linear discriminant function. Therefore, we must first verify this equality is true before proceeding. We test H_o :

$\Sigma_1 = \Sigma_2$ versus $H_1 : \Sigma_1 \neq \Sigma_2$ using the likelihood ratio test. For this test our test statistic is $\frac{m}{c}$, where

$$m = \sum_{i=1}^{2}(n_i - 1)\ln|S_p| - \sum_{i=1}^{2}(n_i - 1)\ln|S_i| \text{ and}$$

$$\frac{1}{c} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}\left(\frac{1}{(n_1 - 1)} + \frac{1}{(n_2 - 1)} - \frac{1}{(n_1 + n_2 - 2)}\right).$$

We calculate the test statistic and obtain $\frac{m}{c}$=14.53, which gives us a corresponding p-value of 1. Therefore, testing at α=.05, we have overwhelming evidence to accept H_o and conclude that $\Sigma_1 = \Sigma_2$. This justifies using a linear discriminant model.

We first run the analysis on both the raw data - the original 12 variables - and the five principal components' scores. The AER for the raw data is 5.7% while the AER for the principal components' scores is 14.9%. This difference in AER is a result of only accounting for 72.9% of the variance when using the principal components. Therefore, we will use the raw data in running the discriminant analysis. The following table shows the results:

Classified Group	**True Group Non-playoff**	**True Group Playoff**
Non-playoff	35	1
Playoff	4	47
Total N	39	48
N Correct	35	47
Proportion	.897	.979
N =87	N Correct = 82	Proportion Correct = 94.3%

Having an AER of 5.7% tells us that our model is fairly accurate in classifying our training sample. The AER is an estimate of the total probability of misclassification (TPM). The results of this model tell us that the squared distance between populations ($\hat{\Delta}_p^2$) is 9.6396. Using this distance we calculate the TPM to be 6.03%. We now use discriminant analysis on data outside the training sample in order to predict how a team will finish. For this purpose we chose 2002 NBA season data.

When first running discriminant analysis on the 2002 data, our results predict that every team will make the playoffs. Obviously, this suggests a flaw in our model. A review of the raw data shows that different variables take on dramatically different values. For example, FG% is between 0 and 1, while Average Home Crowd take on values in the high thousands. Hence, it makes sense to standardize all variables, including the training sample, in an attempt to place them on a comparable scale. Running the analysis on the transformed training sample gives the same results as the table above. The analysis on the transformed 2002 data gives more desirable results than previously, as can be seen in the following table:

Team	Predicted Population	True Population
Boston	Playoff	Playoff
Miami	Non-playoff	Non-playoff
New Jersey	Playoff	Playoff
New York	Non-playoff	Non-playoff
Orlando	Playoff	Playoff
Philadelphia	Playoff	Playoff
Washington	Playoff	Non-playoff
Atlanta	Non-playoff	Non-playoff
Chicago	Non-playoff	Non-playoff
Cleveland	Non-playoff	Non-playoff
Detroit	Playoff	Playoff
Indiana	Playoff	Playoff
Milwaukee	Non-playoff	Playoff
New Orleans	Playoff	Playoff
Toronto	Non-playoff	Non-playoff
Dallas	Playoff	Playoff
Denver	Non-playoff	Non-playoff
Houston	Non-playoff	Non-playoff
Memphis	Non-playoff	Non-playoff
Minnesota	Playoff	Playoff
San Antonio	Playoff	Playoff
Utah	Playoff	Playoff
Golden State	Non-playoff	Non-playoff
LA Clippers	Non-playoff	Non-playoff
LA Lakers	Playoff	Playoff
Phoenix	Playoff	Playoff
Portland	Playoff	Playoff
Sacramento	Playoff	Playoff
Seattle	Playoff	Non-playoff

This model is able to classify 26 of the 29 teams correctly, giving an error rate of 10.34%.

Conclusion

We began with 12 variables for 87 teams in the NBA. In order to get an idea of what makes a playoff team, we used PCA to reduce the dimensionality of our data. We feel that the 5 components we selected give us a better understanding of the attributes of a playoff team. In classifying teams in the training sample, our discriminant model did a good job. Although our error rate is higher for predicting teams outside the training sample, the model is still adequate. After reflecting upon our model, we find ways in which it can be improved. In order to scale down the Average Home Crowd variable, we could have divided it by the number of fans each arena can hold. Also, we could have

looked at player composition of each team. For instance, how many "all-stars" does a team have? Aside from these suggested improvements, we feel our model has shown how multivariate statistics can be applied to the NBA.

Acknowledgements

This research could not have been completed without the help of many people. First of all, we would like to thank Dr. Vasant Waikar for introducing us to the world of multivariate statistics as well as guiding us through this research process. We would also like to thank our Graduate Assistant Candace Porter for her support and friendship. In addition, many thanks go to Dr. Tom Farmer for helping us with the composition of this paper. Finally, we would like to thank all the SUMSRI staff for giving us such a great experience.

References

[1] Basketball Reference. Retrieved June 26, 2003, from www.basketballreference.com

[2] Bender, Patricia. Retrieved June 30, 2003, from www.dfw.net/~patricia/misc/salaries00.txt

[3] Bender, Patricia. Retrieved June 30, 2003, from www.dfw.net/~patricia/misc/salaries01.txt

[4] Bender, Patricia. Retrieved June 30, 2003, from www.dfw.net/~patricia/misc/salaries02.txt

[5] Bender, Patricia. Retrieved June 30, 2003, from www.dfw.net/~patricia/misc/salaries03.txt

[6] National Basketball Association. Retrieved July 5, 2003, from www.nba.com/playoffs2001/

[7] National Basketball Association. Retrieved July 5, 2003, from www.nba.com/playoffs2002/

[8] National Basketball Association. Retrieved July 5, 2003, from www.nba.com/playoffs2003/

[9] Olsauskas, Dalius. Retrieved July 2, 2003, from http://ok.Kalnieciai.lt/nba9900/results/playoffs.htm

[10] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis 4th Edition*. Prentice Hall. 1998. Upper Saddle River, NJ.