A Multivariate Analysis of Basketball Player Attributes

Anum Damani

March 14, 2025

Motivation & Research Questions

In many sports, players are evaluated based on their physical and performance attributes. In basketball, specifically, the height, weight, and skills of a player play a crucial role in their success on the court.

The main research questions of this analysis are:

- How effectively can Principal Component Analysis (PCA) reduce the dimensionality of the data while preserving as much variance as possible, and what underlying relationships are revealed by the principal components?
- How effectively can Canonical Correlation Analysis (CCA) reduce the dimensionality of the data while maximizing correlation, and what underlying relationships can be determined through the canonical variates?

Description of Data Set

The dataset was obtained from Kaggle (See References). This dataset contains basketball player statistics across several NBA seasons, from 1996 to 2023. The original dataset contains 12,844 observations and 21 variables. In this dataset, each observation represents a basketball player. There are no missing values. For this analysis, the non-numeric variables in the dataset were excluded, specifically the name of the player, team abbreviation, name of the college the player attended, name of the player's birth country, draft year, draft round, draft number, and NBA season. After checking the correlations between all numeric variables, the variables with very low correlations were excluded; specifically, age, net_rating, and ts_pct (shooting efficiency) were removed. Therefore, I proceed with the following 10 numeric variables:

- player_height: Height of player (in centimeters)
- player_weight: Weight of player (in kilograms)
- gp: Number of games played in the season
- pts: Average number of points scored
- reb: Average number of rebounds
- ast: Average number of assists
- oreb_pct: Percentage of offensive rebounds

- dreb_pct: Percentage of defensive rebounds
- usg_pct: Percentage of team plays utilized by player
- ast_pct: Percentage of goals assisted

Hence, the resulting dataset has 12,844 observations and 10 numeric variables. Figure 1 is a correlation plot of the 10 numeric variables.

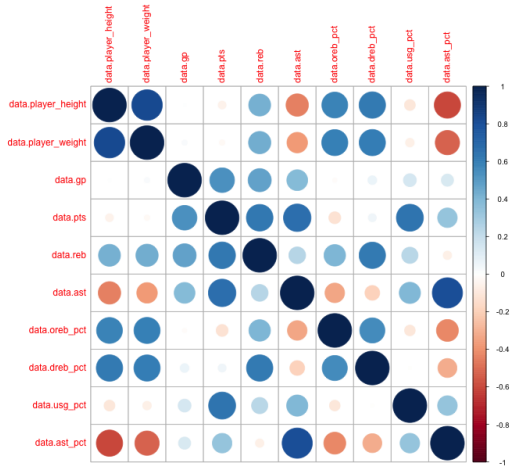

Figure 1: Correlation Plot

Principal Component Analysis

Analysis

Principal Component Analysis (PCA) is a dimensionality-reduction method that aims to preserve as much variation in the dataset as possible. It transforms the original, correlated variables into a smaller set of uncorrelated variables (principal components), which capture the variance-covariance structure of the original data. This allows for more interpretability of relationships between variables. It is appropriate for this dataset because there are moderate to strong correlations among the 10 variables. For example, it is clear that:

- player height is highly, positively correlated with player weight
- player height is moderately, negatively correlated with assists
- points are moderately, positively correlated with rebounds
- assists are moderately, positively correlated with points

These 10 variables can be condensed into a smaller set (PCs) using this method. Therefore, PCA is helpful for determining underlying relationships within the data.

Results & Interpretation

Prior to analysis, the data was standardized. Below is Figure 2, a scree plot, which shows a clear elbow formed by PC1 and PC2, suggesting that the first two principal components explain the most variation in the data. It is clear that the proportion of variance explained drops after the

second principal component. PC1 explains 39.15% of the variance, and PC2 explains 29.44% of the variance. Together, PC1 and PC2 explain 68.58% of the variation in the data. Based on this scree plot, I proceed with only the first two PCs. Figure 3 is a plot of the first two PCs, which appears elliptical.

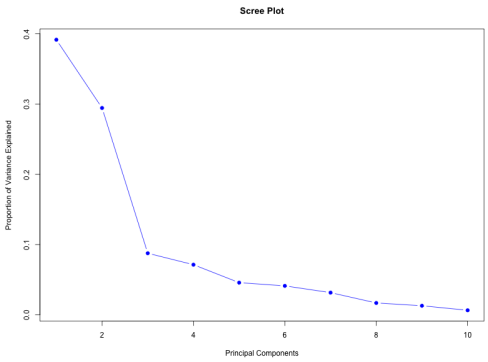

Figure 2: Scree Plot

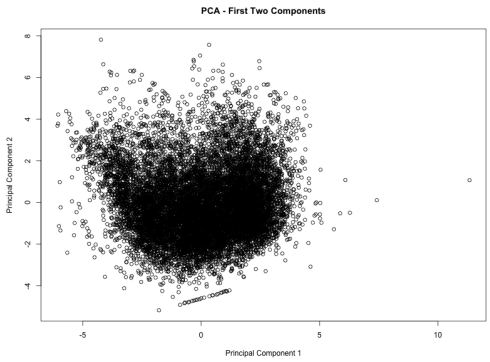

Figure 3: Plot of First Two PCs

Table 1: PC Loadings and Correlations for PC1 and PC2

Table 2: PC Loadings for PC1 and PC2

Variable	PC1 Loadings	PC2 Loadings
player_height	**-0.4398**	-0.1235
player_weight	**-0.4216**	-0.1518
gp	0.0604	-0.3724
pts	0.1536	**-0.5089**
reb	-0.1782	**-0.4923**
ast	0.3464	-0.3418
oreb_pct	-0.3781	-0.1104
dreb_pct	-0.3521	-0.2402
usg_pct	0.1552	-0.3335
ast_pct	0.3986	-0.1543

Table 3: Corr Between PCs & Variables

Variable	PC1 Corr	PC2 Corr
player_height	**-0.8701**	-0.2119
player_weight	**-0.8343**	-0.2605
gp	0.1196	-0.6389
pts	0.3040	**-0.8730**
reb	0.3526	**-0.8447**
ast	0.6853	-0.5864
oreb_pct	-0.7481	-0.1894
dreb_pct	-0.6966	-0.4121
usg_pct	0.3071	-0.5722
ast_pct	0.7886	-0.2649

Table 2 shows the PC Loadings for PC1 and PC2, and Table 3 shows the correlations between the PCs and original variables.

The PC1 loadings reveal that player height and player weight contribute to PC1 the most because they have the highest absolute loadings, which are -0.4398 and -0.4216, respectively. In addition, player height and player weight are strongly, negatively correlated with PC1, with correlations -0.8701 and -0.8343, respectively. Therefore, more negative PC1 values correspond to taller and heavier players. Conversely, more positive PC1 values correspond to shorter and less heavy players. Also, ast_pct has a moderate absolute loading (0.3986) and a strong, positive correlation with PC1 (0.7886); this means that higher ast_pct values contribute to higher, positive PC1 values. In addition, oreb_pct and dreb_pct have moderate PC1 loadings (-0.3781 and -0.3521, respectively) with strong, negative correlations with PC1 (-0.7481 and -0.6966, respectively). So, higher rebound percentages correspond to lower PC1 values. Finally, ast has a moderate PC1 loading (0.3464)

with a moderate, positive correlation (0.6853), which means that higher number of average assists correspond with higher PC1 values.

The PC2 loadings reveal that points and rebounds contribute to PC2 the most because they have the highest absolute loadings, which are -0.5089 and -0.4923, respectively. Also, points and rebounds are strongly, negatively correlated with PC2, with correlations -0.8730 and -0.8447, respectively. Therefore, more negative PC2 values correspond to basketball players that are successful at scoring points and rebounding. Conversely, more positive PC2 values correspond to basketball players that are less successful at scoring points and rebounding. It can be noted that the variable gp has a moderate loading (-0.3724) with moderate, negative correlation with PC2 (-0.6389), indicating that higher amounts of games played corresponds with lower PC2 values.

Canonical Correlation Analysis

Analysis

Canonical Correlation Analysis is a dimensionality reduction method that is helpful for analyzing the relationships between two sets of variables. Dataset includes 2 types of variables: physical attributes and performance attributes. CCA provides summaries for the relationship between physical attributes and performance attributes, while preserving the key information. It finds canonical variate pairs (linear combinations of original variables) that maximize correlation between two variable sets.

Results & Interpretation

Let $Z^{(1)}$ be the set of standardized variables corresponding to physical attributes of basketball players, specifically height and weight. Let $Z^{(2)}$ be the set of standardized variables corresponding to the performance attributes of basketball players, specifically number of games played, points, rebounds, assists, percentage of offensive rebounds, percentage of defensive rebounds, percentage of team plays utilized, and percentage of goals assisted.

$$Z^{(1)} = \begin{bmatrix} Z_1^{(1)} \\ Z_2^{(1)} \end{bmatrix} = \begin{bmatrix} \text{player\_height} \\ \text{player\_weight} \end{bmatrix}, Z^{(2)} = \begin{bmatrix} Z_1^{(2)} \\ Z_2^{(2)} \\ Z_3^{(2)} \\ Z_4^{(2)} \\ Z_5^{(2)} \\ Z_6^{(2)} \\ Z_7^{(2)} \\ Z_8^{(2)} \end{bmatrix} = \begin{bmatrix} \text{gp} \\ \text{pts} \\ \text{reb} \\ \text{ast} \\ \text{oreb\_pct} \\ \text{dreb\_pct} \\ \text{usg\_pct} \\ \text{ast\_pct} \end{bmatrix}$$

Here, $p = 2$ and $q = 8$, so there are $min(p, q) = 2$ sample canonical correlations and sample canonical variate coefficient vectors. The first canonical correlation ρ_1* is 0.8001144 and the second canonical correlation ρ_2* is 0.1565184. The first canonical correlation ρ_1* is the highest possible correlation between the weighted sums of variables in sets $Z^{(1)}$ and $Z^{(2)}$. The first pair of canonical variates convey a substantial amount of information about the relationship between the two sets. The second canonical correlation ρ_2* is much lower than ρ_1*, which means that the second pair of canonical variates may not convey much information about the relationship between the two sets.

The first pair of sample canonical variates $(\hat{U}_1, \hat{V}_1)$ are:

$$\hat{U_1} = -0.6686568z_1^{(1)} - 0.3749867z_2^{(1)},$$
$$\hat{V_1} = 0.03252165z_1^{(2)} - 0.03547987z_2^{(2)} - 0.21348676z_3^{(2)} - 0.09957098z_4^{(2)} - 0.27273937z_5^{(2)} - 0.35273886z_6^{(2)} - 0.03752319z_7^{(2)} + 0.45272971z_8^{(2)}$$

The second pair of sample canonical variates $(\hat{U_2}, \hat{V_2})$ are:

$$\hat{U_2} = 1.624352z_1^{(1)} - 1.716102z_2^{(1)},\ \hat{V_2} = -0.07898288z_1^{(2)} + 0.23610403z_2^{(2)} - 0.14592824z_3^{(2)} - 0.10950441z_4^{(2)} - 0.75388364z_5^{(2)} - 0.05859607z_6^{(2)} - 0.20675359z_7^{(2)} - 0.85298614z_8^{(2)}$$

As stated above, the first canonical pair conveys a substantial amount of information about the relationship between the two sets (with $\rho_1* = 0.8001144$), whereas the second canonical pair does not convey much information ($\rho_2* = 0.1565184$). I proceed by focusing on the first canonical pair $(\hat{U_1}, \hat{V_1})$. The coefficients of $\hat{U_1}$ and $\hat{V_1}$ can be analyzed. The coefficients of $\hat{U_1}$ reveal that player height contributes more to $\hat{U_1}$, with player height being slightly more dominant than player weight. Looking at the absolute values of the coefficients of $\hat{V_1}$, it is clear that the percentage of goals assisted ($z_8^{(2)}$), percentage of defensive rebounds ($z_6^{(2)}$), percentage of offensive rebounds ($z_5^{(2)}$), and average number of rebounds ($z_3^{(2)}$) contribute most to $\hat{V_1}$.

Table 4: Corr between $Z^{(1)}$ and $\hat{U_1}, \hat{V_1}$

$Z^{(1)}$ Variables	$\hat{U_1}$	$\hat{V_1}$
$z_1^{(1)}$ $(player\_height)$	-0.98	-0.78
$z_2^{(1)}$ $(player\_weight)$	-0.93	-0.74

Table 5: Corr between $Z^{(2)}$ and $\hat{U_1}, \hat{V_1}$

$Z^{(2)}$ Variables	$\hat{U_1}$	$\hat{V_1}$
$z_1^{(2)}$ (gp)	-0.01	-0.01
$z_2^{(2)}$ (pts)	0.05	0.06
$z_3^{(2)}$ (reb)	-0.45	-0.60
$z_4^{(2)}$ (ast)	0.44	0.54
$z_5^{(2)}$ $(oreb\_pct)$	-0.62	-0.77
$z_6^{(2)}$ $(dreb\_pct)$	-0.64	-0.80
$z_7^{(2)}$ (usg_pct)	0.09	0.12
$z_8^{(2)}$ (ast_pct)	0.60	0.75

Table 4 shows the correlations between $Z^{(1)}$ and $\hat{U_1}$. The physical attributes, player_height and player_weight, are strongly correlated with $\hat{U_1}$, with correlations -0.98 and -0.93, respectively. The strong correlations between the physical attributes and $\hat{U_1}$ align with the interpretation of the coefficients stated above. Therefore, player_height and player_weight strongly impact $\hat{U_1}$, the canonical variate representing physical attributes.

Table 5 shows the correlations between $Z^{(2)}$ and $\hat{V_1}$. It is clear that dreb_pct, oreb_pct, and reb are moderately to strongly, negatively correlated with $\hat{V_1}$, with correlations -0.80, -0.77, and -0.60, respectively. In addition, ast_pct is strongly, positively correlated with $\hat{V_1}$, with correlation 0.75. This aligns well with the interpretation of the coefficients stated above. The percentage of defensive rebounds, percentage of offensive rebounds, and average number of rebounds contribute strongly impact $\hat{V_1}$, the canonical variate representing performance attributes.

$$R_{11} - \text{sampleCov}(\tilde{z}^{(1)}) = \begin{bmatrix} 0.046 & -0.081 \\ -0.081 & 0.145 \end{bmatrix}$$

$$R_{22} - \text{sampleCov}(\tilde{z}^{(2)}) = \begin{bmatrix} 0.00 & -0.01 & 0.01 & 0.00 & 0.03 & 0.01 & 0.01 & 0.03 \\ -0.01 & 0.03 & -0.02 & -0.01 & -0.10 & -0.02 & -0.03 & -0.08 \\ 0.01 & -0.02 & 0.02 & 0.01 & 0.09 & 0.02 & 0.02 & 0.08 \\ 0.00 & -0.01 & 0.01 & 0.00 & 0.03 & 0.01 & 0.01 & 0.03 \\ 0.03 & -0.10 & 0.09 & 0.03 & 0.36 & 0.07 & 0.09 & 0.31 \\ 0.01 & -0.02 & 0.02 & 0.01 & 0.07 & 0.01 & 0.02 & 0.06 \\ 0.01 & -0.03 & 0.02 & 0.01 & 0.09 & 0.02 & 0.02 & 0.08 \\ 0.03 & -0.08 & 0.08 & 0.03 & 0.31 & 0.06 & 0.08 & 0.26 \end{bmatrix}$$

$$R_{12} - \text{sampleCov}(\tilde{z}^{(1)}, \tilde{z}^{(2)}) = \begin{bmatrix} 0.00 & -0.01 & 0.01 & 0.00 & 0.02 & 0.00 & 0.01 & 0.02 \\ -0.00 & 0.01 & -0.01 & -0.00 & -0.04 & -0.01 & -0.01 & -0.03 \end{bmatrix}$$

Next, the matrices of errors of approximation are created using the first canonical pair. The first canonical pair effectively summarizes (reproduces) the intra-set correlations in R_{12} due to small error values (closer to zero). However, there are larger values for the errors for R_{11} and R_{22}, which means that the first canonical pair do not effectively summarize (reproduce) the sampling variability in the original $Z^{(1)}$ and $Z^{(2)}$ variable sets.

The proportion of total standardized sample variance in $Z^{(1)}$ explained by $\hat{U}_1$ is 90.48%. The proportion of total standardized sample variance in $Z^{(2)}$ explained by $\hat{V}_1$ is 27.85%. The proportion of total standardized sample variance in $Z^{(1)}$ explained by $\hat{U}_2$ is 9.52%. The proportion of total standardized sample variance in $Z^{(2)}$ explained by $\hat{V}_2$ is 8.97%.

The following two hypothesis tests are conducted to better understand the significance of the canonical correlations. These tests are performed at the $\alpha = 0.05$ significance level.

	Test 1	**Test 2**
H_0	$\Sigma_{12} = 0, (\rho_1^* = \rho_2^* = 0)$	$\rho_1^* \neq 0, \rho_2^* = 0$
H_1	$\Sigma_{12} \neq 0$	$\rho_2^* \neq 0$

Table 6: Hypothesis Tests

The first test is the likelihood ratio test. Since the Bartlett's test statistic (13440.38) is greater than the critical value (26.29623), we reject H_0 at the $\alpha = 0.05$ significance level. So, this suggests that at least one canonical correlation is not equal to zero. Since $\rho_1* = 0.8001144$ is high and much larger than $\rho_2* = 0.1565184$, ρ_1* must be already statistically significant and is defining the direction of the null hypothesis.

Since the null hypothesis of the first test is rejected, the second hypothesis test can be performed at the $\alpha = 0.05$ significance level. The result of the second test reveals that the test statistic (318.4094) is greater than the critical value (14.06714), which means we reject H_0 at the $\alpha = 0.05$ significance level. This indicates that the second canonical correlation is statistically significant.

Therefore, the rejection of both null hypotheses suggests that both ρ_1* and ρ_2* are significant (nonzero). Both ρ_1* and ρ_2* capture the relationship between the physical attributes and performance attributes. Therefore, the relationship between physical attributes and performance attributes is statistically significant.

Conclusion

Therefore, the PCA and CCA analyses clearly addressed the research questions regarding effectiveness of the methods and the ability to reveal underlying relationships. Both PCA and CCA share some similarities. PCA and CCA have a common goal of dimensionality reduction. PCA reduces from 10 dimensions to 2 dimensions (principal components), explaining 68.58% of the variance. CCA reduces data from 10 dimensions to 2 dimensions (2 canonical variate pairs), with one pair having a stronger canonical correlation than the other.

In addition, both PCA and CCA have a common goal of enhancing interpretability. PCA transforms the original, correlated variables into a smaller set of uncorrelated variables (principal components), which capture the variance-covariance structure of the original data. CCA finds canonical variate pairs (linear combinations of original variables) that maximize correlation between two variable sets. Both PCA and CCA transform original variables into linear combinations of uncorrelated variables, which are called PCs in PCA and canonical variates in CCA.

However, it is important to highlight the differences between PCA and CCA, as seen by the analysis and results above. The main goal of PCA is to maximize variance. PC1 explains 39.15% of the variance, and PC2 explains 29.44% of the variance (68.58% together). The player_height, player_weight, ast_pct, oreb_pct, dreb_pct, and ast variables contribute most to PC1. Also, the pts, reb, and gp variables contribute most to PC2.

On the other hand, the main goal of CCA is to maximize correlation. After interpreting the correlations between $Z^{(1)}$ and U_1, it is clear that player_height and player_weight have the strongest correlations with U_1. After analyzing the correlations between $Z^{(2)}$ and V_1, it is evident that dreb_pct, oreb_pct, ast_pct, and reb have the strongest correlations with V_1. The proportion of total standardized sample variance in $Z^{(1)}$ explained by U_1 is 90.48%, while the proportion in $Z^{(2)}$ explained by V_1 is 27.85%. The canonical correlations are $\rho 1* = 0.8001144$ and $\rho_2* = 0.1565184$, and through the hypothesis tests, it is clear that both $\rho 1*$ and ρ_2* are statistically significant. Therefore, the relationship between physical and performance attributes is statistically significant.

References

Dataset link: https://www.kaggle.com/datasets/justinas/nba-players-data