# A Study on the Dynamic Relationship between Performance Scoring and Outcomes in Sports Competitions

Haitao Li [1, *], Zhijun Li [2], Yuqi Jin [2]

[1] School of Mechanical Engineering, Hefei University of Technology, Hefei, China, 230009

[2] School of Computer Science and Information Engineering (School of Artificial Intelligence), Hefei University of technology, Hefei, China, 230009

* Corresponding Author Email: 2021216866@mail.hfut.edu.cn

**Abstract.** In this paper, the relationship between kinetic energy and match results in tennis matches was explored by using principal component analysis (PCA) and the entropy weighting method. By analyzing the match data of Wimbledon 2023, the paper describes the data processing methods in detail, including the processing of missing data and the application of the MICE algorithm. It was found that there is a correlation between changes in kinetic energy during matches and match success, thus providing practical guidance for tennis coaches and players in adjusting their strategies. In addition, the role of kinetic energy in determining sports outcomes was analyzed, demonstrating the validity of the methodology employed. This study provides a new research idea and framework for the field of sports science.

**Keywords:** MICE algorithm, PCA, Entropy Weight Method, Tennis game strategy.

## 1. Introduction

In today's highly competitive sports, in-depth understanding and analysis of athlete performance have become central to improving athleticism, training programs, and strategic decision-making. With the development of big data and quantitative analytics, researchers and coaches are focusing more on utilizing scientific methods to assess athletes' performance and potential. Athletes' performance is influenced by various static factors like physical fitness, technique, and tactics, as well as dynamic factors including psychological state, opponent's strength, and momentum. Therefore, it is of great significance to construct a comprehensive evaluation system, that can fully reflect the performance of athletes under different conditions, to optimize the training methods and improve the performance.

To solve the problem of kinetic analysis of tennis matches, we use principal component analysis (PCA) and go through the entropy weighting method to optimize and determine the weights of the factors. PCA, as a data dimensionality reduction tool, can extract the most important few principal components from multiple variables, which is used to reveal the simple structure hidden behind the complex data. For example, Wan, G. (2019) demonstrated how to improve the accuracy of PCA in data analysis through an improved PCA with maximum entropy method, which is crucial for us to understand the complexity of race data [1].

Meanwhile, the entropy weighting method, as an objective assignment method to determine the weight of each indicator by calculating the entropy value of the indicator, is widely used in multi-indicator decision analysis. For example, Zhu, Y. (2020) emphasized the effectiveness of the entropy weighting method in decision-making in his research, showing how to deal with a large amount of data by the entropy weighting method, to obtain more information in the decision-making process [2].

Combining these two methods, we can more accurately identify and analyze the key factors affecting the results of tennis matches. As shown in Xu, X. (2017), the troubleshooting approach through information entropy and relative principal component analysis provides a novel way to process and parse the data, which is especially important for us to analyze the kinetic energy changes in tennis matches [3].

Further, Ye, C (2018) provided an innovative analysis of network structure through entropy component analysis, and its methodology is equally applicable to understanding complex interactions and dynamics in tennis matches [4]. In addition, Liang-Chen, W. (2010) demonstrated the effectiveness of this methodology in processing and analyzing large-scale datasets by combining PCA and entropy weighting to provide a comprehensive assessment of the economic development potential of multiple cities, which provides us with an important reference for analyzing tennis matches [5].

Through an extensive literature review [6-10], we found that the entropy weight method and PCA have not only been successfully applied in many fields, such as economic analysis and troubleshooting but also provide a solid theoretical and methodological foundation for kinetic energy analysis in tennis matches. These studies demonstrate the powerful ability of the entropy weight method and PCA in analyzing complex data sets and revealing the deep structure behind the data, which provides us with effective tools and methods to study the kinetic energy in tennis matches.

In this study, we use principal component analysis (PCA) and entropy weight method to deeply explore the relationship between match kinetic energy and outcome by collecting and preprocessing tennis match data. Our goal is to reveal a possible correlation between changes in kinetic energy and match success, exploring the impact it may have during the match. We hope that this finding will provide some practical guidance for tennis coaches and players' strategy adjustment and that it provides a new methodology and idea for related research within the field of sports.

## 2. Data processing

### 2.1. Missing data statistics

We collected detailed data on matches at Wimbledon over the year 2023 and analyzed them initially. We found that there are several variables with missing data in the collected data, and we used different schemes to deal with different missing data. The missing data are serve_width (direction of serve), serve_depth (depth of serve), return_depth (depth of return), and speed__mph (speed of serve), and the specific missing rates are shown in Figure 1 below.

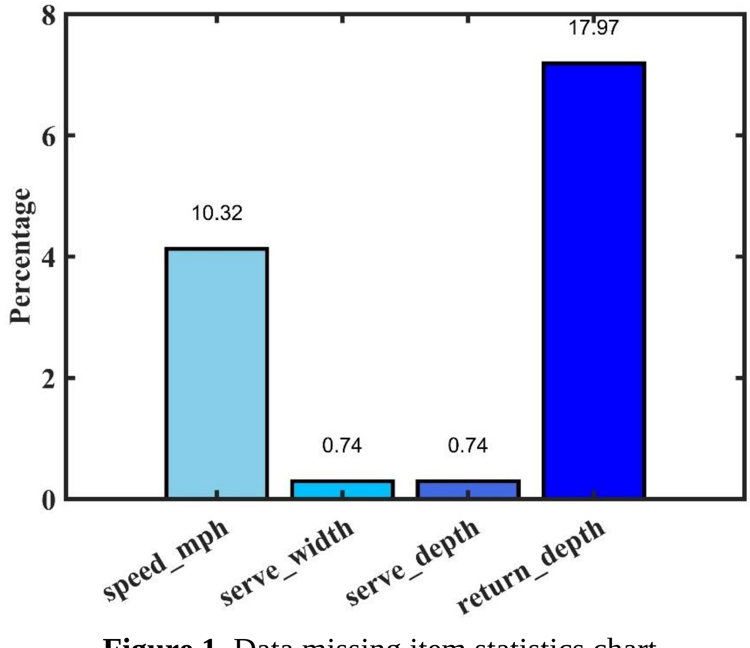


**Figure 1.** Data missing item statistics chart

Among them, the missing rate of serve_width and serve_depth is less than 1%, which is a character variable, and through statistical analysis, the direction of serve and the depth of serve tend to be stable in most cases, so here we choose to deal with the way of filling in the plurality.

The data volume of return_depth is as high as 17%. However, we found through statistical analysis, that this parameter only has two values, and cannot accurately describe the value, basically has no effect on the problem. Therefore, here we decided to delete this variable and disregard the influence of this factor.

The missing rate of speed_mph is about 10%, but when we categorize the data according to the number of races, we find that the missing data are concentrated in the two races 2023-wimbledon-1310 and 2023-wimbledon-1311. So, we chose to delete the data from these two games and interpolate the remaining data.

### 2.2. MICE linear interpolation

The MICE algorithm is widely used in statistical analysis for its flexibility in dealing with missing values in data. The algorithm constructs conditional distributions by iteratively using existing data, after we exclude the data from the games 2023-wimbledon-1310 and 2023-wimbledon-1311. The missing rate in the statistical other game fields is about 3%. The data is linked to physical strength and fluctuates up and down, so we use linear regression to go for prediction and predictive interpolation. Compared with the traditional single interpolation method, MICE enhance the reliability of statistical inference by creating multiple datasets, thus reflecting the uncertainty introduced due to missing data in the final analysis.

In the MICE algorithm, the whole process can be described by the following formalization in Figure 2:

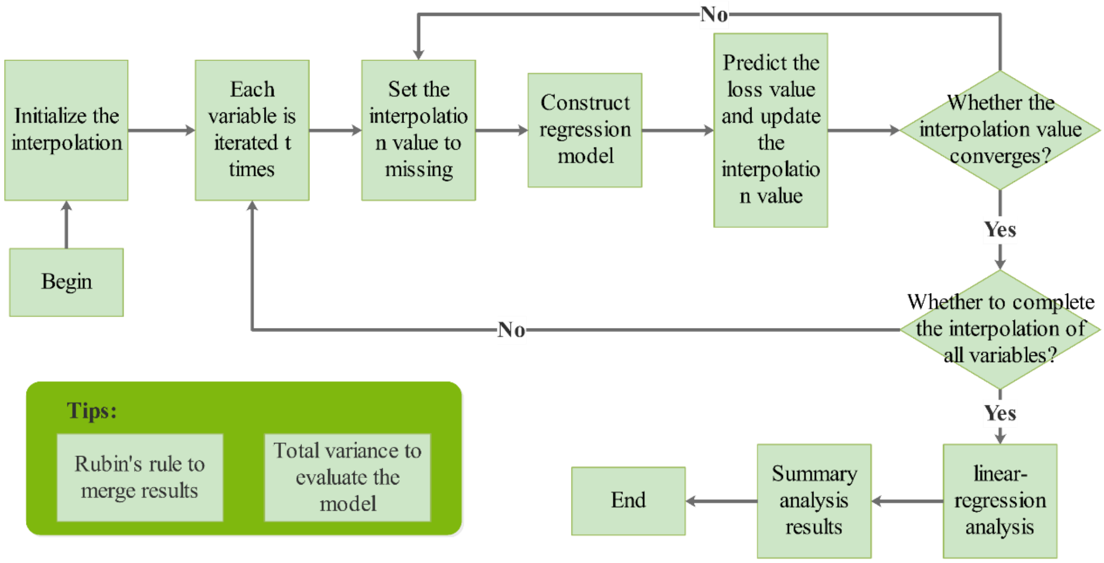


**Figure 2.** MICE flow chart

At the same time, to better evaluate the explanatory ability of the model and the variability of the data, we introduced the total variance to evaluate the model:

$$\mathrm{Var}(\theta) = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{1}$$

Where $\bar{U}$ is the average of the variances of the parameter estimates from the analysis of the $m$ complete data and $B$ is an estimate of the variance between the parameter estimates.

The value of $Var(\theta)$ was calculated to be 144.48, indicating a strong and excellent interpretation using this method. Figure 3 below shows the overall interpolation, with the blank area in the figure showing the excluded data for the two games. Also use the red scatter to indicate interpolation points and the blue points to indicate data points.

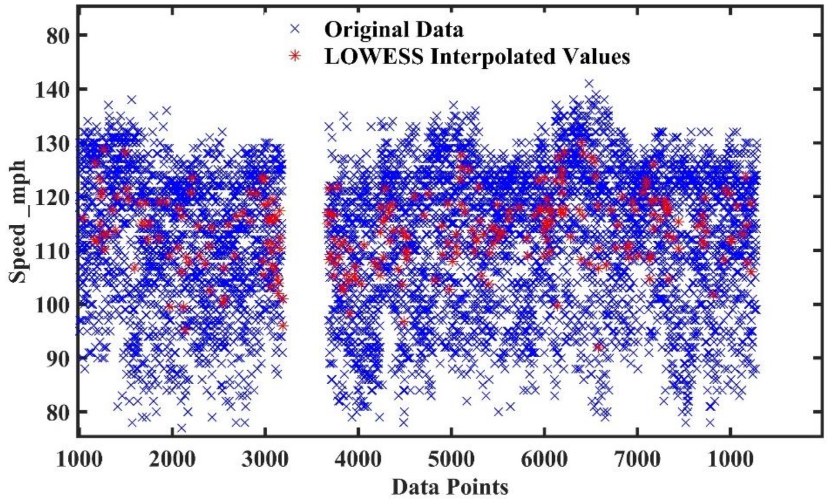


**Figure 3.** Interpolated scatter plot

### 2.3. Data precheck

Before unfolding the principal component analysis, in order to check whether the processed data can support the subsequent analysis, we here carry out the KMO test and Bartlett's sphericity test on the data, and the test results are shown in Table 1 below.

**Table 1.** KMO and Bartlett's Test

| KMO and Bartlett's Test | |
|---|---|
| KMO Measure of Sampling Adequacy | 0.6336 |
| Bartlett's Test of Sphericity Approx. Chi-Square | 667415.24 |
| P | 0.00 |

The result of the KMO test shows that the value of KMO is 0.6636, meanwhile, the result of Bartlett's spherical test shows that the significance P-value is 0.000, which presents significance at the level of rejection of the original hypothesis, i.e., it shows that there is a correlation between the variables, and that the results of the principal component analysis are valid and referable.

## 3. Exploring Temporal Variations in Athlete Performance

### 3.1. The establishment of the simulation model

To reduce the data complexity and retain the features of the original data as much as possible, we first use PCA to reduce the dimensionality of the data.

We imported the normalized tennis player datasets of the two players into Python separately, calculated their covariance matrix $R$ and eigenvectors $\lambda_n$, and computed the cumulative $y_j$ contribution $b_j(j = 1,2, \dots, n)$ and $y_1$, $y_2$,..., $y_n(p \leq n)$ of each principal component by the following formulae contribution rate $\alpha_p$.

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{n} \lambda_k} \qquad (j = 1, 2, \dots, n) \tag{2}$$

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{n} \lambda_k} \quad (p \leq n) \tag{3}$$

We extracted the components with eigenvalues greater than 1 according to Kaiser's criterion, and got 9 principal components in total, with 72.44% cumulative variance contribution of P1 and 72.57% cumulative variance contribution of P2. Therefore, here, we can consider that the models of Athlete 1 and Athlete 2 are similar. Among them, the variance contribution of each principal component of P1 is shown in Figure 4 below.

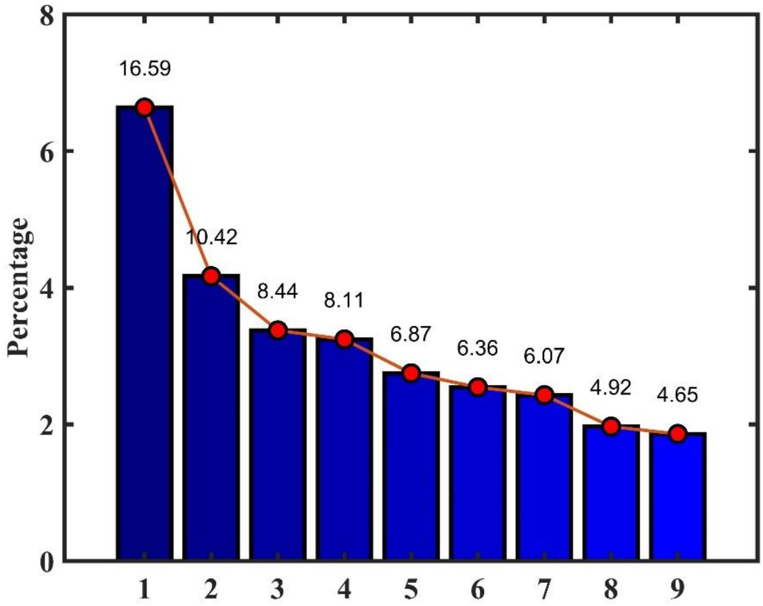


**Figure 4.** Statistical chart of variance contribution rate

### 3.2. Factor Load coefficient and principal component Classification

To show the importance of the hidden variables in each principal component, we plotted a table of factor loading coefficients. Figure 5 below shows the loading coefficients of each parameter in principal component 1 of P1. It reflects the degree of influence of different parameters on the factor.

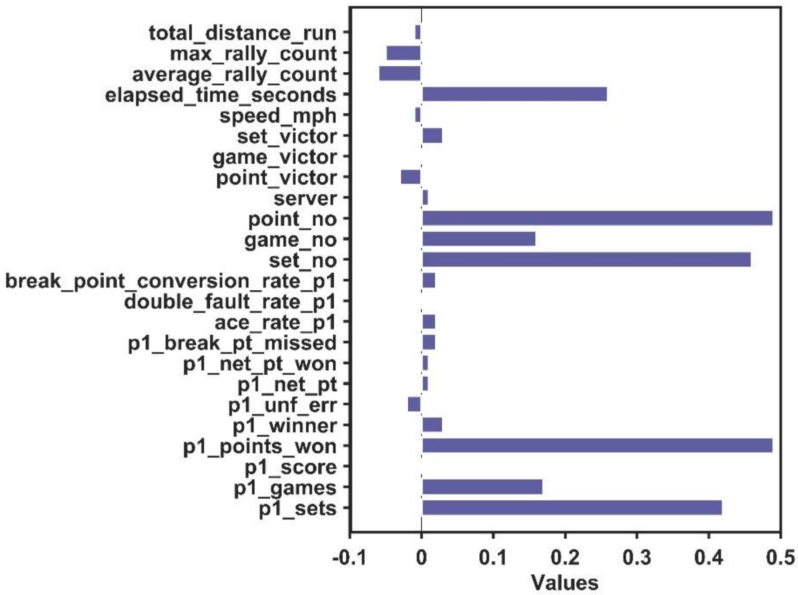


**Figure 5.** Load matrix numerical diagram

By analyzing the factor loading coefficients, we classified them into nine categories shown in Figure 6 and selected the three variables with the highest loadings as their strong positive loadings, i.e., the three factors with the greatest impact.

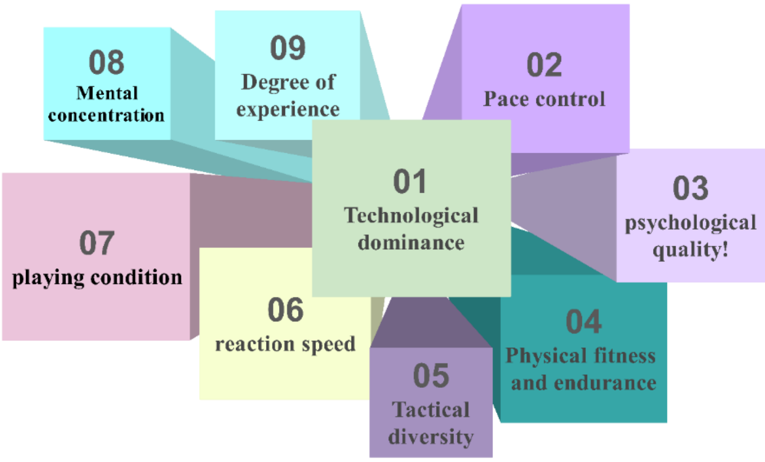


**Figure 6.** Schematic diagram of nine principal components

### 3.3. Scoring model based on entropy weight method and PCA

#### 3.3.1. Computed weight matrix

Since the load matrix obtained after the principal component analysis above has been standardized, only the absolute value of the load is needed here to eliminate the effect of the load being negative. Firstly, the calculation of specific gravity and information entropy is carried out, in the calculation of the coefficient of variation and weight through the relevant formula, and finally, the weight is adjusted by combining the variance contribution ratio:

$$\mathrm{W}_{j,\,final} = W_j * VC_j \tag{4}$$

$W_{j,final}$ is the final weight of the $j$th original variable after considering the variance contribution rate, and $VC_j$ is the variance contribution rate of the $j$th principal component. The array of weights for P1 obtained from the final calculation is shown in Table 2 below:

**Table 2.** Weights array table

| Number & Rate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 34.70% | 14.52% | 8.00% | 6.95% | 5.45% | 9.46% | 10.79% | 6.32% | 3.82% |

#### 3.3.2. Principal Component Factor Score Calculation

For each principal component score $C_i$, the score can be calculated by the following formula:

$$C_i = w_{i1} \times x_1 + w_{i2} \times x_2 + \ldots + w_{in} \times x_n \tag{5}$$

Where $\omega_{ji}$ is the loading of variable $x_j$ on the ith principal component. $x_j$ is the value of variable $j$. The total score factor $F$ calculation can be calculated by the following formula:

$$F = k_1 \times C_1 + k_2 \times C_2 + \ldots + k_m \times C_m \tag{6}$$

$k_i$ is the weight of the $i$th principal component, calculated by dividing the variance contribution of that principal component by the sum of the variance contributions of all retained principal components.

The final score is obtained by weighting and summing each principal component by the principal component contribution share going. In order to explore the effect of the model, we take the data of the first game to calculate the momentum (score) of p1 and p2, and plot the image7 as follows, where the blue line represents the score of players 1 and the red line represents the score of players 2. We find that their momentum sometimes fluctuates strongly up and down, but generally shows an upward trend, which is also consistent with our usual situation of the game getting more and more anxious the later it gets.

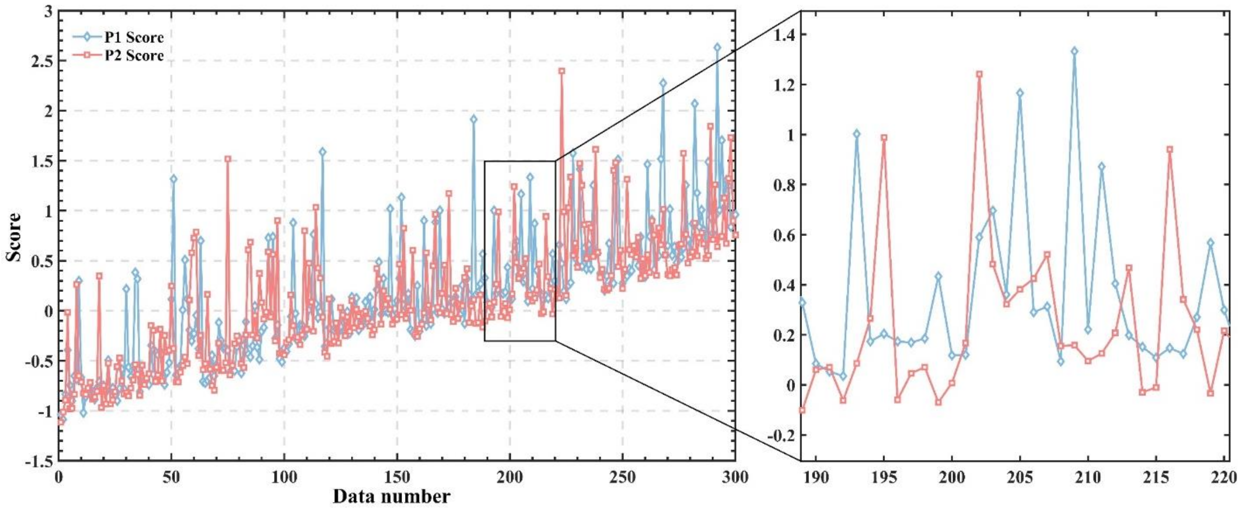


**Figure 7.** Dynamic score chart

Also, time is described by the time period corresponding to the period between the two data points. Simply based on the momentum score shown above, it is possible to visualize how good or bad a player is performing in that time period. In addition, we also notice that there are sudden fluctuations above the graph, and these fluctuations are due to the presence of high-impact balls such as ACEs at this point in time.

### 3.3.3. The Role of Momentum in Determining Sports Outcomes

Based on the above model, we introduced the score difference of two athletes to compare with our rating difference to analyze whether there is interpretability in the course of the game. Figure 8 below shows a composite line plot of score difference versus rating difference, with red and blue representing score difference and rating difference, respectively.

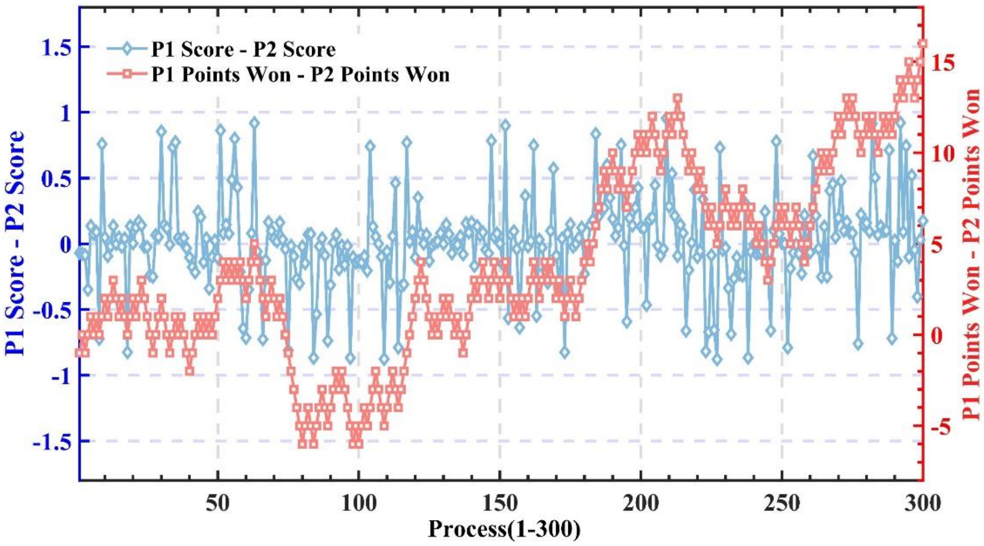


**Figure 8.** Comparison chart of the score difference and rating difference

Observing the trend of its change, we find that the bias of the scoring difference can reflect the trend of the score difference, and the trend of the change between the two scores is the same. This shows that our model unearths the fact that athletes' strength will be affected by the actual situation during the game, and the result of the game is not random, but related to the momentum of both sides during the game, which is our score.

When one side has more momentum, its score slowly grows. When the momentum starts to weaken and the score difference positive or negative changes, it can be noticed that its scoring situation also starts to gradually decline. This is enough to reflect that there is a connection between the relationship between winning and losing and is related to the player's current set of states, which make up his momentum and thus affect the outcome of the match.

In addition, we analyzed whether there is some relationship between momentum difference and score difference by calculating different correlation coefficients, which are shown in Table 3 below. It indicates that there is a moderate correlation between the two variables. We can assume to a certain extent that there is a relationship between volatility and players' success. And the reason for the low correlation coefficient may be due to the lag between the two variables or there are other possible influences.

**Table 3.** Phase relation table

| Correlation index | Values |
|---|---|
| Spearman rank correlation coefficient method | 0.327 |
| Kendall rank correlation | 0.412 |
| ANOVA | 248.88 |

## 4. Conclusion

First, we adopt an interpolation strategy for data with different missing variables. In this, numerical variables were interpolated using the MICE algorithm. The processed data were split into P1 athletes and P2 athletes to facilitate subsequent analysis and processing. Before using PCA, we carried out the KMO test and Bartlett sphericity test on the data and then used principal component analysis for dimensionality reduction after passing it, nine principal components were extracted, and the total variance contribution rate of both P1 and P2 was around 72%. Subsequently, we established the entropy weighting method-principal component analysis method model to get the weights of each index and the score model. We visualized the performance of players in different periods with line graphs and found that the performance of players in the game showed dynamic changes over time, and the higher the score of players in the same period, the better their performance.

To assess the role of momentum, we compared the difference in scores between the two players with the difference in ratings we obtained and then calculated various rank correlation coefficients between the two data as well as the F-statistic for the analysis of variance (ANOVA). The data showed a positive correlation between fluctuations and player success, so we concluded that "momentum" has a role to play in the game, proving the feasibility of the method.

This paper provides a research idea and framework for use in the field of sports science. This study reveals the role of momentum in determining exercise outcomes, which can help solve problems in areas such as coaching training program scheduling.

## References

[1] Wan G, Maung C, Schweitzer H. Improving the accuracy of principal component analysis by the maximum entropy method [C]//2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019: 808 - 815.

[2] Zhu Y, Tian D, Yan F. Effectiveness of entropy weight method in decision-making [J]. Mathematical Problems in Engineering, 2020, 2020: 1 - 5.

[3] Xu X, Wen C. Fault diagnosis method based on information entropy and relative principal component analysis [J]. Journal of Control Science and Engineering, 2017, 2017.

[4] Ye C, Wilson R C, Hancock E R. Network analysis using entropy component analysis [J]. Journal of Complex Networks, 2018, 6 (3): 404 - 429.

[5] Liang-chen W, Xue-feng Z, Hui W. Method of synthetic evaluation based on the principal component analysis and entropy weight[C]//2010 International Conference on Computer Application and System Modeling (ICCASM 2010). IEEE, 2010, 8: V8 - 312 - V8 - 315.

[6] He R, Hu B, Yuan X T, et al. Principal component analysis based on non-parametric maximum entropy[J]. Neurocomputing, 2010, 73(10-12): 1840 - 1852.

[7] Jenssen R. Kernel entropy component analysis [J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 32 (5): 847 - 860.

[8] Vásquez-Correa M C, Rodas H L. A robust approach for principal component analyisis [J]. arXiv preprint arXiv: 1903.00093, 2019.

[9] Deng X, Tian X. Entropy principal component analysis and its application to nonlinear chemical process fault diagnosis [J]. Asia-Pacific Journal of Chemical Engineering, 2014, 9 (5): 696 - 706.

[10] Long Y, Shi X, Xiao Z, et al. An Outlier Fault Diagnosis Method Based on Principal Component of Entropy Weight[C]//2021 2nd International Conference on Computer Engineering and Intelligent Control (ICCEIC). IEEE, 2021: 158 - 161.