

Tennis Player Momentum Analysis and Match Flow Prediction Model Based on PCA

Ruikang Liang^{1,*}, Liancai Shen¹, Xiyu Zeng², Yiwang Zhong³

¹School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou, China, 510006

²School of Materials and Energy, Guangdong University of Technology, Guangzhou, China, 510006

³School of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China, 518060

* Corresponding Author Email: 13676191364@163.com

Abstract. Momentum is a fluctuating phenomenon that occurs during the course of a game when the seemingly dominant side scores consecutive points or influences the course of the entire game. And as tennis is a popular sport all over the world, exploring the effects and mechanisms of athletes' momentum on match changes and outcomes in tennis matches can provide reasonable match advice to athletes. To this end, this paper systematically establishes a prediction model based on tennis players' momentum analysis and match direction. In this paper, this text first cleaned the data for the 2023 Wimbledon tennis tournament obtained from the American Collegiate Mathematical Modeling Contest website and supplemented the missing values using cubic Hermite interpolation. The processed data is then analyzed using principal component analysis, where a linear combination of each principal component is defined as a momentum value, thus quantifying the momentum and creating a mathematical model between the indicator and the momentum. Finally, a momentum plot of the player is created to visualize the relationship between the player's momentum and the outcome of the race.

Keywords: Tennis, Momentum, PCA, Three Times Hermit Interpolation Method.

1. Introduction

With the comprehensive popularization of sports, tennis is developing into one of the most ornamental and popular sports in the world. Every year, the Association of Tennis Professionals (ATP) holds professional tennis tournaments in over 30 countries around the world. As representatives of the highest level of tennis, Wimbledon, Australian Open, US Open, and French Open have been loved by a large number of fans.

But at the same time, with the continuous development of informatization and digital technology, it is not new to introduce advanced Internet high-tech technologies such as big data into the sports field, predict the results of relevant sports events, and assist in scientific training and event preparation.

Studying the important indicators of players on the court and accurately analyzing the trends and results of tennis matches provide useful insights for further in-depth excavation of the winning laws of professional tennis programs^[1]. Traditional tennis match outcome prediction models only consider the direct effect of indicators on the outcome. For example, logistic regression prediction^[2], decision tree prediction model^[3] and Bayesian prediction^[4] all have certain shortcomings.

In this paper, 12 metrics that have a large impact on player momentum during tennis matches are selected. The relationship between the momentum of the players of both sides of the match and the indicators after dimensionality reduction is constructed separately by using PCA, and then the trends and results of the match are analyzed by obtaining the momentum of the players of both sides of the match according to the constructed model.

2. Model construction

2.1. Description of the variable

To quantify the variation of athlete performance with the number of games, this text extracted a total of twelve feature indicators. These twelve characteristic indicators can be divided into two categories and four subcategories. All the data for the twelve metrics were obtained from the official website of American College Mathematical Modeling: <https://www.comap.com/contests>.

this text believe that a player's scoring ability is mainly attributed to their own technical and psychological factors^[5].

Because the serve is the fastest stroke and the game revolves around scoring points^[6], this text believe that technical factors can be further divided into serving performance and receiving performance.

At the same time, psychological factors should not be underestimated, as two different psychological states will inevitably arise in advantageous and disadvantageous situations. Therefore, this text believe that psychological factors can be divided into dominant and disadvantaged states.

Considering the convenience and rigor of subsequent analysis, this text have uniformly transformed the selection of indicators into numerical indicators. The twelve indicators this text have selected are shown in the table below.

And this text first standardized the selected indicators and found that all indicators have numerical values, so all indicators are retained. In the end, this text selected twelve indicators, including Aces in one game^[7] etc, as the descriptive system for players' momentum. Please refer to the table 1 below for details.

Table 1. Key notations used in this section

Variable	Meanings
X_1	Aces in one game
X_2	Double points won in one game
X_3	Triple points won in one game
X_4	Break serves
X_5	Breaking points
X_6	Holding serves
X_7	Successful serves in one game
X_8	The number of sets ahead in this match
X_9	The number of games ahead in this set
X_{10}	Speed of serve (miles per hour; mph)
X_{11}	Number of shots during the point
X_{12}	Player's distance ran during point (meters)
M_1	Player1's momentum value
M_2	Player2's momentum value

2.2. Data preprocessing

By conducting descriptive statistical analysis and relevant checks on the dataset, the outliers are first cleared. This text have found that the competition with the number 2023-wimbledon-1303 has been suspended and postponed for 24 hours. Considering that "momentum" is a continuously changing quantity, and delaying for 24 hours will inevitably lead to varying degrees of impact on the "momentum" on the human body, therefore, in the case of sufficient experimental data, it is chosen to directly delete the data with this number.

Then, this text found that there are missing the values of speed mph/serve width/serve depth and return depth in the dataset, so this text need to interpolate the missing values.

This text first analyze each variable and other related features. This text believe that variables with NA values, such as speed mph/serve width/serve depth and return depth, are closely related to their own physical strength.

Therefore, each match can be divided into regions which are calculated using Hermite(3D) interpolation to deal with missing values. This interpolation method can be used for smooth interpolation between given data points.

Taking 2023-wimbledon-1301 as an example, the interpolated results for some of the indicators are shown in Figures 1 and 2 below:

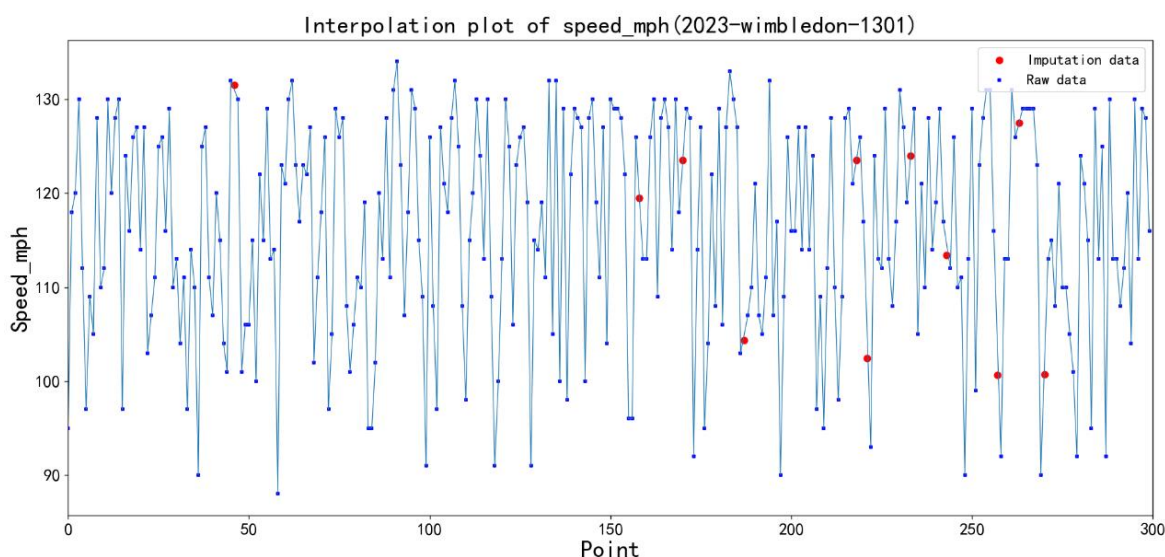


Figure 1. Interpolation plot of speed_mph

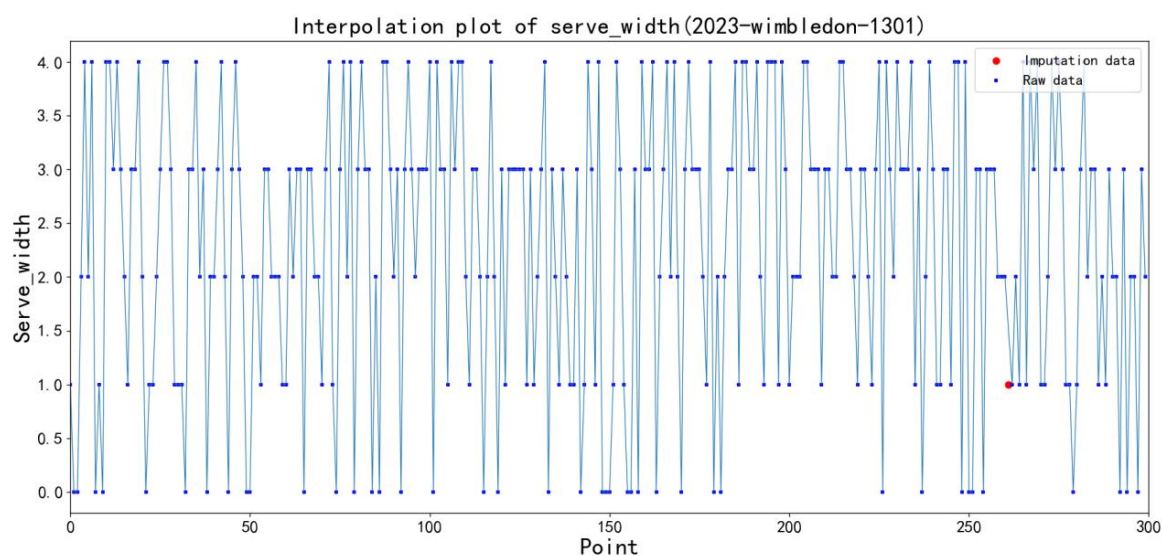


Figure 2. Interpolation plot of serve_width

Meanwhile, during the processing, this text found that the speed/mph values in the 2023-wimbledon-1310 and 2023-wimbledon-1311 numbered areas were both NA so that had no reference significance. Therefore, this text chose to remove this numbered area without affecting the overall judgment. The above is the entire process of our data preprocessing.

2.3. Description of the model

In the process of modeling and analyzing mathematical problems encountered in daily life, this text often encounter the problem of establishing a mathematical model that requires too many input variables, which greatly increases the difficulty of solving and analyzing the model. To solve this

problem, people have proposed a method called principal component analysis, which uses a small number of new variables to replace the original variables while retaining as much information as possible. Principal Component Analysis (PCA) is a common multivariate statistical method in statistics, which is often used for dimensionality reduction of data and can be used in financial analysis^[8], medical research^[9] and other fields. It can extract and integrate multiple indicators into a few new indicators, which are linear combinations of the original indicators and have weak correlation with each other. It can retain most of the information in the original data, helping us greatly reduce the difficulty of model analysis^[10].

Generally speaking, when using principal component analysis for data dimensionality reduction, assuming we have p indicators and n samples, we can construct a sample matrix x with a size of n times p .

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p) \quad (1)$$

Assuming we want to extract the data into a new set of variables $z_1, z_2, \cdots, z_m (m \leq p)$ each new variable can represent a linear combination of the original indicators.

$$z_1, z_2, \cdots, z_m (m \leq p) \begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases} \quad (2)$$

Based on each principal component this text obtain, this text need to observe and analyze the coefficients before each indicator. If the coefficients before each indicator are relatively large, it indicates that the indicator has a significant impact on the principal component; If the coefficient in front of the indicator is relatively small, it means that the indicator has a relatively small impact on the principal component.

The modeling process in this paper is shown in Figure 3 below.

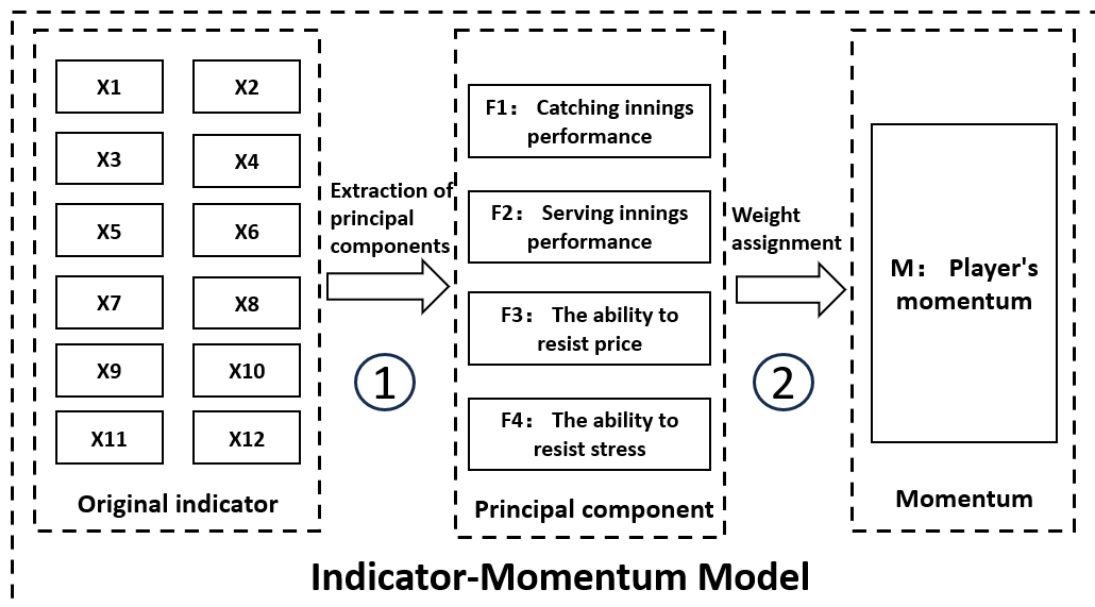


Figure 3. Model diagram

3. Results

3.1. KMO and Bartlett tests

As shown in the table, the KMO test value is 0.615, which is greater than 0.5. This indicates that the selected twelve indicators have a certain correlation and weak partial correlation, making them suitable for factor analysis. The Bartlett sphericity test result $P=0.000$, less than 0.01, indicates that the correlation coefficient matrix is not an identity matrix and rejects the assumption that each variable is independent. Both test results indicate that the twelve selected indicators are suitable for factor analysis.

3.2. Extract raw variables

As shown in the table 2, out of the twelve indicators, 10 indicators are all above 0.7, which fully indicates that these indicators can be fully expressed by common factors.

Table 2. KMO and Bartlett tests

KMO sampling suitability quantity		0.615
Bartlett's sphericity test	Approximate chi square	577142
	Freedom	66
	Significance	0.000

Although the original information extracted from the number of breaking points and the number of net wins in the current session is less than 0.5, this text believe that these two indicators can fully represent the corresponding factors to a certain extent, so this text choose to keep them. Overall, these twelve indicators can be well expressed by common factors, meeting the requirements of factor analysis. The indicators and their corresponding public factors are shown in the table 3 below.

Table 3. Common factor variance

Variable	Meanings	Extract value
Z_1	Aces in one game	0.752
Z_2	Double points won in one game	0.910
Z_3	Triple points won in one game	0.707
Z_4	Break serves	0.962
Z_5	Breaking points	0.496
Z_6	Holding serves	0.790
Z_7	Successful serves in one game	0.933
Z_8	The number of sets ahead in this match	0.784
Z_9	The number of games ahead in this set	0.403
Z_{10}	Speed of serve (miles per hour; mph)	0.912
Z_{11}	Number of shots during the point	0.881
Z_{12}	Player's distance ran during point (meters)	0.902

3.3. The proportion of explainable variance

The initial eigenvalues of the first four factors of the twelve selected indicators are greater than 1, and the cumulative variance contribution rate reaches 78.603%. Although it is not higher than 80%, the main reason is that when extracting the original variables, this text left two indicators: the number of breaking points and the number of net wins in the current market. Therefore, this text believe that 78.603 meets the conditions combined with the following figure 4, so this text extracted the first three factors as the main factors, denoted as F_1 , F_2 , F_3 , and F_4 , respectively. The variance contribution rate of F_1 is 31.974.

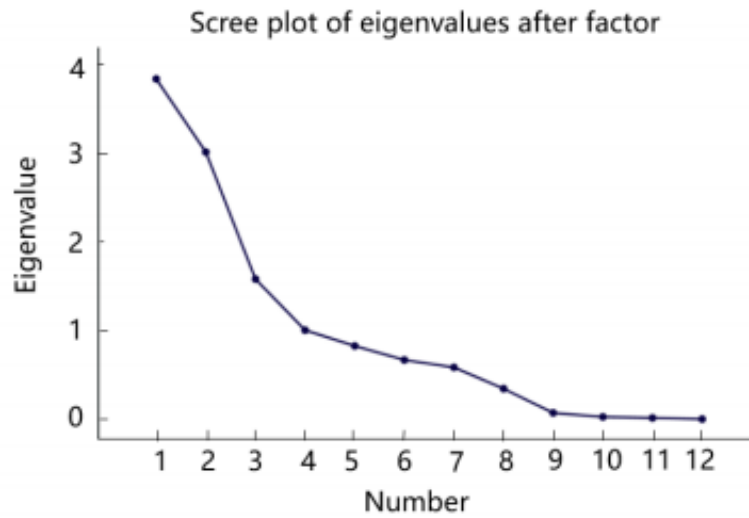


Figure 4. Gravel map

3.4. Common factor load factor

As shown in the table, F_1 has a relatively high loading capacity in indicators such as the number of serve breaks (X_4), number of break points (X_5), number of hitting rounds (X_{10}), and running distance (X_{11}). F_2 has a relatively high load on indicators such as the number of ACE balls played (X_1), the number of consecutive 2 points scored (X_2), the number of consecutive 3 points scored (X_3), the number of successful serve attempts (X_7), and serve speed (X_9). The number of ACE shots fired by F_3 (X_1), the number of saved break points (X_6), the number of sets currently taken (X_8), and the number of winning games in the game (X_{12}) are all indicators with high load. F_4 has a relatively high loading capacity in ACE balls (X_1) and currently has a high lead count (X_8).

Therefore, this text named F_1 as the ability to perform in receiving innings, F_2 as the ability to perform in serving innings, F_3 as the ability to carry pride, and F_4 as the ability to withstand pressure.

3.5. Extract raw variables

Based on the common factor loading coefficient, this text calculated the corresponding factor score coefficients and finally obtained the calculation formulas for each factor. The results are as follows:

$$\left\{ \begin{array}{l} F_1 = -0.59 X_1 + 0.154 X_2 - 0.02 X_3 + 0.259 X_4 + 0.172 X_5 + 0.83 X_6 - 0.006 X_7 \\ \quad + 0.087 X_8 - 0.008 X_9 + 0.298 X_{10} + 0.293 X_{11} - 0.006 X_{12} \\ F_2 = 0.462 X_1 + 0.297 X_2 + 0.311 X_3 - 0.004 X_4 - 0.008 X_5 - 0.279 X_6 + 0.092 X_7 \\ \quad + 0.1 X_8 - 0.003 X_9 - 0.016 X_{10} - 0.025 X_{11} + 0.101 X_{12} \\ F_3 = -0.262 X_1 + 0.009 X_2 - 0.01 X_3 - 0.094 X_4 - 0.078 X_5 + 0.484 X_6 + 0.32 X_7 \\ \quad - 0.115 X_8 - 0.045 X_9 + 0.123 X_{10} + 0.088 X_{11} + 0.302 X_{12} \\ F_4 = 0.270 X_1 - 0.060 X_2 - 0.155 X_3 + 0.062 X_4 - 0.020 X_5 + 0.017 X_6 - 0.39 X_7 \\ \quad + 0.751 X_8 - 0.422 X_9 + 0.029 X_{10} + 0.049 X_{11} + 0.002 X_{12} \end{array} \right. \quad (3)$$

Ultimately, the quantitative calculation formula for power was successfully derived:

$$M_{player1} = \frac{0.2864 F_1 + 0.19979 F_2 + 0.19115 F_3 + 0.10868 F_4}{0.78603} \quad (4)$$

$$M_{player2} = \frac{0.40893 F_1 + 0.17915 F_2 - 0.10352 F_3 + 0.10027 F_4}{0.79187} \quad (5)$$

3.6. Analyzing of the results

Based on the established regression formula for power M, this text substitute the data and calculate the real-time power maps figure 5.(a) and cumulative power maps figure 5.(b) of players 1 and 2 in each game of 2023-wimbledon-1701, which is a visual description of the player's performance.

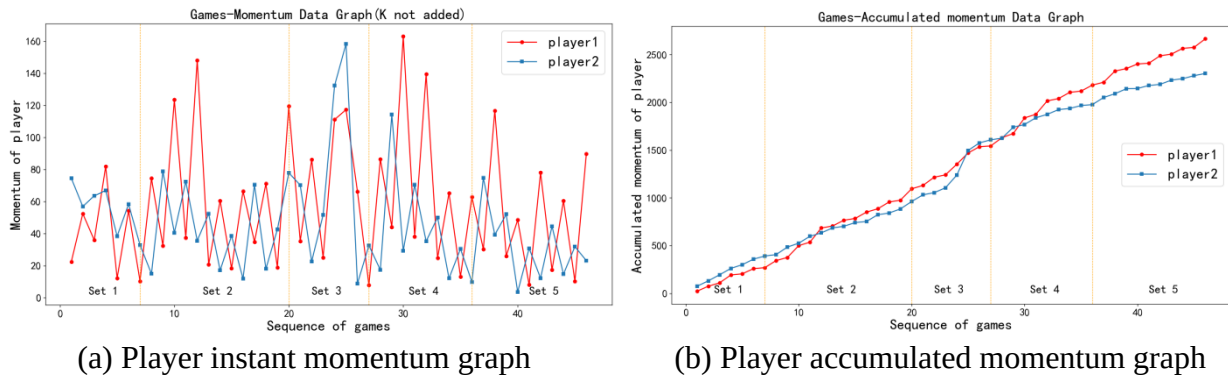


Figure 5. Player momentum graph

This text believe that dynamic momentum maps can fully quantify the real-time performance of players in Game units. It can be seen that in the real-time power map, the player's momentum value is a oscillation curve. According to tennis rules, the serving side will rotate on a per game basis. Therefore, the characteristic of oscillation in the real-time power map fully indicates that the scoring ability of the serving game is higher than that of the receiving game.

At the same time, the cumulative power map can fully quantify the direction of the situation throughout the game. If the cumulative momentum value of one side within a certain range continues to be higher than that of the other, it indicates that the higher side is more likely to win the game, and the intersection of the two curves can be considered as representing a change in the situation and a change in the winner of the game.

As a result, problem 1 was fully resolved, and this text utilized a performance evaluation system based on principal component analysis to quantitatively describe the athlete's momentum. And visualized the performance and competition direction of the players from the perspectives of single and cumulative momentum values.

4. Conclusions

In this paper, outlier data were cleaned up in the data preprocessing session while missing values were supplemented using cubic Hermite interpolation. The KMO and Bartlett tests were performed on the processed data, and the results showed that the 12 selected indicators were suitable for factor analysis. Principal component analysis was then used to integrate the four principal components, define the linear combination of each principal component as a momentum value, quantify the momentum and establish a mathematical model between the indicators and the momentum. Finally, a momentum plot of the player's game is made to help the player visualize and analyze the relationship between the player's momentum and the outcome of the game.

However, the factors affecting the momentum are far more than these, and the 12 indicators chosen in this paper are not able to fully cover the whole process of the game, failing to realize the comprehensive quantification of the momentum, and the model still has a lot of details that need to be improved.

References

- [1] Luo Wei quan, Zhang Lei. Research on the construction of winning factor model of professional tennis players [J]. Journal of Guangzhou Sports Institute, 2020, 40(03): 78-81.
- [2] MA Ji cai, CHEN Qing guo, LIANG Cairong, et al. Model construction and research analysis of match-winning factors of professional tennis players [J]. Bulletin of Sports Science and Technology Literature, 2023, 31(04): 66-68+118.
- [3] MENG Fanming, HUANG Wenmin. Quantitative analysis of winning factors of female tennis players based on decision tree method [J]. Journal of Jilin Institute of Physical Education, 2019, 35(05): 43-47.
- [4] HE Wensheng, ZHANG Liwei, ZHANG Liancheng. Technical analysis of winning factors of the top 3 male tennis players in the world [J]. Journal of Wuhan Institute of Physical Education, 2011, 45(09): 67-73.
- [5] Tang Linfang, Xu Xibin. Exploration on the improvement way of tennis serve technology [J]. Sports Goods and Technology, 2021(22): 35-36.
- [6] Xie Hanxi, Chen Qian, Li Jie, et al. A study on the differences in perceptual anticipation of tennis players under different visual masking paradigms [J]. Journal of Guangzhou Sports Institute, 2023, 43(05): 113-121.
- [7] MA Ji cai, CHEN Qing guo, LIANG Cairong, et al. Modeling and research analysis of match-winning factors of professional tennis players [J]. Sports Science and Technology Literature Bulletin, 2023, 31(04): 66-68+118.
- [8] Xie Xingguo, Bai Yujing. Construction of county financial comprehensive stress index based on principal component analysis--Taking Sunan County of Gansu Province as an example [J]. Gansu Finance, 2022(05): 49-54.
- [9] GUO Jiaxin, YANG Xiru, YAN Dong, et al. Research on Chinese medicine evidence of progressive gastric cancer based on principal component analysis and cluster analysis [J]. China Journal of Traditional Chinese Medicine Information, 2024, 31(04): 153-156.
- [10] Liu Qinlong. Reflections on the application of principal component analysis in sports research [J]. Journal of Shanghai Institute of Physical Education, 2003(05): 78-79+82.