Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Quantitative modelling of momentum in tennis based on factor analysis

Yiwei Ge *

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China, 730050

* Corresponding author: 17346876683@163.com

Abstract. Tennis, with its playful and elegant nature, has perennially captivated individuals of all ages. Beyond being a mere game, it represents a clash of momentum. In modern sports, the term "momentum" is used to describe a player's drive and momentum during a game. In the tennis match, it can affect the player's rhythm of the match and the result, is the key factor to determine the result of the match. But the quantification of momentum has no definitive reference in sport. This study attempts to propose a model that can be used to quantify the momentum of tennis players based on factor analysis. In this paper, data from the men's singles match at Wimbledon was first taken and subjected to missing and outlier detection and processing to ensure the reliability of this data. Secondly, the paper selected all the indicators that may be related to a player's momentum for factor analysis. In this paper, the proportion of variance contribution is used as the weight of each factor to calculate the composite score as the quantitative value of momentum. The analysis and modelling process in this paper could provide new insights into the understanding and development of the sport of tennis, offering a perspective and direction to quantization. The results of the factor analysis may provide valuable perceptions for coaches and athletes to help them better understand the game situation and develop strategies for future games.

Keywords: momentum; quantitative model; factor analysis.

1. Introduction

"Momentum is one of the most important factors in the game of tennis, it can make you bounce back from defeat or keep you calm in victory." So writes Alistair Higham in his book *Momentum: the hidden force in tennis*.

In the field of physics, momentum is defined as the product of an object's mass and its velocity when in motion. Interestingly, an object already in motion requires less force to maintain its movement compared to a stationary object of the same mass. This concept has been adopted in sports where "momentum" is used to describe the impact of the preceding events on subsequent ones. If a match or game is progressing in a certain direction, it is likely to continue unless there is a shift in momentum. The notion is that successful events build momentum for a player or team, increasing the likelihood of future success. This perception of momentum is commonly observed in sports, with commentators often referring to it, including in the sport of tennis [1]. In tennis matches, the term "momentum" refers to a player's psychological or physiological drive throughout a game, which can impact a player's rhythm. The existence of momentum may change the course of a match at some point, which has attracted the attention of many related scholars and sports practitioners around the world.

Although many studies examine if players in sports and especially in tennis benefit from a psychological or physiological drive (momentum), most of the research in tennis sports has focussed on the influence of momentum on point outcomes more than the determination of factors that may affect or react to players' momentum [2-3]. Therefore, this paper aims to explore the essential mechanism of "momentum". Focusing on the race indicators related to momentum factors, this study aims to explore the key "momentum" factor in the transition between victory and defeat, and to establish a comprehensive evaluation index model to quantify the momentum behind the game. The

entire process of modelling may provide new insights into the understanding and development of the sport of tennis, offering a new reference for the field of quantization of momentum.

2. Materials and Methods

2.1. Data Collection and Pre-processing

The data used in the modelling process of this paper mainly comes from the official website of COMAP(https://www.mathmodels.org/Problems/2024/MCM-C/Wimbledon_featured_matches.csv), which gives the datasets of each round of Wimbledon men's singles matches after the first two rounds of the 2023 Wimbledon. These datasets provide a complete record of the match data for each round of the tournament, and include key performance metrics, match outcomes, set information, and various other crucial data points. These data provide almost all the information which this paper needs to build the model. But before using, these data need to be checked first to avoid missing values or outliers that may cause errors in the calculation.

2.1.1. Missing value detection.

This paper uses python's pandas to detect and find that there are a lot of disk matches with many missing values on the following four indicators as shown in Table 1 below. Since all these missing data are from certain complete one-deck matches, and the individual matches are independent of each other, it is not possible to measure these missing values with data from other disk matches, and therefore there is no suitable method to fill in these missing values, and this paper directly excludes these missing values.

Table 1. Missing value detection

Variables	return_depth	serve_depth	serve_width	speed_mph
Description	The position in which a player returns the ball after an opponent's serve	The position in which a player serves the ball in the longitudinal position of the court	The position in which a player lands the ball in the lateral position of the court when serving	The speed of the ball
Quantity	>1000	54	54	752

2.1.2. Outlier detection.

Considering the possibility of anomalies in the data collected by the website, box plots were introduced to detect these outliers. Box plot theory does not require restrictions on the data itself, is not affected by extreme values, can visually characterize the discrete distribution of the data, and provides a criterion for identifying outliers, i.e., values greater than the upper bound set by the box plot or less than the lower bound set by the box plot are identified as outliers, where an outlier is often defined as a value less than Q1-1.5IQR or greater than Q3+1.5IQR (where Q1 is the first quartile, Q3 is the third quartile, and IQR is the interquartile spacing).

This paper performs IQR test on the data and use the box plot function in MATLAB to generate box plots and identify outliers. Through the box plot we can see that there are four groups of data with outliers (as shown in Figure 1), the other data indicators are not abnormal. And the paper develops the following analyses and treatments.

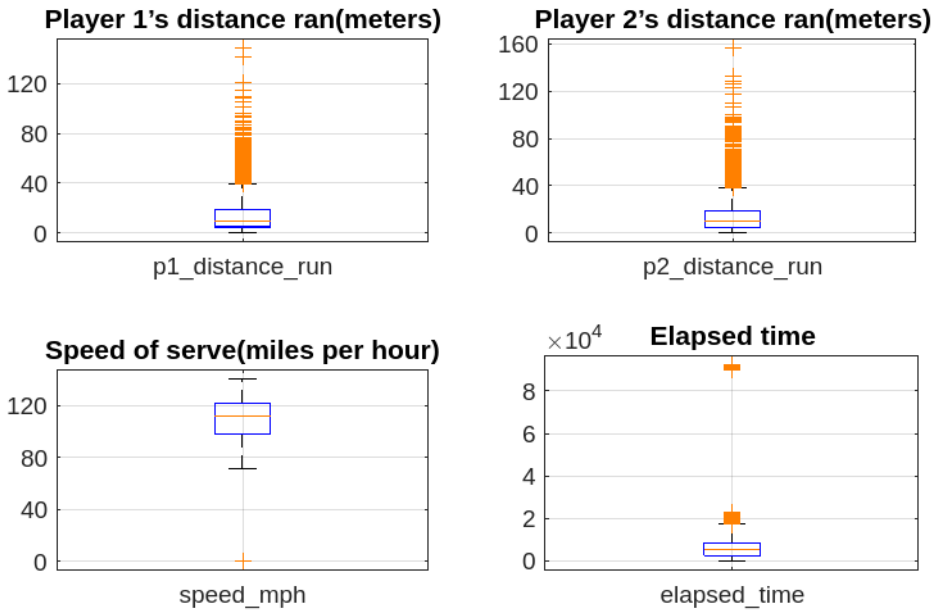

Figure 1. box plots

p1_distance_run&p2_distance_run (Player 1's & Player 2's distance run during point):

The maximum values for these two metrics are 148.723 meters and 156.856 meters, respectively, and these extremely large values could point to some very long rounds of play, which is something that can happen in high-level tennis matches, so they are not excluded.

speed_mph (speed of serve):

Some values of 0 on this indicator are labeled as outliers by the box plot, whereas in actual matches a serve speed of 0 means that the ball hit the net and landed on the ground. They are therefore consistent with reality and are not treated.

elapsed_time (time elapsed since start of first point to start of current point):

The outliers shown in the fourth subfigure can be seen to have significant deviations, and the lengths of time corresponding to these data are grossly out of line with reality. Therefore, for this anomalous dataset, this paper chose to exclude all data corresponding to the same game.

2.2. Introduction to factor analysis

Factor analysis is a statistical technique utilized to investigate the connections between observed variables and reveal hidden latent factors that account for these connections. Essentially, it seeks to simplify a dataset by pinpointing the shared factors that impact the observed variables. Factor analysis can be either exploratory or confirmatory. Exploratory factor analysis (EFA) is utilized to uncover the underlying structure of the data without any predetermined hypotheses, whereas confirmatory factor analysis (CFA) examines specific theories or models [4]. In this paper, the technique employed is exploratory factor analysis, which will be simply referred to as factor analysis in the latter part.

2.2.1. Mathematical modelling of factor analysis.

The mathematical model of factor analysis assumes that there are p original variables, denoted by x_1, x_2, ..., x_p, and to reduce the dimensionality of the data, each original variable is denoted by a linear combination of m (m<p) factors (F_1, F_2, ..., F_m), i.e.:

$$\begin{aligned} \boldsymbol{x_1} &= \boldsymbol{a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1} \\ \boldsymbol{x_2} &= \boldsymbol{a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2} \\ &\cdots\cdots \\ \boldsymbol{x_p} &= \boldsymbol{a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p} \end{aligned} \tag{1}$$

The matrix form is:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = (a_{ij})_{pm} \cdot \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} \tag{2}$$

The same can be expressed as:

$$X = AF + \varepsilon \tag{3}$$

where $\boldsymbol{F}$ is the common factor that appears in each variable's linear expression; $\boldsymbol{A}$ is the factor loading matrix, and a_{ij} denotes the loading of the i-th original variable on the j-th factor; $\boldsymbol{\varepsilon}$ denotes the fraction of the indicates the portion of the original variable that cannot be explained by the factor [5].

2.2.2. Main steps of factor analysis.

The main steps of factor analysis are shown in Figure 2 below:

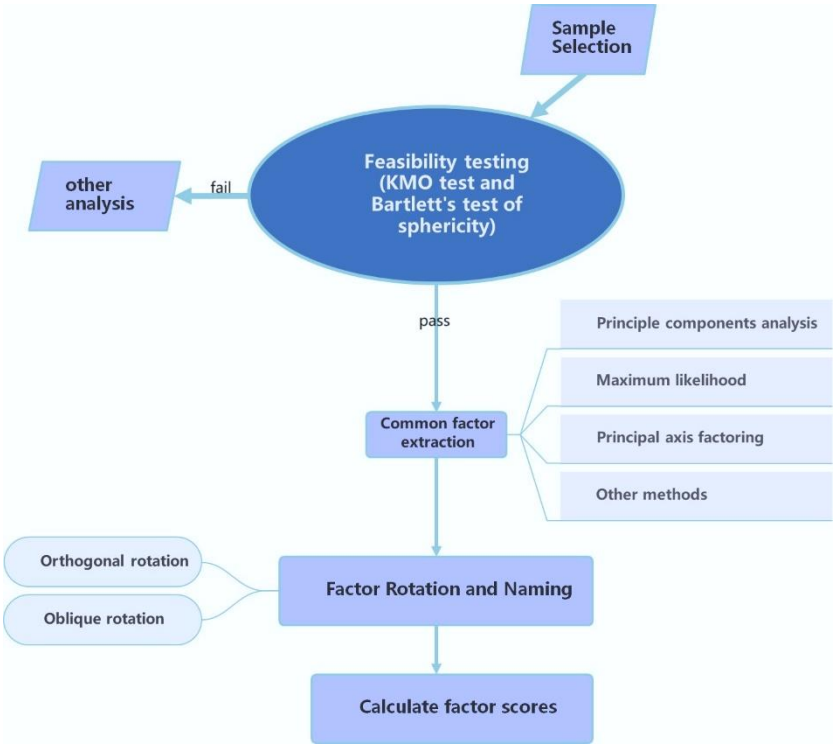

Figure 2. Flowchart of factor analysis

3. Modelling and solving

3.1. Indicator selection

Momentum is now widely taken to be defined as: a positive or negative change in cognition, physiology, affect, and behavior caused by a precipitating event or series of events that will result in a shift in performance in matches [6]. Accordingly, this paper chose men's singles tennis players as the

subjects of this study, using the match at Wimbledon 2023 as the data source for this modelling. This paper made some translations on the original dataset and extracted the following indicators (as shown in Table 2 below), which are closely related to the player's momentum, as the samples of this model.

Table 2. Selected indicators

Indicators	Description
ace	whether or not an ace was issued in this round (1 yes, 0 no)
distance_run	opponent's running distance minus that player's own running distance (multiplied by -1 if the player loses)
error_rate_of_serve_no.1	error rate of the first serve on the turn to serve
lead_score	current game leading score
net_pt_won	if the player won the point while at the net
rally_count	number of shots during the point (multiplied by -1 if the player loses)
server_pt	server or not (1 yes, 0 no)
speed_mph	speed of serve

The rationale for choosing these metrics as the sample for this modelling is as follows:

ace: Serving an ace can provide a significant confidence boost to the server. It demonstrates their ability to control the point right from the start and can give them a psychological edge over their opponent.

distance_run: Reflects the amount of activity of the player on the field. The relative length of the running distance may show whether the player is passive in the game or not [7].

error_rate_of_serve_no.1: The serve is completely under your control and not under the influence of your opponent. The miss on the first serve was a lost opportunity and a blow to the psyche. The error rate of the first serve is indicative of the player's physical exertion and mental state.

lead_score: Reflects a player's advantage on the scoreboard, which directly correlates to winning or losing the game and affects the player's psychological state. Leading in a competition may provide a psychological advantage [8].

net_pt_won: Shows a player's initiative and technical ability at the net, which is often a behavior that can be performed by aggressive players (scoring at the net tends to require a high level of agility and skill). Like the ace ball, this paper sees this action as characteristic of a high-momentum player.

rally_count: Reflects to some extent the player's dedication to winning the round. Players achieving a victory after a longer round creates an increase in psychological momentum. Except for ace ball of course, which has already been considered. Considering that the player's perception of the number of strokes is greater than that of the length of the match, and that there is a positive correlation between the two, this paper chose this metric instead of metric "elapsed_time".

server_pt: The player who serves has a much higher probability of winning the point/game. When players take their turn on the initiating side, their hearts will be more in the right place.

speed_mph: The speed of the serve reflects the player's offense, the higher the speed of the serve, the harder it is to be caught by the opponent; Fluctuations in the speed of the serve can reflect the player's fitness level and mental level to a certain extent.

3.2. Standardization of data

When conducting factor analysis, there may be differences between the selected sample data due to the dimensionality of the order of magnitude and scale effects. For example, indicator "distance_run" is measured in meters, while indicator "speed_mph" is measured in meters per second. Sample data therefore needs to be standardized to eliminate the effects of dimensionality and size on each indicator. The standardization of this paper is carried out with the help of the default Z-score standardization of

SPSS software, and the standardization can be carried out automatically by importing the sample data into SPSS.

3.3. Feasibility testing

To construct a valid factor model, this paper uses KMO test and Bartlett's test of sphericity to assess whether the original variables have feasibility and practicality. The KMO(Kaiser-Meyer-Olkin) test aims to test whether the original variables have feasibility and whether they can meet the requirements of factor analysis. Table 3 shows the results of Bartlett's test of sphericity and KMO test obtained by using SPSS software.

Table 3. KMO and Bartlett's Test

KMO Measure of Sampling Adequacy		0.665
Bartlett's Test of Sphericity	Approx. χ^2	94182.328
	df	28 28
	Sig.	0.000

As shown in Table 3, the values of KMO test and Bartlett's test of sphericity are 0.665 and 94182.328, respectively. Usually, when the KMO value is greater than 0.5 and the significance level of Bartlett's value is less than 0.01, it means that there is a high correlation between the variables. When the calculated KMO value is greater than 0.5, the data selected in this paper meets the prerequisite requirements of the factor analysis method, and thus can be analyzed using the factor analysis method [9].

3.4. Common factor extraction

In this paper, principal component analysis is used to extract the factors. Principal Component Analysis (PCA) is a commonly used factor analysis method for reducing the dimensionality of data and identifying the main structures in the data. Its basic idea is to map the original data to a new coordinate system through linear transformation, which makes the data as dispersed as possible under the new coordinate system, i.e., retaining the variance information of the data as much as possible. After processing and analyzing by SPSS software, the variance of all the original factors can be obtained. This paper makes a gravel plot containing the eigenvalues of the components of the data, as shown in Figure 3 below.

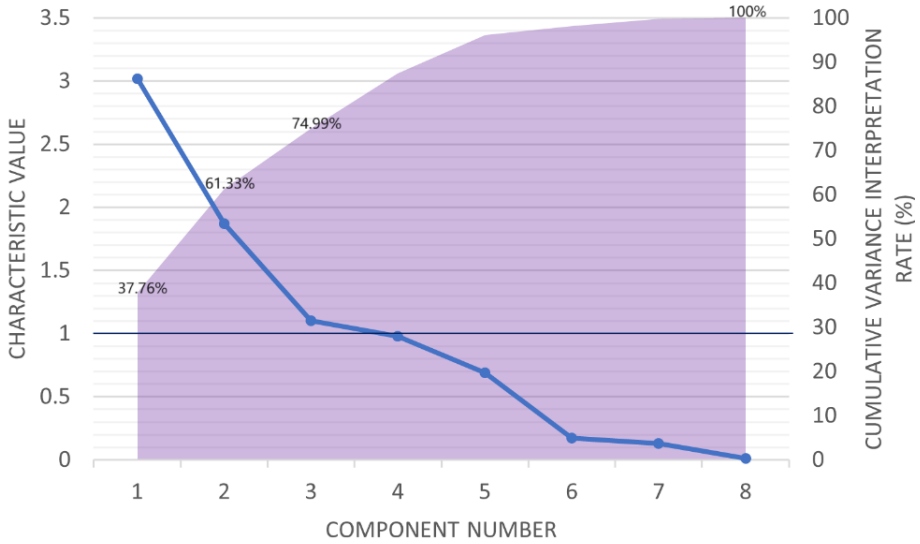

Figure 3. Scree plot

After calculating the variance of the common factor, it is known that there are 3 factors with eigenvalues greater than 1, and the cumulative degree of explanation is 74.99%. From the gravel plot, we can see that the eigenvalues from the 3rd component onwards drop very low, i.e. the 4th component is the "inflection point" of this fold. Before this one is the primary factor and after this one is the secondary factor. Therefore, these 3 factors are retained.

3.5. Common factor naming

Factor analysis requires that the extracted metrics have actual meanings. After extracting the common factors from the sample data, it is necessary to name each common factor according to the connotations it represents, since each common factor represents a different meaning. For this study, the metric nomenclature allows for a more nuanced understanding of the composition of the factors influencing momentum as the object of study. Among the components of each extracted male factor, if the factor loadings of a certain category of indicators are high, the male factors can be named according to the connotations of this large category.

This paper analyzed the components of the three identified common factors by using the principal factor analysis mode in the SPSS software to obtain the component matrices of the three common factors using statistical analysis of the three previously selected significant principal components. By rotating the initial factor loading matrices, the interpretation of the meanings can be done more easily. This paper uses the maximum variance method of Kaiser's normalization for rotation, which is the most used method of rotation, where the factors remain orthogonal but every effort is made to maximize the variance difference of the factors to facilitate the interpretation of the factors. This paper accordingly obtained the rotated component matrix and made the following matrix heat map (Figure 4).

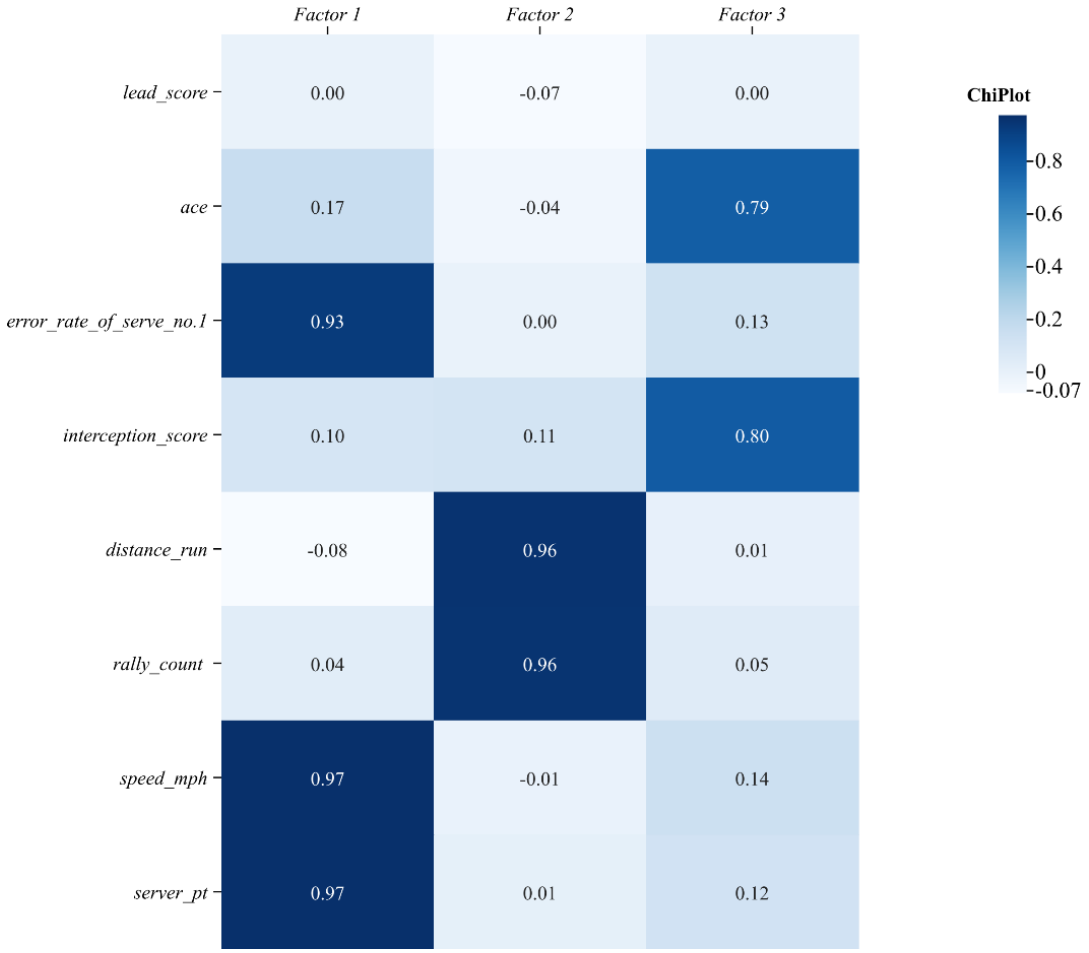

Figure 4. Matrix heat map

The higher the absolute value of the factor loadings, i.e. the darker the colour of the corresponding block in the heat map, which means that the factor has more information related to that variable. By looking at the matrix as above, we can conclude that:

For the first factor variable, first serve error rate, serve speed and whether to serve or not, these three indicators have larger loadings, which are closely linked to the serve. Therefore, this factor is defined as Serve Factor;

For the second factor variable, running distance and number of strokes have larger loadings, which are closely linked to the batting. Therefore, this factor is defined as Batting Factor;

For the third factor variable, interception score and irresistible ball have larger loadings, which belongs to the external factors that are beyond one's control. So, this factor is defined as External Force Factor.

It is worth noting that for indicator "lead_score", the loading values for all three factors are low, which means that this indicator played a minor role in this analysis. It can be inferred from this that professional tennis player's momentum is less disturbed by the score, especially in high-level tournaments like Wimbledon. However, the conclusion may not apply to ordinary players and tournaments, as they may not have undergone professional mental training.

3.6. Calculation of factor scores and composite score

The calculation of each common factor score is based on the factor score coefficient matrix and standardized raw variables [10]. This paper analyzed and estimated the obtained factor score coefficient matrices according to the regression analysis function of the SPSS software. The factor scores coefficient matrices of each factor processed by the SPSS software are shown in Table 4:

Table 4. Factor Score Coefficient Matrix

Indicators (X_i)	*Serve Factor*	*Batting Factor*	*External Force Factor*
lead_score(X_1)	-0.002	-0.039	0.002
ace(X_2)	-0.077	-0.062	0.638
error_rate_of_serve_no.1(X_3)	0.343	0.007	-0.057
net_pt_won(X_4)	-0.102	0.019	0.651
distance_run (X_5)	-0.016	0.517	-0.031
rally_count (X_6)	0.022	0.516	-0.017
speed_mph(X_7)	0.358	0.006	-0.052
server_pt(X_8)	0.362	0.018	-0.071

The above table contains the score coefficients of several indicators selected in this paper, and this paper use x_1 to x_8 to represent the values of these 8 related indicators and get the following factor score model:

$$\begin{aligned} \boldsymbol{F_{serve}} &= \boldsymbol{-0.002x_1 - 0.077x_2 + 0.343x_3 - 0.102x_4 - 0.016x_5 + 0.022x_6 - 0.358x_7 + 0.362x_8} \\ \boldsymbol{F_{batting}} &= \boldsymbol{-0.039x_1 - 0.062x_2 + 0.007x_3 + 0.019x_4 + 0.517x_5 + 0.516x_6 + 0.006x_7 + 0.018x_8} \\ \boldsymbol{F_{external}} &= \boldsymbol{0.002x_1 + 0.638x_2 - 0.057x_3 + 0.651x_4 - 0.031x_5 - 0.017x_6 - 0.052x_7 - 0.009x_8} \end{aligned} \quad (4)$$

After calculating the score function of the three public factors mentioned above, this paper combined the proportion of variance contribution of the three public factors interpreted by SPSS software as the weight of each factor, and then applied the weighted average method to sum up to calculate the final composite score. The following is the expression for the composite score based on this weighting:

$$F_{momentum} = \frac{(37.756\%F_{serve} + 23.577\%F_{batting} + 13.656\%F_{external})}{74.998\%} \quad (5)$$

Where $F_{momentum}$ is the composite of the three extracted common factors, i.e., the momentum this study is trying to quantify. Figure 5 below illustrates the composition of the momentum score derived from this equation.

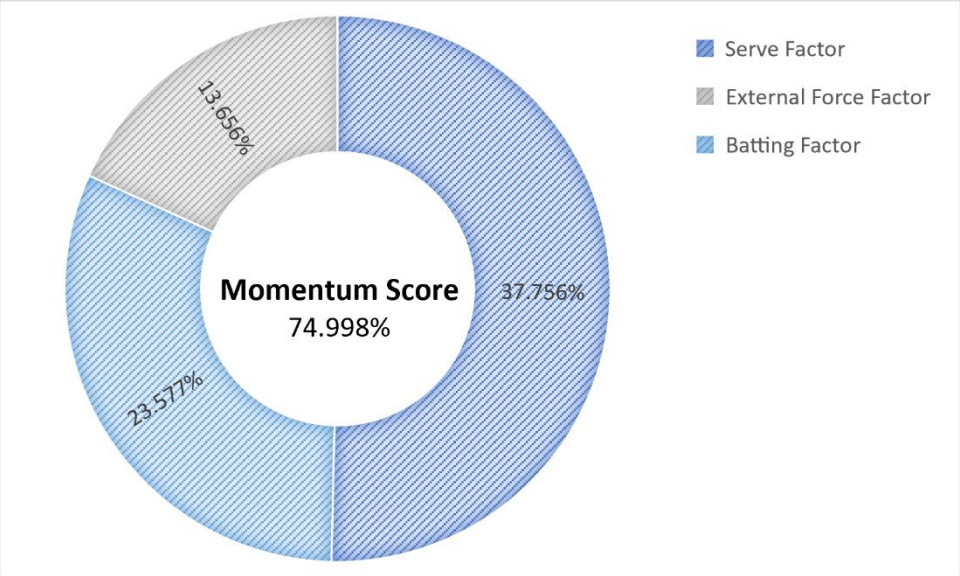

Figure 5. Momentum Composition Diagram

In this way, this paper ends up with an expression for the momentum score based on factor analysis. The essence of calculating the momentum score is a series of assigning weights and summing the values of the selected indicators.

3.7. Discussions

A tennis player's momentum is certainly not constant throughout the match. The expressions this paper has obtained above are based on the tennis player's entire match's tournament data set. To get the momentum of a player in a specific round, we just need to substitute the player's event data for that round. Note that the values substituted here are the normalized values. For the opponent's momentum score expression, we need to substitute in the opponent's entire match data set to analyse it separately. In addition, the expression this paper obtained above is only for one of the players who participated in the tournament.

The disadvantage of this model is that it is not possible to calculate a player's momentum value in real time during a match, but only to calculate and analyse the momentum score expression for a specific player based on all the round-by-round match data for the whole match after the match. Substituting in the event data of the specified player for a specific round, the player's momentum value for that round can be obtained.

4. Conclusion

This study is committed to in-depth analysis of players' performance in tennis matches. By analyzing the match data in a refined way, this paper constructed a comprehensive system of technical indicators to quantify the player's momentum in different game segments and reveals three implicit factors that affect the momentum score. But quantifying a tennis player's poise is a relatively subjective task. This is because different definitions of momentum could result in different analyses. Meanwhile, although this paper considered most technical indicators that may be related to the momentum, there are some subjective indicators that are difficult to capture. For example, a player's chants after each round win and body language during the game are factors that can reflect a player's momentum to a certain extent, but there is a lack of means to collect them and statistical standards. Furthermore, the relationship between momentum status and performance outcomes in sports is complex and may involve interactions with other cognitive factors beyond psyche [11-12]. It may lack rigour to extrapolate a

player's momentum status backwards from match performance alone. Therefore, this study only provides a perspective and direction to quantization and provide valuable perceptions for coaches and athletes to help them better understand the game situation and develop strategies for future games.

References

[1] Goyal A, Simonoff J S. Hot Racquet or Not? An Exploration of Momentum in Grand Slam Tennis Matches [J]. arXiv e-prints, 2020: arXiv: 2009.05830.

[2] Peng R, Li Z. MLFEF: Machine Learning Fusion Model with Empirical Formula to Explore the Momentum in Competitive Sports [J]. arXiv preprint arXiv: 2402.12149, 2024.

[3] Lin J, Shao P, Zhang Q. Advancing Tennis Analytics: Comprehensive Modeling for Momentum Identification and Strategic Insights [J]. International Journal of Computer Science and Information Technology, 2024, 2 (1): 104 - 117.

[4] Watkins M W. Exploratory factor analysis: A guide to best practice [J]. Journal of black psychology, 2018, 44 (3): 219 - 246.

[5] Lu Qi. Research on Performance Evaluation of AEnergy Company Based on Factor Analysis [D]; Xi'an Shiyou University, 2023.

[6] Taylor J, Demick A. A multidimensional model of momentum in sports [J]. Journal of Applied Sport Psychology, 1994, 6 (1): 51 - 70.

[7] Bai Zhenta. An Analysis of the Physical Consumption of Players in Tennis Mens Singles Matches and the Indicators related to Match Outcomes [J]. Bulletin of Sport Science & Technology, 2022, 30 (6): 61 - 5.

[8] Page L. The momentum effect in competitions: field evidence from tennis matches[C]//Econometric Society Australasian Meeting. 2009.

[9] Shi Wenhao. Research on Financial Performance EValuation of C AutomobiIe Company based on lhctor analysjs method [D]; Xi'an Shiyou University, 2023.

[10] Geng Yuanyuan. Research on financial performance evaluation of Vanke based on factor analysis [D]; Hebei Agricultural University, 2021.

[11] Bühren C, Steinberg P J. The impact of psychological traits on performance in sequential tournaments: Evidence from a tennis field experiment [J]. Journal of Economic Psychology, 2019, 72: 12 - 29.

[12] Morgulev E. Success breeds success: Physiological, psychological, and economic perspectives of momentum (hot hand) [J]. Asian Journal of Sport and Exercise Psychology, 2023.