Problem Chosen
C

2024
MCM
Summary Sheet

Team Control Number
2400384

The Art of Momentum Mastery in Tennis

Summary

Tennis is a popular sport played by millions of people in clubs and on public courts. Affected by multiple factors, the trend of the game is usually unimaginable and players' performance is often unexpected. This paper studies the importance of each point and the impact of momentum on the match by establishing two models: **S**core-capture and **P**erformance-identification **M**odel (SPM) and **M**omentum **S**wing **P**rediction **M**odel (MSPM).

The SPM model is divided into two parts. The first part is **Score-capture model**, which aims at detecting the point changes and calculating the swing of winning probability when different game scores occur. Evolved from Markov Chain, Score-capture Model can not only find the match winning probability at any point, but also depict the probability of one score changes to another since the current state is closely related to the former state. By ensuring the state transition matrix's stability, we successfully established three Markov chains, which reveal the relationships between point to game, game to set, and set to match. Due to their different winning-calculation system, we analyzed the coefficient between these three chains and adjust the weight to correctly use this link.

The second part of SPM is a **performance-identification model**, which can estimate a player's performance at any point of the match by performing descriptive statistical analysis to the player's critical moments during the game and comparing his current winning state with the predicted winning state we calculated by using the score-capture model. When doing the calculation, we noticed that the performance score only swings from a small range, so we did **statistics normalization** to broaden its swinging range, better visualizing a player's performance change during the match. What is more, to make our model more accurate, we set an **auto-adapted parameter** based on leverage to change the weight of a player's critical moment performance and general performance, making our calculation more practical and precise.

In order to better understand the flow of the game, we analyzed the existence and influence of momentum and make it quantitative, making the momentum easier to understand and detect. In the process of **momentum quantization**, we did MCMC sampling to do random competition simulation. By comparing the randomly generated matrix with momentum-considered matrix under **KS test**, we noticed the impact of momentum and used leverage and winning probability to calculate it. What is more, we established a momentum swing prediction model based on **LSTM neural network**, and used sequence correlation to extract relevant features that may have impact on momentum. By training the set with high efficiency under supervision and implementing cross-validation to evaluate the performance of the model, our model reached an accuracy of 76.3%. Furthermore, we discussed the model's **universality** and applied it to other matches in the end.

Key words: Markov Chain, momentum, LSTM neural network, prediction

Contents

1. Introduction

"Psychological momentum is the positive or negative change in cognition, affect, physiology, and behavior caused by an event or series of events that affects either the perceptions of the competitors or, perhaps, the quality of performance and the outcome of the competition." – **The Oxford Dictionary of Sports Science & Medicine** [1]

1.1 Problem Background

Tennis, a game in which two opposing players use tautly stung rackets to hit a ball over a net on a rectangular court, is a popular sport played by millions of people in clubs and on public courts. Aiming at describing which player is in control at the point of the match, momentum plays a significant role in playing tennis. It had been reported that a swing in momentum can make a difference in the probability of winning the match. To expand advantages and stimulate the potential of the athletes, many game review and analysis have been completed and winning strategies have developed rapidly in recent years. Since it is difficult to measure momentum precisely, a momentum-capture-and-analysis model is needed to help predict the swing of the match.

1.2 Restatement of the problem

Considering the background of the question and the limitations, we need to develop a model that can identify the process of the game as points occurs and analyze the current situation to test the performance of the players.
After through in-depth analysis and research on the background of the problem, we can specify that our article should cover the following aspects:

- Generate a model that can capture the flow of the play when points have been made and distinguish the player with greater advantage at a given time, then apply it to at least one of the matches and provide a relevant visualization to illustrate the match flow.
- Use the model to assess the role of "momentum" in the match and explore whether the switch in the game and the continuous success of one player is random or not.
- Use the data provided for at least one match to establish a model that can help to indicate the swing of momentum in the match. Find the and apply it to find a better strategy for one player to play against another based on the given "momentum" swings in the past.

1.3 Literature Review

Numerous statistical prediction models have been applied to tennis matches, which mainly focus on Markov Chain (MC), a mathematical system that change status due to certain probabilistic rules, and Long Short-Term Memory (LSTM), a recurrent

neural network that is applicable to classification and predicting date based on time series. The work of Klaassen and Magnus (2001) [2] showed an analytical equation for match-win probability given each individual player's score status by constructing a hierarchical Markov model, which revealed the link between each games' winning probability. What is more, Helmut and Cornel (2017) [3] pointed out that once players lose control over a match, their chances of winning the next set are significantly below their opponents, which can be attributed to the thing called anti-momentum. Since LSTM can remember information for extended periods of time, Dumovic and Howarth (2017) [4] designed a neural network model to make predictions based on features including weather, the mentation of the player and so on with approximately 69.6% accuracy. These findings laid a solid foundation for our research and we developed Markov Chain and LSTM on this basis.

1.4 Our work

The problem requires building a model to capture the flow of the score and identify the performance of the player. We firstly build a Score-capture and Performance-identification Model (SPM) evolved from Markov Chain. Then based on the findings of the model, we establish a developed model of momentum that can predict the swing of the match and offer relevant advice. Finally, we tested our model by applying it into more matches and even other sports to prove its universality.

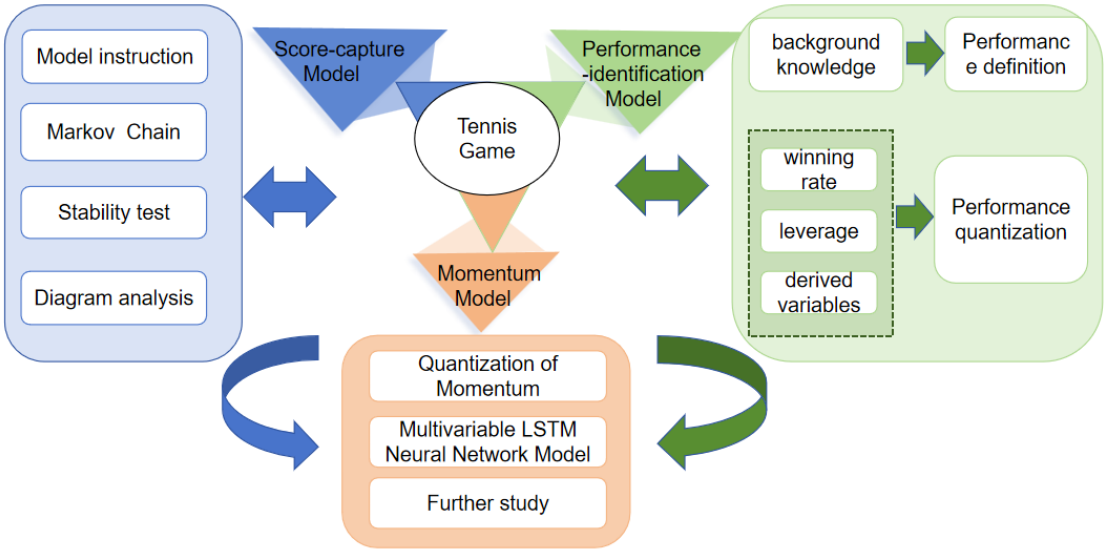


Figure 1: Diagram of our main work

2. Assumptions and Notations

2.1 Assumptions

In practical application, tennis competitions have many complex situations. To simplify the analysis of our problem, we make the following reasonable assumptions:

- **Assumption 1: The model can only be applied to professional competition.**
  Explanation: The strength of amateur players is not stable, and amateur players are more likely to be affected by various external factors, their victory and defeat can often be fortuitous. In addition, the record of professional events is more accurate, and its data is easier to operate and analyze.
- **Assumption 2: The pre-game odds can reflect individual strength and consider it as part of exploring the win rate.**
  Explanation: The pre-game odds are calculated by bookmakers based on a complex of factors such as the recent performance of the players and the world rankings and so on, thus reflecting the strength of the players
- **Assumption 3: The trend of the game is not affected by the weather, the audience reaction and the location of the venue.**
  Explanation: Athletes are under the same environment during the process of the game, though different individuals will have different reactions to different weather, the impact on the game can also be a part of the individual strength, which we think has been taken into account in the pre-game odds.

2.2 Notations

- Symbols and Descriptions

Symbols	Descriptions
P_{pm}	the probability of winning the match on this point
$P_m(S_1, S_2)$	the probability of a particular set state
$P_m(G_1, G_2)$	the probability of a particular game state
$P_m(P_1, P_2)$	the probability of a particular point state
P_{sm}	the winning probability on this point based on the player's strength
P_m	the probability of winning the match
L	A particular point's impact on the winning rate of the match
d_{ACE}	the score one ACE get at this point
d_{sp}	A player's specific performance score at this point
D_{sp}	A player's specific performance score at this point after normalization
d_{gp}	A player's general performance score at this point
D_p	A player's performance score at this point

- Technical term and Concept

Technical term	Concept
ACE	A legal serve which the returner does not manage to get their racquet to, which always results in the server winning a point.
unforced error	A missed shot or lost point (as in tennis) that is entirely a result of the player's own blunder and not because of the opponent's skill or effort.
breaking success rate	breaking success times/obtaining breaking points

breaking rate	breaking innings/receiving innings
holding rate	holding innings/serving innings
leverage	A concept that measures the importance of a single point to the final outcome of a tennis match by quantifying how much a player's probability of winning the match changes.
momentum	A concept aims to describe which player is in control at any point of the match

3. Score-capture and Performance-identification Model

Score-capture and Performance-identification Model (SPM) is a model evolved from Markov Chain that can calculate the game odds when different scores occur based on the initial pre-game odds. After the Markov Chain calculates the winning probability trend of score changing under thousands of match information searched online, the player's performance can be obtained by comparing the winning probability with the player's strength, which can be reflected by pre-odds.

3.1 Data Preprocessing

Before data analysis and model establishment, we need to guarantee the availability of data.

- **Data substitution.** We noticed that the scores of two players sometimes range from 1 to 9, which is not the standard score such as 0, 15, 30 and 40. This happens when the two players reached a tie of 6:6 in a set and they are on the 13^{th} game. To better calculate the changes of the score, we replace 1 with 15, 2 with 30, 3 to 9 with 40 or AD.
- **Data collecting.** Since we need to build a score-capture Model based on Markov Chain, we collected the match results and scores of Wimbledon Gentlemen's Singles Competition during 2016 to 2023 (except 2022 since the information is not available online), with the numbers of games amounting to more than 2300, to make our model more general and accurate. Furthermore, we collected the information of pre-odds provided by the gambling company to help calculate the player's winning probability.
- **Data clustering.** To estimate the performance of the player more comprehensively, we add up the number of a player's ACE, unforced error and cumulative running distance and so on. In addition, we also create derived variables such as leverage, breaking success rate, breaking rate and holding rate.

3.2 Score-capture Model Based on Markov Chain

Score-capture Model can detect the winning probability changes when different game scores occur, it can show the probability of one score changes to another.

3.2.1 Markov Chain

Markov Chain is a stochastic model describing a series of possible events, where the probability of each event depends only on the state reached by the previous event.

One of its difference from other general stochastic process is that it experiences transitions from one state to another according to certain probabilistic rules and it must be "memory-less".
The Markov Chain is a stochastic sequence that can be written as follows:

$$\{x_n, n = 1,2,\dots\}$$

And the probability of the next state is only related to the current state:

$$P\{x_{n+m} = a \mid x_1 = a_1, x_2 = a_2, \dots, x_{n+m-1} = a_n\} = P\{x_{n+m} = a \mid x_{n+m-1} = a_n\}$$

To be more specific, every next state is related to the current state and some of it can be recurrent. What is more, as n becomes larger, the dependence of $P\{x_n\}$ has on the initial state becomes less and less, after a long enough time, the initial state is negligible.

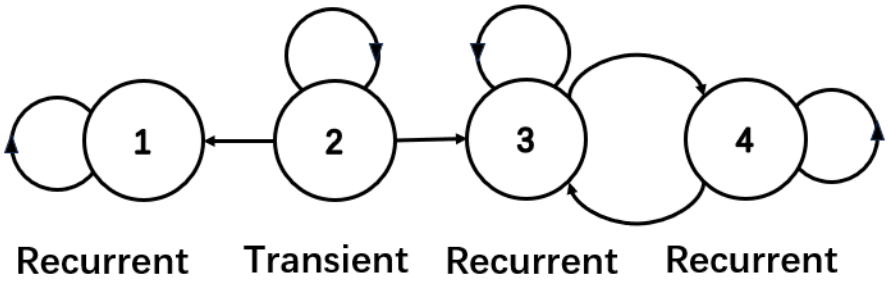


Figure 2: Markov probability transfer diagram

The state of our model is transient most of the time because the game score is not likely to go back with the process of the game. However, we consider one case recurrent, which happens when the two players have reached a tie at 40:40 during the game. In this case, if player 1 wins, the score will turn to AD:40, if he continues to win, he will win this game, if he loses, the score will go back to 40:40, instead of AD: AD. The specific flow of the game will be shown later.

3.2.2 Markov Stability Test

Once the flow of the chain is decided, we can have the state transition matrix P, where $P(i,j)$ means the probability of state j under the condition state i, namely $P(j|i)$.
Due to the character of P, the values of each element in the state transition matrix are only related to the column number, and the values of the elements in the same column are the same, which can be expressed as the matrix's stability. We use the following methods to check whether our state transition matrix is stable or not.

$$\lim_{n\to\infty} P^n = \begin{Bmatrix} \pi(1) & \pi(2) & \dots & \pi(j) & \dots & \pi(K) \\ \pi(1) & \pi(2) & \dots & \pi(j) & \dots & \pi(K) \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \pi(1) & \pi(2) & \dots & \pi(j) & \dots & \pi(K) \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ \pi(1) & \pi(2) & \dots & \pi(j) & \dots & \pi(K) \end{Bmatrix}$$

$$\lim_{n\to\infty} P_{ij}^n = \pi(j)$$

Think of the matrix P as a combination of π

$$\pi(j) = \sum_{i=0}^{K} \pi(i) P_{ij}$$

3.2.3 Score-capture Model establishment

Tennis competition can be divided into three parts: game, set and match, as a result, we build three sequences which represents the probability of set changes, game changes and score changes respectively with six variables as follows:

$$P_m(S_1, S_2),\ P_m(G_1, G_2), P_m(P_1, P_2)$$

where S_1 means the current number of set player 1 wins, S_2 means the current number of set player 2 wins, G_1 means the current number of games player 1 wins, G_2 means the current number of games player 2 wins, P_1 means the current score player 1 wins, P_2 means the current score player 2 wins.

Take point as an example, since every game starts at (0,0) and a player must win four or more points by a margin of at least two to win a game, we have enumerated and analyzed the following cases:

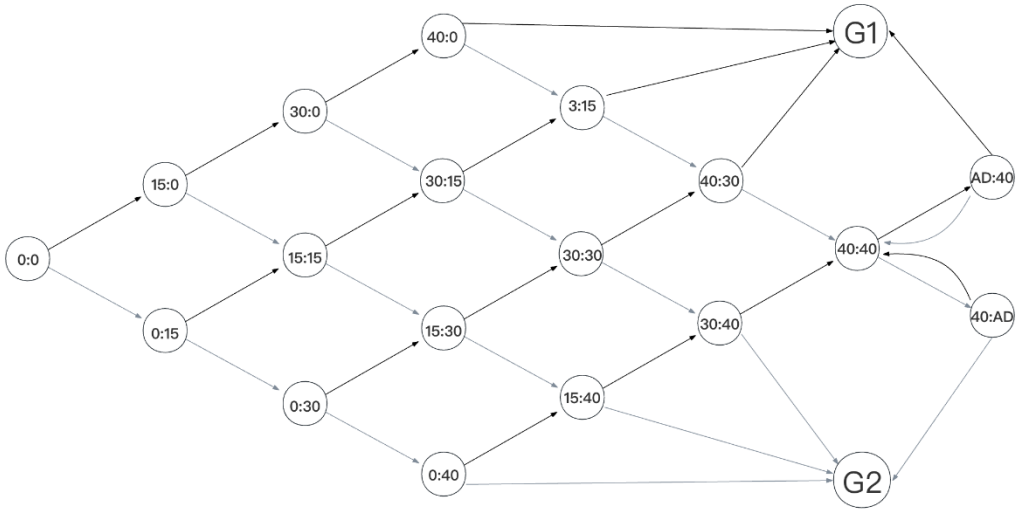


Figure 3: Markov Chain of a game

The above picture depicts the flow of the game, where "G1" means player 1 wins, "G2" means player 2 wins. By using the Markov Chain model, we can have the corresponding probability of one score change to another.

Step 1: Name each score a state. For computational convenience, we think that 0:0 stands for state 1, 15:0 stands for state 2, 0:15 stands for state 3, 30:0 stands for state 4 and so on. Labeling from top to the bottom, from left to right of the flow diagram. In addition, since we can have infinite ties, we think the next state of "AD:40" is either "win" or "40:40".

Step 2: Insert the information into the model. Then, we can have the probability of state i change to state j by using the formula as below:

$$P_m(G_{ij}) = \frac{n_{ij}}{n_i}$$

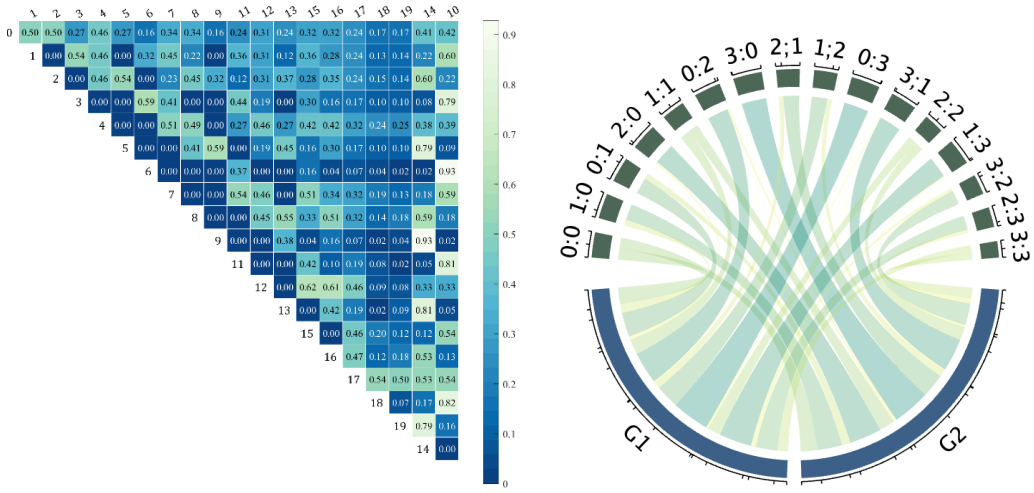
where n_{ij} means the number of times state i changes to state j, n_i means the number of times the score reaches state i.

Step 3: Export the probability of state changing and get the state transition matrix. Examine the stability of the transition matrix. The model is established after we have the stable transition matrix and we can use visualization techniques to make statistics easier to understand and analyze.

Step 4: Integrate the probability changes between each game and set, by using the following formula we can get the impact of a point on the match:

$$P_{pm} = P_m(S_1, S_2) + (1 + P_m(G_1, G_2) \times (1 + P_m(P_1, P_2)))$$

3.2.4 Score-capture Model Analysis



(a) transition probability matrix (b) chord diagram relationship

Figure 4: the impact of different points on game

Take the point of a game as an example. The figure above shows the relationship between each point in a game. The figure on the left is a thermodynamic chart, where the lighter the color is, the closer the two points relationship are. From this we can discover that the point 40:0 is the most likely to reach victory for player 1 and 0:40 is the most likely for player 1 to lose, AD:40 is very likely to win while 40: AD is very likely to lose. Furthermore, the probability of player 1 to win or lose a point is different when facing different points. For example, we can see that for 30:15 changes to 30:30, the changing probability is 46%, while the changing probability for 30:30 turns to 30:40 is 61%, with a difference of 15%. The flow of the points is symmetric without considering the strength of the player. The right figure is a chord diagram, where G1 means player 1 wins and G2 means player 2 wins. When the linking color is darker, it is more likely to for one point to reach another.

3.3 Performance-identification Model

3.3.1 Basic knowledge

Before we start analyzing the performance of the player, we need to understand the meaning of following situations:

- Break point: the returner wins the game by scoring the next point.
- Hold: the server wins the game.
- ACE: the server wins the point by just hitting their serve with the returner can't get it back into court and doesn't even get a racket on the ball
- Double fault: the server misses both their first and second serve, which will automatically lose the point then and there.
- Net point: a point won or lost that is approaching the net.
- Leverage: the impact of a particular point on winning the match

After we know the basic knowledge of special cases in tennis, we can figure out how to identify the player's performance.

3.3.2 Winning rate calculation

We have collected the pre-odds information searched online, since the gambling company needs to make profit, we assume that the information it provided has a $k\%$ swing, which means that the pre-odd of the player is the reciprocal of $1-k\%$ winning rate. Therefore, we have the winning rate based on the player's strength:

$$P_{sm} = \frac{1}{\frac{P_{odd}}{1-k\%}} = \frac{1-k\%}{P_{odd}}$$

In our model, we take k as 10, since Forrest (2003) has pointed out in his research that the gambling company usually deduct an 8 to 11 percentage from the sum of money as a profit.

Now, we not only have the winning rate P_{pm} calculated by the score-capture model, which is a general statistic based on thousands of games, but also have the winning rate P_{sm} specialized for individual, by combining these two winning rates, we can get a rather practical winning rate P_m. In reality, the better a player's skill is, the less his winning probability is influenced by the general winning rate, therefore, we can have the practical winning rate P_m as follows:

$$P_m = \frac{P_{sm}}{P_{sm}+P_{pm}} \cdot P_{sm} + \frac{P_{pm}}{P_{sm}+P_{pm}} \cdot P_{pm}$$

3.3.3 Derived Variables, Leverage and Performance Quantization

The performance of the player can not be estimated only by his winning rate, the move he made on the game such as the number of a player's ACE should also be taken into consideration. We create a scoring system for the number of ACE, double faults, break point won and unforced- error. For ACE, we multiple the increase pf the score ACE get as the number of ACE accumulated, which can be described as follows:

$$d_{ACE} = 4$$

where d_{ACE} means the score a player gets based on the number of the ACE.

Similarly, we can have the score of double faults, break point, break point won, unforced-error, first serve won, second serve won and successive victory on the point:

$$d_{DF} = -2$$

$$d_{BP} = 0.5n_{BP}$$

$$d_{BPW} = 2n_{BPW}$$

$$d_{UE} = -2(n_{UE}-1)$$

$$d_{FW} = \begin{cases} 2 \times 0.9, & when\ the\ player\ is\ server \\ 2 \times 1.1, & when\ the\ player\ is\ turner \end{cases}$$

$$d_{SW} = \begin{cases} 1 \times 0.9, & when\ the\ player\ is\ server \\ 1 \times 1.1, & when\ the\ player\ is\ turner \end{cases}$$

$$d_{SV} = 2(n_{SV} - 1)$$

Explanation:

- Since double fault and unforced error are negative factors, we add a minus when estimating their score.
- n_{BP} means the number of BP a player has made in the match at the point, others are similar.
- ACE is the most difficult to get, which is related to the player's current performance closely, we give it the highest rate of rise. Double faults, first serve won and second serve won is a very common phenomenon in a match, then we give them a fixed score as a result.
- When ACE happens, though this situation is contained in the first serve won, we only calculate the score ACE get to avoid repeat scoring.
- Since the player serving has a much higher probability of winning the point/game, we calculate d_{FW} differently when the player changes from server to returner. The following figure will explain this phenomenon more clearly (We take 2023-wimbledo-1701 as an example).

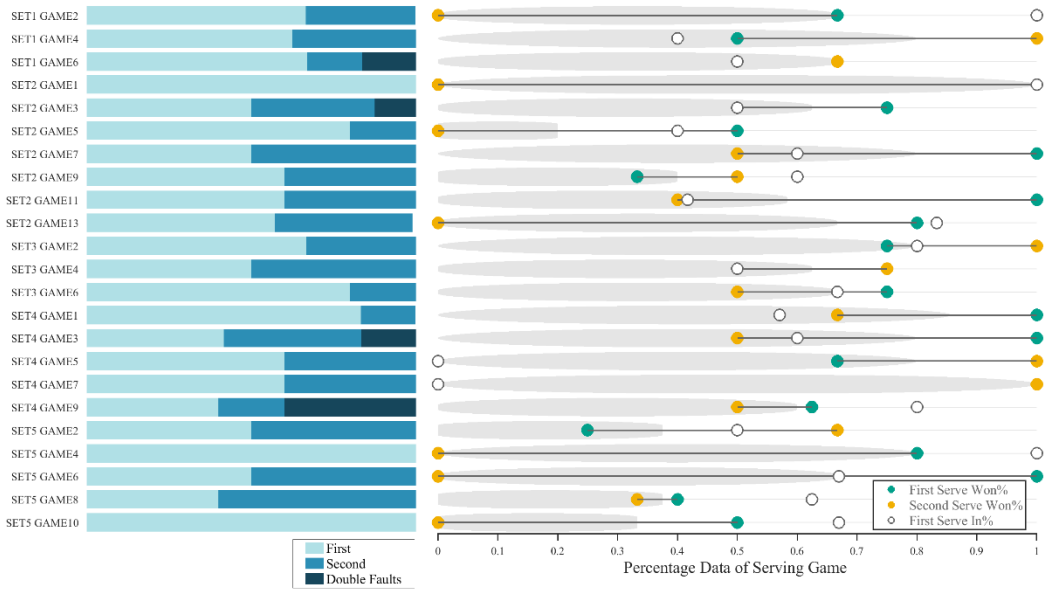


Figure 5: Carlos Alcaraz's serving performance

We collected and calculated the data of Carlos Alcaraz's serve in 5 sets, in which he was server for 24 games. The light blue in the above figure means the probability of first serve in, median blue means the probability of second serve in, and dark blue means the probability of double faults. By this we can see that most of Alcaraz's first serve is successful, but his performance is unstable since sometimes the probability of double faults is also very high, like Set 4 Game 9. On the right hand of the figure, green point means the probability of first serve won, yellow point means the probability of second serve won, while the grey background means the probability of serve won. By this we can see that the probability of first serve won is usually higher than second serve

won. The swing of the grey background also provides the evidence of the unsteadiness of Alcaraz's performance. Similarly, by analyzing the break point data, we can have the returner's performance. Combining these two together, we decided to set the probability of server and returner changes to 10%.

Now we have the scores with specialized coefficient, add all these factors up, we can have the player's specific performance score d_{sp} at this point:

$$d_{sp} = d_{ACE} + d_{DF} + d_{BP} + d_{BPW} + d_{UE} + d_{FW} + d_{SW} + d_{SV}$$

Calculate the specific performance score at every point, then do normalization:

$$D_{sp} = \frac{d_{sp} - d_{sp\,min}}{d_{sp\,max} - d_{sp\,min}}$$

Carlos Alcaraz		Novak Djokovic
9	ACES	2
7	DOUBLE FAULTS	3
94/150 62.67%	FIRST SERVE%IN	**64.13% 118/184**
67/94 71.28%	WON% ON 1ST SERVE	61.86% 73/118
28/49 57.14%	WON% ON 2ND SERVE	**58.73% 37/63**
5/19 26.32%	BREAK POINT WON%	**33.33% 5/15**
25/38 65.79%	NET POINT WON%	59.65% 34/57
35	UNFORCED ERRORS	**30**
6606.523	DISTANCE	6195.168
39.32	DISTANCE COVERED/PT	37.32
168	TOTAL POINT WON	166

Figure 6: Statistics collection and comparation between the two players

The figure above directly shows the statistics of individual's specific performance and the comparation between the two players. We can see that Carlos Alcaraz is obviously better at doing ACE, which will bring him 36 scores, while Djokovic only get 18 scores according to the specific performance score system mentioned above.

We have calculated P_m, the practical winning rate at this point, then we need to compare it to the actual winning state to get the player's general performance score D_{gp}:

$$D_{gp} = \begin{cases} 100 \times (1 - P_m), & if\ the\ player\ wins \\ 100 \times (0.5 - P_m), & if\ the\ player\ loses \end{cases}$$

For example, according to our model, player 1 has a winning percentage P_m of 65%, however, he loses the game, then his general performance score will be $100 \times (0.5 - 65\%) = -15$.

Leverage is the amount by which the player's game winning probability changes given the next point. For instance, a player's probability of winning the match will increase by 5% (0.05) if he wins the next point. But, if he loses that point, his probability of winning the match will decrease by 2% (-0.02). The difference between these match-winning probabilities is 7 percentage points (0.05 – (-0.02) = 0.07). Therefore, we assume this point has a leverage of 0.07.

Now, all the information needed for estimating performance has been calculated. We can now do performance quantization to get the player's performance score D_p by using the following formula:

$$D_p = L \cdot \left(\frac{-ln(-L+1)}{-ln(-L+1)+1} D_{sp} + \frac{1}{-ln(-L+1)+1} D_{gp} \right)$$

3.3.4 The Performance Results and Visualization

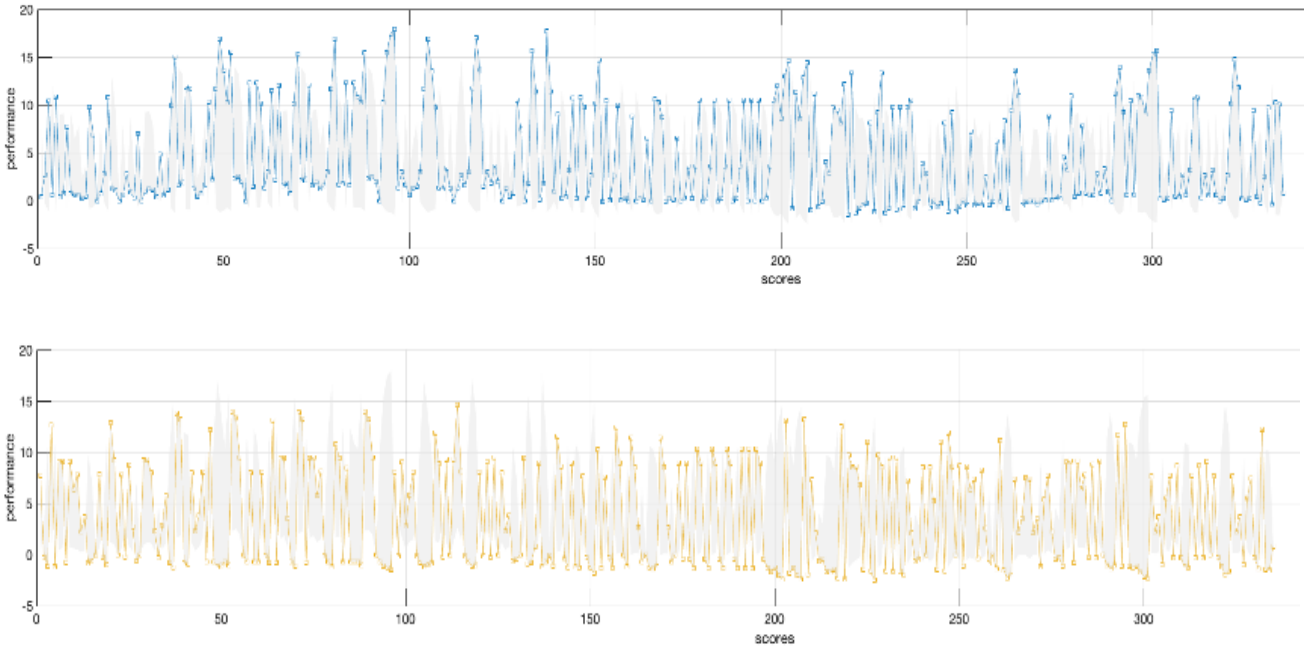


Figure 7: The real-time performance chart of the two players

Table 1: The top five with the highest performance difference

1701 Line No.	50	96	97	106	138
Set	0:1	0:1	0:1	0:1	0:1
Game	0:0	3:3	3:3	4:4	6:6
Point	2:1	2:0	3:0	2:1	5:6
Performance_ Diff	18.0987176	18.9119673	19.3856514	18.0977844	18.9119673

These two graphs illustrate the real-time performance changes of the two players during the 2023 1701 season. We observed that at crucial scores, such as the transition from set 0:0 to 0:1, the performance disparity between the two is most pronounced.

4. Analysis on the existence of momentum

4.1 Quantization of Momentum

4.1.1 What is Momentum?

Momentum is a concept that aims at describing which player is in control at a particular point of the match, in our model, we define momentum as an exponentially weighted

moving average (EMA) of the combination of leverage and the probability of winning the match at the point.

4.1.2 Calculate Momentum

Firstly, me need to check the smoothness of the combination of leverage and the probability of winning the match at the point, which we named C_{lp}:

$$C_{lp\_t} = \gamma \cdot L_t \cdot P_{m\_t}$$

where C_{lp_t} means the combination of leverage and winning rate at this point, L_t means the impact of this point on the match and P_{m_t} means the player's probability of winning the match at this point. Draw the graph of C_{lp}, we can get the following figure:

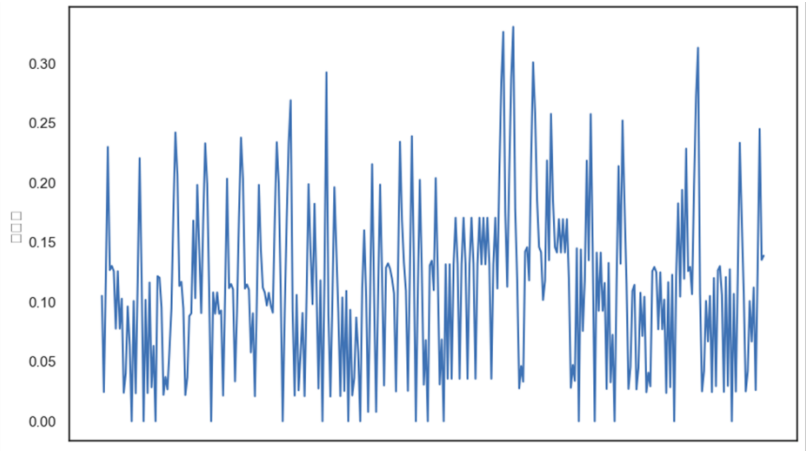


Figure 8: Time series diagram of C_{lp}

The figure shows the flow of C_{lp} with the process of the game. From the figure we can roughly justify that this time series is stable. To make our findings more precise, we do the ADF test with drift and trend on the series to check whether it is stationary at level:

$$\Delta C_{lp\_t} = \beta_1 + \beta_2 t + \alpha C_{lp\_t-1} + \sum_{k=1}^{K} \theta_k \Delta C_{lp\_t-k} + \mu_t$$

After calculation, we can find out that the coefficient of the trend β_2 is significant, then, we use this equation to determine stationarity.

Since the statistics of C_{lp} is stable, we can now calculate momentum by exponential moving average:

$$EMA_t = (LP_{m\_t} - EMA_{t-1}) \times \alpha + EMA_{t-1} \times (1 - \alpha)$$

While

$$\alpha = \frac{2}{N+1}$$

Where α is the smoothing factor, N is the period length of EMA.

By the recursive expression, we can have the following expression:

$$EMA_t = \alpha A_t + (1-\alpha)\alpha A_{t-1} + (1-\alpha)^2 \alpha A_{t-2} + \ldots$$

Since C_{lp} is stable, we should choose shorter N and bigger α. In our model, we take N as three. Therefore, we have the quantization of momentum, which is EMA_t. When we list these statistics in a matrix, we will get the momentum transition matrix P_{mom}

4.2 MCMC Sampling

Markov Chain Monte Carlo, short for MCMC, is a sampling model that can deal with high-dimensional random variables. In a tennis match, we have three units which are points, games and sets when analyzing the winning state of the match, therefore, using MCMC to do random gambling for random process simulation would be an appropriate method.

In 4.1, we have already got the momentum transition matrix P_{mom}, however, this matrix is not stationary since it cannot satisfy the following equation:

$$\pi(i)P_{mom}(i,j) \neq \pi(j)P_{mom}(j,i)$$

where $\pi(x)$ is the stable target.

Therefore, we need to revamp the above formula to get the stable momentum transition matrix:

$$\pi(i)P_{mom}(i,j)\alpha(i,j) = \pi(j)P_{mom}(j,i)\alpha(j,i)$$
$$\alpha(i,j) = \pi(j)P_{mom}(j,i)$$
$$\alpha(j,i) = \pi(i)P_{mom}(i,j)$$

In this way, we can get the corresponding stable Markov state transition matrix P of momentum, where

$$P(i,j) = Q(i,j)\alpha(i,j)$$

where Q is an arbitrary state transition matrix and α is the acceptance rate

4.2.1 Random Process Simulation

To detect the existence of momentum, we first do the random competition simulation. During this random test, the winner of the point is totally random, but the player serving will have a 10% higher probability of winning. The following algorithm shows the common steps for establishing a random competition simulation.

Algorithm 1 MCMC Sampling Process

Input: state transition matrix P
Output: sample set $(x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2-1})$

```
def mcmc_sampling(P, pi, n1, n2)
x_t = np.random.choice(len(pi), p=pi)
samples = []
for t in range(n1 + n2):
    x_star = np.random.choice(len(pi), p=P[x_t])
    u = np.random.uniform(0, 1)
    alpha = pi[x_star] * P[x_star][x_t]
    if u < alpha:
        x_t = x_star
    if t >= n1:
        samples.append(x_t)
    end if
end for
return samples
```

However, the efficiency of MCMC sampling may be low if the initial acceptance probability of α is too small. Therefore, while promising the equation still holds, we expand the matrix by a factor of 5:

$$\pi(i)P(i,j) \times 0.5 = \pi(j)P(j,i) \times 1$$

What is more, we make a few improvements to normalize α:

$$\alpha(i,j) = min\{\frac{\pi(j)P(j,i)}{\pi(i)P(i,j)}, 1\}$$

Finally, we get the samples of the difference of momentum at each point.

4.2.2 The coach might not be true

The Markov chain model we have developed can be sufficiently used as an example to validate this coach's idea. In tennis, to verify whether the change of momentum is just a random event. A traditional way we can use the KS test method, which is a non-parametric test to check whether a sample comes from a specific probability distribution or the state. The distribution matrix of our Markov chain model can well represent a specific probability distribution, thus we can use this method to test the momentum theory.

Firstly, we assume that momentum is a random event in a match, this is important so we can use the reality distribution matrix to test whether momentum is a random event. The randomly generated state distribution matrix for momentum would be compared with the state distribution matrix from the actual matches' data. To achieve this process, the following 6 steps enable us to show the difference in momentum flow between the two scenarios through data 。

Step 1: Generation of State Distribution Matrix

In this step, a state distribution matrix is constructed based on nature of momentum in the system under consideration. This matrix serves as a representation of the probabilistic transitions within the system, capturing the random flow of momentum. The constructed matrix is prepared for subsequent comparative analyses.

Step 2: Utilization of Tennis Data and Determination of Critical Value

The available tennis data denoted as x_i, is employed in this stage, where n represents the sample size. At this juncture, a critical value α.is established, marking a significance threshold for subsequent statistical assessments. The critical value is pivotal in hypothesis testing and decision-making processes.

Step 3: Computation of Mean μ, and Standard Deviation δ.

Statistical measures, namely the mean μ, and standard deviation δ, are calculated from the provided tennis data.

Step 4: Assignment of Ordinal Numbers

Each data point in the dataset is assigned an ordinal number.

Step 5: Empirical Distribution Function Calculation

The empirical distribution function (EDF) is computed, representing the cumulative probability distribution derived directly from the observed data.

Step 6: Computation of Cumulative Distribution Function (CDF) of Standard Normal Distribution

Calculate the cumulative distribution function of the standard normal distribution.

$$\phi(z) = \frac{1}{2}\left(1 + \text{erf}\,(\frac{z}{\sqrt{2}})\right)$$

Step 7: Calculate the value of fabs, D, and determination of α

$$fabi(i) = |F(i) - CDF(i)|$$

$$D = \max\{fabis(i)\}$$

4.2.3 Hypothetical Conclusions

Table: Validation of the Impact of Momentum

Data	n	Mean	Standard Deviation	KS Statistic	p-value
Random_momentum	100	1.234423022	0.135542426	0.728333333	1.4581E-28
Reality_momentum	120	0.990709406	0.11736885		

Based on the results of the KS test, we reject the null hypothesis (H0), indicating a significant deviation between our data and the hypothesized theoretical distribution. At a significance level of 95%, the observed statistic surpasses the critical value, providing sufficient evidence to reject the null hypothesis. with a p-value of less than 0.0001 in the KS test, far below the 0.05 significance level.

Thus, we can confidently say that the existence and flow of momentum is not a random event.

5. Momentum-detection Model

5.1 Sequence Correlation

In our model, we divided our characteristic mainly into six parts:

- Score differential: continuous score, continuous loss score
- Ability to serve: the number of ACE, the serve speed, first serve won
- Fault: the number of double-fault, the number of unforced error
- Critical Points: break point, net point
- Athletic ability: running distance
- Drop point: depth and width

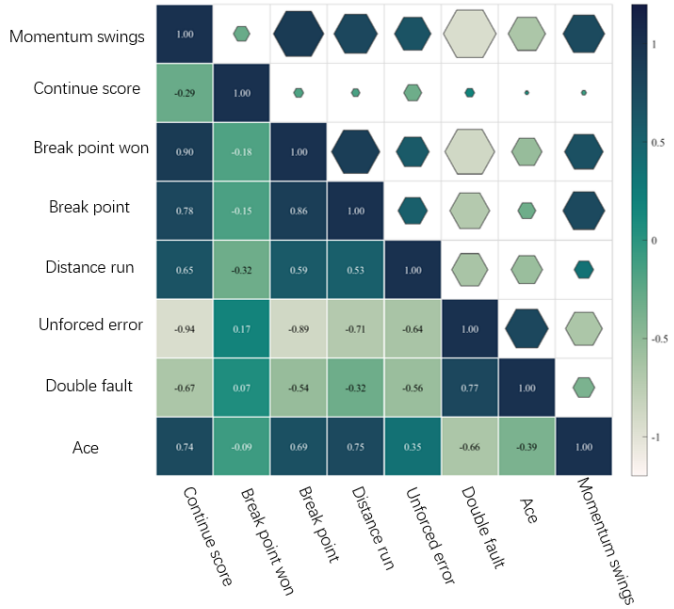


Figure 9: thermodynamic chart of characteristics

In the figure above, we can see that ace is strongly correlated with unforced error and distance run, which may be related to the particularity of ace. double faults are not highly correlated with most indicators, and may have a low incidence and low correlation with on-field scores.

5.2 Multivariable LSTM Neural Network Model

Since we have multiple variables, we need to establish multivariable LSTM neural network model.

Long Short-Term Memory Networks, short for LSTM, is a kind of temporal recurrent neural network, which is specially designed to solve the long-term dependence problem of general RNN (recurrent neural network).

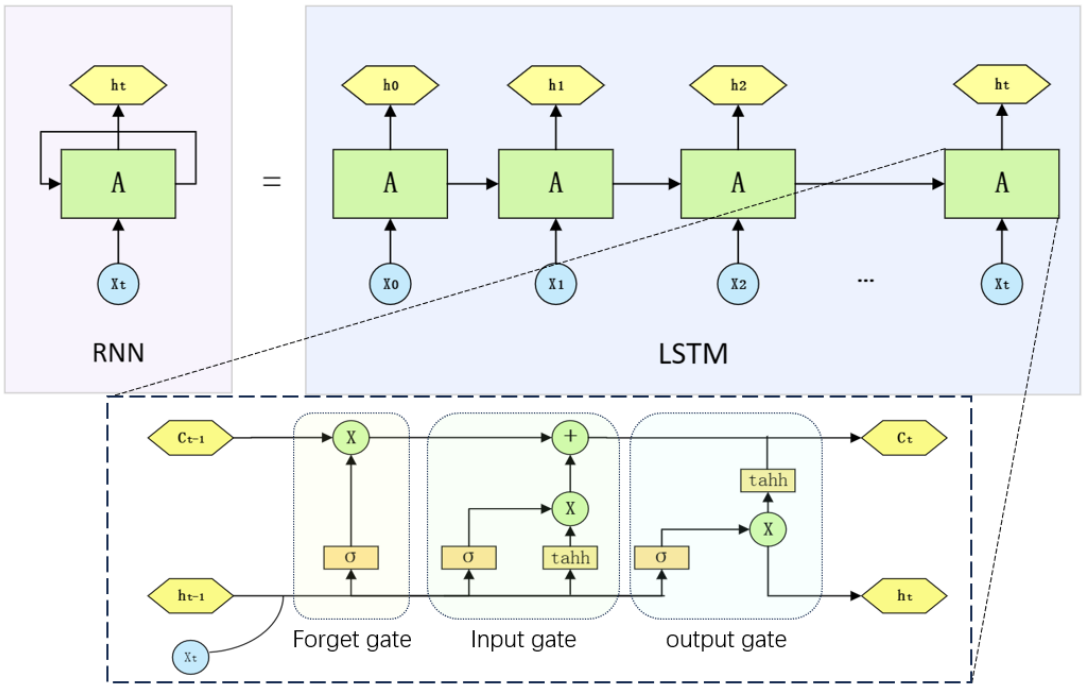


Figure 10: structure chart of LSTM

From the figure above, we can know that based on the RNN model, the LSTM model solves the problem of RNN short-term memory by adding gates, so that the recurrent neural network can use the long-distance time series information with high efficiency. The LSTM adds three logic control units, namely Forget Gate, Input Gate and Output Gate, to the RNN infrastructure, and each of them is connected to a multiplication element. By setting the weights at the edges of the memory unit of the neural network connected with other parts, we can control the input and output of the information flow and the state of the cell unit.

The functions of these three doors are:

- Input gate i: How much information does the candidate state at the current time need to be saved
- Forgetting Gate i: How much information does it take to forget to control the internal state of the previous moment

- Output gate i: How much information needs to be output to the external state h to control the internal state at the current moment h_t

When $f_t = 0$, $i_t = 1$, the memory unit clears the historical information and writes the candidate state vector \ widetilde $\widetilde{c_t}$, but at this time, the memory unit c_t is still related to the historical information from the previous moment, when $f_t = 1$, $i_t = 0$, the memory unit will copy the content from the previous moment and not write new information.

The "gate" in LSTM network is a type of "soft" gate, with values between (0, 1), indicating that information is allowed to pass through in a certain proportion. The calculation method of the three gates is:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

where σ is a Logistic function with an output interval of (0,1), which is x_ The current input at time x_t, h_{t-1} represents the external state of the previous moment.

Final result display:

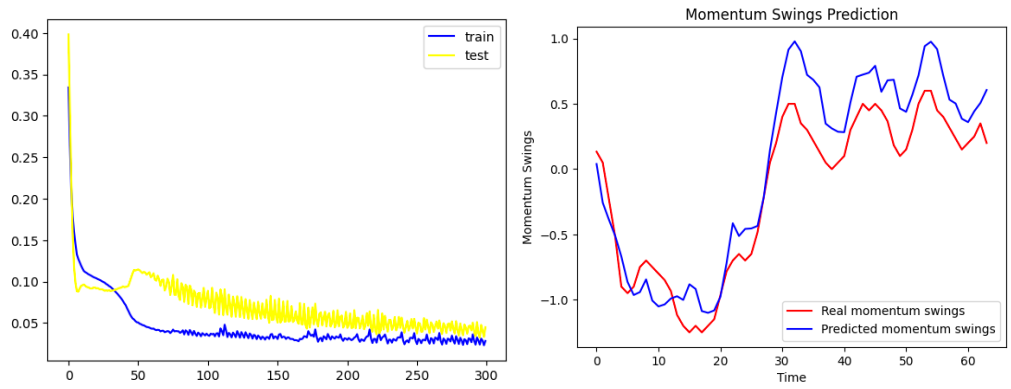


Figure 11: the results of model training and model predictions

The left image shows the model training results, and the right image shows the model prediction results.

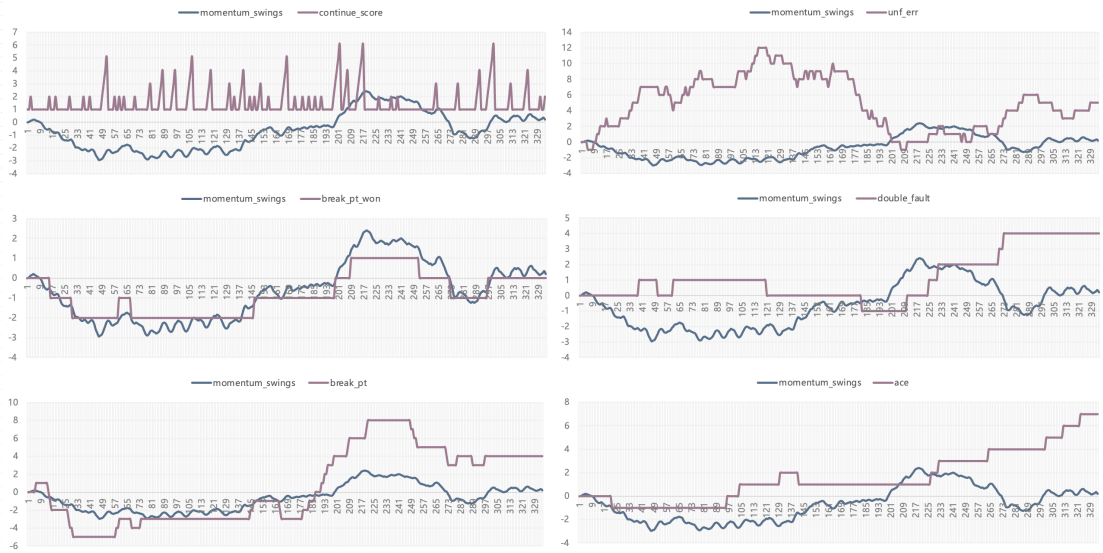


Figure 12: Changes in momentum

5.3 Suggestions for coaches-make well use of momentum

1. **Conduct in-depth analysis of competition data:**
Conduct a thorough examination of competitive data: Assess the athletes' capabilities, comprehend their strengths and weaknesses, scrutinize various data metrics, and devise tailored strategies to address diverse competitions.
2. **When momentum is totally against a palyer:**
It is advisable to recommend a deliberate approach, slowing down the pace, and wait for their opportunities.
3. **When your teammate hold the momentum:**
they could encourage the player to increase the energy and play more aggressively.
4. **When the momentum is neutral**:
both players are advised to patiently await opportunities to seize the initiative.
5. **When momentum is in favour of the player:**
Exercise caution and avoid becoming excessively confident
6. **When their opponent gained the momentum**:
The coach may want to advise his player to change the trategy to mix up their game
7. **Be well aware of the stage of momentum:**
Remain cognizant of the momentum's stage, such as who controls it, and take appropriate measures accordingly.

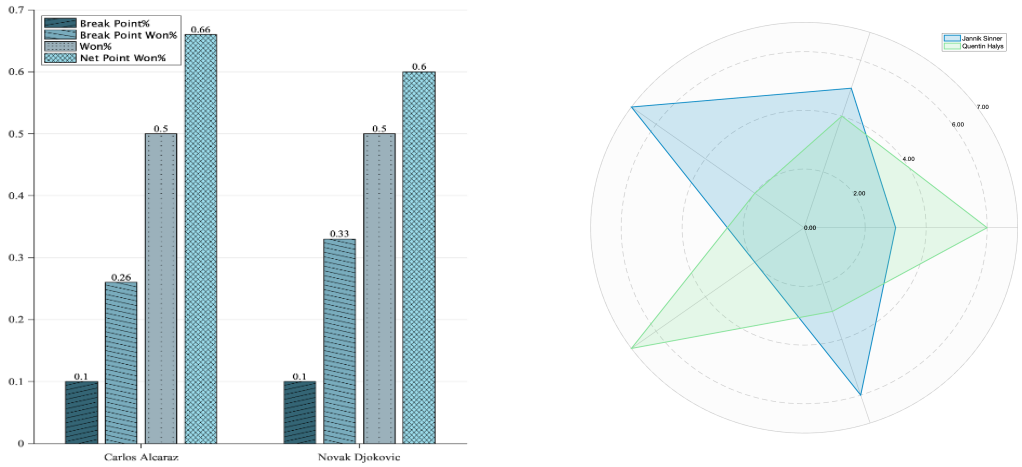


Figure 13: Difference of break_points

6 Model Generalization: A Cross-Domain Study

6.1 Data Generalization

To enhance the model's generalizability, we expanded the evaluation to include the men's doubles tournament, assessing the accuracy of variables incorporated in the tennis model training. Additional data from Wimbledon doubles tennis open was sourced from the repository: https://github.com/JeffSackmann/tennis_slam_pointbypoint/tree/master(https://github.com/JeffSackmann/tennis_slam_pointbypoint/tree/master). Given the similarity in

rules between doubles and singles, we utilized the doubles data in the Markov model for question one, calculating the leverage and momentum of each ball.

6.2 Interesting Insights from Testing the Markov Model

- **First Finding: Break points have less impact in doubles matches**

By comparing the Markov model with singles matches, we find that doubles matches have significantly less impact on certain break points than singles matches, and it is clear that singles matches are more dominant on serve and very difficult to re-turn the situation once the serve is lost.

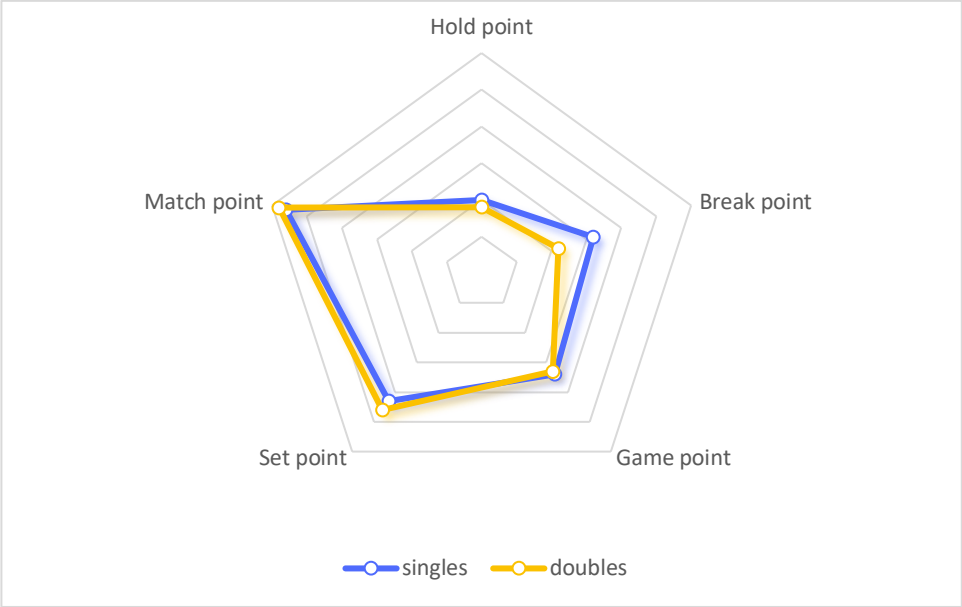


Figure 14: Radar diagram of the difference between men's and women's tennis nets

- **Second Finding: Serving is more dominant in singles matches**

Doubles matches have significantly less impact on certain break points than singles matches. This suggests that in a doubles scenario, break points do not change the pattern of the match as easily as they do in singles.

6.3 Model Leverage Adjustment in Doubles Tennis

A mechanism for dynamically adjusting weights was introduced into the model to flexibly adapt break point leverage based on real-time match conditions. This mechanism allowed us to optimize in real-time according to factors such as score, momentum fluctuations, and opponent performances. Following modifications to the break leverage in the doubles match model, illustrated in the specific scenario where Player1 faced a defeat against Player2 with scores of 5-7 and 4-6, and considering the fluctuations in momentum throughout the match, we proceed to refine and elaborate on these adjustments.

6.4 Model Generalization in other matches

For other matches, such as table tennis, badminton, etc., due to the different rules of the game, the importance of each point for the game is not the same, so it is necessary

to adjust the leverage in the model, but the model still has a good generalization ability. The following are the results of momentum fluctuation prediction for tennis doubles matches:

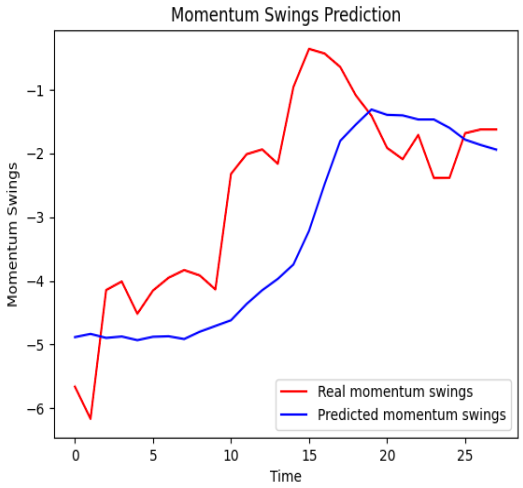


Figure 15: Momentum swing prediction

While the adjusted model shows a degree of generalizability to overarching trends, it is evident that the model's performance is not as robust as it is for individual games. To enhance the model's effectiveness, several avenues for improvement can be explored:

6.5 Possible Directions for Improving the Model:

Our model excels in depicting momentum shifts in tennis matches and can even predict the overall match outcome. However, the training dataset predominantly focuses on objective factors, overlooking the intricate interplay between momentum shifts and psychological elements. In light of this observation, our team proposes the following recommendations:

1. More Complex Features:

Consider incorporating more intricate features into the model, such as player fatigue and psychological pressure during critical moments of the match. These nuanced factors can contribute to a more comprehensive understanding of momentum shifts.

2. Cross-Sport Generalization:

Explore the potential for cross-sport generalization by analyzing the model's applicability in different ball sports. Investigate whether the model can capture universal laws governing momentum transitions across various sports, providing insights beyond the realm of tennis.

3. Conduct In-Depth Player Interviews and Surveys:

Augment the dataset by incorporating qualitative insights through in-depth interviews and surveys with players. This qualitative data can provide a more comprehensive understanding of the psychological aspects influencing momentum dynamics during matches.

By exploring these avenues, our objective is to meticulously refine the model, mitigating its limitations, and propelling it towards enhanced precision in predicting momentum shifts during multiple types of matches.

7. Future Work

7.1Strength and Weakness

7.1.1 Strength

- **Extensive dataset.** We utilized a substantial dataset encompassing point-by-point data from the men's Wimbledon tournaments spanning from 2016 to 2023. Additionally, we incorporated comprehensive odds information from matches across various years. Which serves as a reflection of the historical performance and capabilities of players, greatly improving the accuracy of our model.
- **High flexibility.** We used auto-adapted parameter to change the weight of a player's critical moment performance and general performance, making our calculation more practical and precise.
- **Greater inclusiveness.** We used MCMC to estimate the posterior distribution of parameters of interest through random sampling in the probability space instead of traditional Monte Carlo, which can deal with cumulative distribution function even if it is not integrable and it can solve the problem of "dimensionality disaster", which makes our model more practical to run with high efficiency.
- **High universality.** In our model detection, we proved that our model can be applied to multiple matches.
- **Markov's proficiency in depicting state transitions.** Markov models are especially well-suited for modeling within the context of games. This model excels at capturing the dynamics of discrete states, making it particularly effective for analyzing various scenarios and situations that unfold during a game.

7.1.2 Weaknesses

- **Rules Restriction.** Our model is to some extent influenced by the inherent rules of the sporting events, such as the structure of games, sets, and points. Hence, the universality of our findings may be somewhat compromised due to variations in the types of sporting events analyzed.
- **Psychological Restrictions.** Our model places a significant emphasis on assessing the impact of each point on the overall match. However, the reality can be more intricate, further refinement and consideration of nuanced factors will be crucial to enhance the model's ability.

7.2 Conclusion

In this article, we employed various models to predict the momentum shifts in men's singles tennis matches. Based on our model construction and result evaluation, we have drawn the following conclusions:

- The performance of players can be quantified, allowing us to discern the differences between them. In the 2023 1701 final, performance between Novak Djokovic and Carlos Alcaraz intersected three times, all coinciding with significant changes in set scores, further confirming that substantial score differentials can influence momentum.

8. References

[1] Kent, M. (2012). *The Oxford dictionary of sports science & medicine.* Oxford Oxford Univ. Pr.

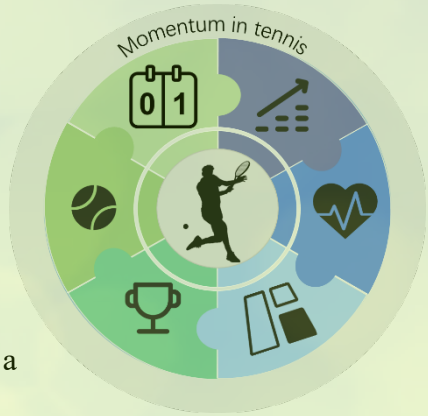
[2] Klaassen, F. J. G. M., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, *148*(2), 257–267.

[3] Dietl, H., & Cornel N. (2017). Momentum in tennis: Controlling the match. *International Journal of Sport Psychology*, *48*(365), 459–471.

[4] Dumovic, M., & Howarth, T. (2017). *Tennis Match Predictions Using Neural Networks*. Standford Univ. Pr.

[5] Forrest, D. (2003). Sport and Gambling. *Oxford Review of Economic Policy*, *19*(4), 598–611.

9. Memorandom



To: coach
From: Team# 2400384
Date: February 6th, 2024
Subject: Fully utilize "momentum" to prepare for a tennis match

---

Dear Sir or Madam:
As more and more people start to play tennis, the strategies of how to play tennis and take control of the match have reached great progress. In our work, we have established **S**core-capture and **P**erformance-identification **M**odel (SPM) and **M**omentum **S**wing **P**rediction **M**odel (MSPM). We are pleased to write this memo to show you how our models work, then we will recommend our strategy for you with our model's result presented in the end.

The Model

★**S**core-capture and **P**erformance-identification **M**odel

Our first model, SPM, is a detecting and evaluation model. Not only can it detect the point changes and calculate the swing of winning probability when different game scores occur, but also can it estimate a player's performance at any point of the match by considering a player's critical moment performance and general performance.

★**M**omentum **S**wing **P**rediction **M**odel

Our second model, MSPM, is a prediction model which can forecast the momentum of the player using relevant features such as leverage and winning rate. Based on LSTM neural network model with supervision, our model has reached an accuracy of 76.3% when we implement cross-validation to evaluate the performance of our model.

Our Strategy

- **Train the players to identify and utilize the turning points in the game.** Our model shows that the point becomes much more important ad the match progress goes. What is more, points 20:30 in game 4 set 2 is a slight turning point to pay attention to as its leverage takes 0.47, which means losing this point will lose 47% of the winning match probability.
- **Analyze the data of the opponent.** Find his weakness and make use of "momentum"