Analisi delle Componenti Principali

I sei "re" per il Six Kings Slam

Giuseppe Pio Zito 583233

1 Introduzione

1.1 Scopo dello studio

Nell'ambito della campagna per il turismo in Arabia Saudita *Visit Saudi*, il principe saudita Mohammad bin Salman vuole organizzare un torneo di esibizione di tennis nel nuovo ed avvenieristico impianto *The Venue*, mettendo in palio il premio più alto della storia del tennis (1,5 milioni di dollari per la sola partecipazione fino ai 6 milioni di dollari per la vittoria). L'iniziativa ha ricevuto l'approvazione della federazione tennistica internazionale (*ATP*) e si terrà tra il 16 e il 19 ottobre 2024.

L'idea del principe è quella di invitare al torneo i sei tennisti in attività ad aver vinto in carriera fino alla stagione 2023 almeno uno Slam (Australian Open, Roland Garros, Wimbledon, US Open), ovvero Novak Djokovic, Rafael Nadal, Daniil Medvedev, Carlos Alcaraz, Stan Wawrinka e Dominic Thiem. Il comitato tecnico organizzatore però sconsiglia questa opzione in quanto Nadal, Wawrinka e Thiem sono ormai a fine carriera e in condizioni fisiche non ottimali. A questo punto il principe si è rivolto alla nostra agenzia per individuare i tennisti adatti a garantire uno spettacolo di alto livello che possa al meglio pubblicizzare le attività turistiche che offre l'Arabia Saudita. Per tenere fede al nome scelto per questo torneo (*Six Kings Slam*), ci viene richiesto di tenere in maggiore considerazione i tennisti che hanno vinto quanti più tornei nel 2023, in particolare i vincitori degli Slam.

1.2 Caratteristiche della tabella

Abbiamo deciso di analizzare le prestazioni nella stagione 2023 dei tennisti nella top 50 del ranking ATP al 01/01/2024 (`https://www.atptour.com/en/rankings/singles?dateWeek=2024-01-01`). I fattori d'ingresso che compongono la tabella riguardano le

partite in singolare valide per tornei del circuito ATP (Slam, Master 1000, ATP500, ATP250, ATP Finals) svolti nel 2023 e sono:
- ***Matches***, partite giocate, può essere visto come un indice del numero di turni superati nei vari tornei (da 1 partita per un eliminazione al primo turno fino a 7 in caso di finale in uno Slam);
- ***Titles***, tornei vinti;
- ***Sets***, set vinti, è più significativo del numero di vittorie in quanto dà maggiore enfasi alle prestazioni negli Slam che si svolgono al meglio dei 5 set rispetto al meglio dei 3 set degli altri tornei;
- ***Tiebreaks***, tiebreak vinti, può essere visto come un indice di abilità complessiva essendo l'unica tipologia di game nel quale i tennisti si alternano tra servizio e risposta;
- ***Aces per match (Apm)***, servizi vincenti a partita;
- ***Break Points Faced per match (BPFpm)***, palle break concesse a partita, insieme al precedente riguardano i game al servizio, ma in senso opposto (da un tennista abile nei game al servizio ci aspetta un alto valore di *Apm* e uno basso di *BPFpm*);
- ***Return Games Won per match (RGWpm)***, game in risposta vinti a partita.
- ***Break Points per match (BPpm)***, palle break a favore a partita, insieme al precedente riguardano i game in risposta;
- ***Unforced Errors per match (UEpm)***, errori non forzati a partita, indica la tendenza all'errore negli scambi;

Il fattore *Titles* è stato raccolto da `https://it.wikipedia.org/wiki/ATP_Tour_2023#Titoli_vinti_per_giocatore`, i restanti da `https://www.ultimatetennisstatistics.com/statsLeaders`. Poiché i dati sono disponibili solo singolarmente, sono stati trascritti manualmente su un foglio di lavoro Excel successivamente convertito in file CSV. L'analisi è stata condotta in R e il codice completo è disponibile in appendice [4].

Oss 1. I fattori d'ingresso scelti si basano sulle tabelle al seguente link `https://www.wheeloratings.com/tennis_atp_stats_last52.html`, dove però la maggior parte dei dati sono espressi in percentuale o normalizzati rispetto a differenti quantità; perciò si è preferito raccogliere i dati dalle fonti precedentemente illustate, i quali sono normalizzati tutti rispetto al numero di partite giocate.

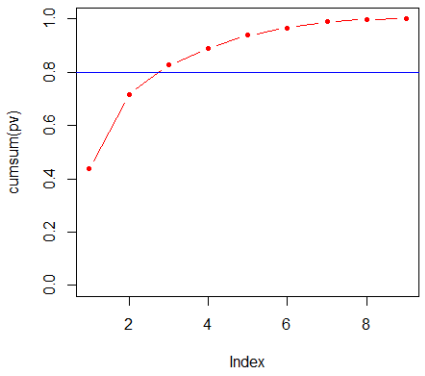
```
> head(tab,10)
                  Matches Titles Sets Tiebreaks  Apm BPFpm RGWpm BPpm  UEpm
Novak Djokovic         60      7  139        31 6.16  4.03  2.97 7.50 4.867
Carlos Alcaraz         77      6  152        19 4.02  5.00  3.37 8.33 5.065
Daniil Medvedev        84      5  150        17 6.32  4.95  3.29 7.33 7.345
Jannik Sinner          76      4  143        17 5.39  4.83  3.13 7.39 6.947
Andrey Rublev          81      2  139        19 7.33  4.95  2.85 7.27 5.889
Stefanos Tsitsipas     71      1  118        26 6.70  3.64  2.11 5.79 4.775
Alexander Zverev       80      2  124        18 7.20  4.71  2.56 7.07 6.438
Holger Rune            65      1  105        22 4.96  5.43  2.51 6.23 7.446
Hubert Hurkacz         68      2  116        32 15.06 4.11  2.02 6.06 7.735
Taylor Fritz           76      2  129        25 8.72  4.63  2.47 6.16 8.224
> summary(tab)
    Matches          Titles           Sets          Tiebreaks          Apm              BPFpm            RGWpm            BPpm            UEpm
 Min.   :26.00   Min.   :0.00   Min.   : 30.00   Min.   : 3.00   Min.   : 2.100   Min.   :3.640   Min.   :1.400   Min.   :4.400   Min.   : 1.654
 1st Qu.:41.00   1st Qu.:0.00   1st Qu.: 55.25   1st Qu.:10.00   1st Qu.: 4.455   1st Qu.:4.880   1st Qu.:2.118   1st Qu.:5.670   1st Qu.: 5.778
 Median :55.00   Median :1.00   Median : 75.00   Median :13.50   Median : 5.830   Median :5.700   Median :2.405   Median :6.200   Median : 6.939
 Mean   :53.42   Mean   :1.16   Mean   : 80.20   Mean   :14.42   Mean   : 6.111   Mean   :5.680   Mean   :2.440   Mean   :6.261   Mean   : 6.807
 3rd Qu.:62.00   3rd Qu.:2.00   3rd Qu.: 97.25   3rd Qu.:18.00   3rd Qu.: 7.298   3rd Qu.:6.205   3rd Qu.:2.717   3rd Qu.:7.035   3rd Qu.: 7.761
 Max.   :84.00   Max.   :7.00   Max.   :152.00   Max.   :32.00   Max.   :15.060   Max.   :8.260   Max.   :3.370   Max.   :8.330   Max.   :10.488
```

Si può subito riscontrare che i dati hanno diversi ordini di grandezza, sarà dunque opportuno svolgere l'analisi sulla tabella standardizzata.

Si può osservare una forte correlazione fra: *Matches* e *Sets*, infatti, per ogni partita si possono vincere da 0 a 5 set, quindi c'è una naturale dipendenza lineare; *RGWpm* e *BPpm*, dovuta al fatto che per vincere un game in risposta bisogna convertire una palla break. Inoltre, vi è una correlazione significativa fra *Sets* e *Tiebreaks*, dovuta al fatto che vincere un tiebreak significa vincere un set. Infine, possiamo osservare che i fattori riguardanti i game al servizio hanno correlazione negativa con quelli riguardanti i game in risposta e che *UEpm* sembra scorrelata da tutti gli altri fattori.

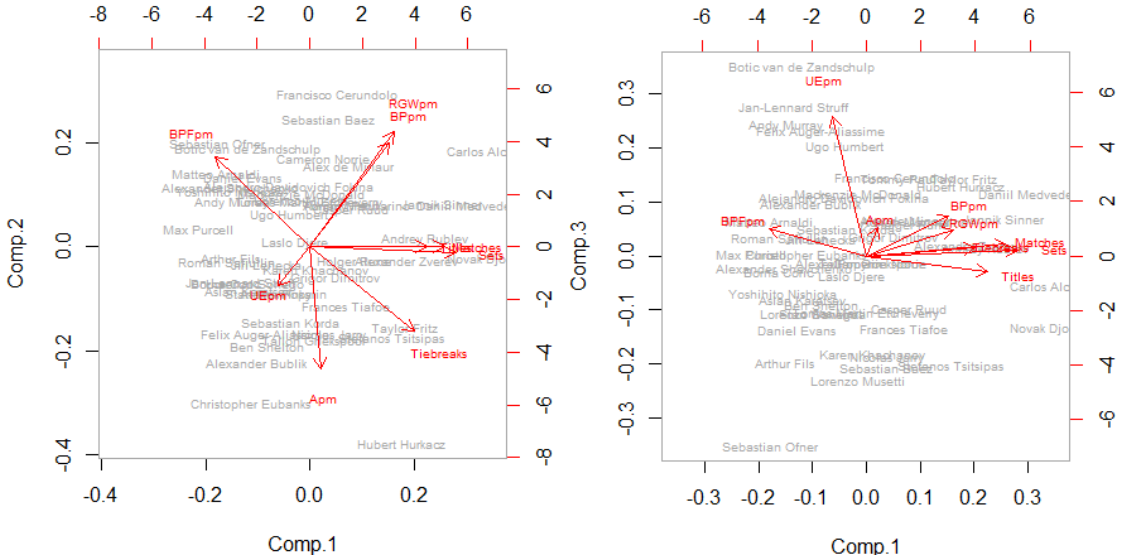
2 Analisi delle Componenti Principali (PCA)

```
> summary(tab.pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4    Comp.5     Comp.6    Comp.7     Comp.8      Comp.9
Standard deviation     1.9824075 1.5839811 1.0003978 0.74982161 0.6665133 0.49712167 0.4633383 0.271578234 0.135015568
Proportion of Variance 0.4366599 0.2787773 0.1111995 0.06247027 0.0493600 0.02745888 0.0238536 0.008194971 0.002025467
Cumulative Proportion  0.4366599 0.7154373 0.8266368 0.88910708 0.9384671 0.96592597 0.9897796 0.997974533 1.000000000
```

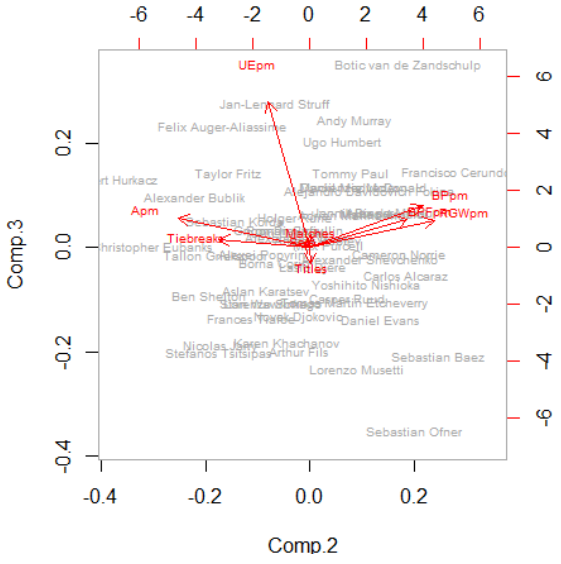


Quantitativamente si può osservare che le prime tre componenti superano la soglia empirica dell'80% di proporzione di varianza cumulata, mentre le restanti danno un contributo basso sotto al 7%. Invece, graficamente non si vede un chiaro "gomito". Allora, per rendere il modello più semplice possibile, studieremo le prime tre componenti.

2.1 Piani principali e interpretazione qualitativa



Dal piano delle componenti 1 e 2 (a sinistra) sembra che *Matches*, *Sets* e *Titles* siano associati alla prima componente, mentre *Apm* sembra associato alla seconda componente. Sono dubbi *Tiebreaks*, *BPFpm*, *RGWpm*, *BPpm* e *UEpm*.
Dal piano delle componenti 1 e 3 (a destra) sembra che *UEpm* sia associato alla terza componente e che anche *Tiebreaks* e *BPFpm* siano associabili alla prima componente. Restano dubbi *RGWpm* e *BPpm*.



Infine, il piano delle componenti 2 e 3 sembra confermare le impressioni sulla terza componente, mentre *Tiebreaks*, *BPFpm*, *RGWpm* e *BPpm* sembrano associabili alla seconda componente.

Oss 2. I versi dei fattori nei grafici precedenti sono coerenti con i segni dei coefficienti della matrice di correlazione [**??**]. In particolare, i fattori maggiormente correlati fra loro sembrano associabili alle stesse componenti principali.

Possiamo quindi concludere che, qualitativamente, *Tiebreaks* e *BPFpm* sono gli unici fattori di difficile assegnazione tra la prima e la seconda componente.

2.2 Matrice dei loadings e interpretazione delle componenti principali

```
> loadings(tab.pca)

Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
Matches    0.455                0.427  0.145  0.186  0.481  0.227  0.515
Titles     0.395        -0.101 -0.495 -0.725         0.117  0.176  0.119
Sets       0.493                0.163                0.237        -0.815
Tiebreaks  0.353 -0.356        -0.116  0.294 -0.733 -0.183 -0.177  0.208
Apm               -0.517  0.185 -0.558  0.336  0.474  0.200
BPFpm     -0.320  0.381  0.178 -0.326  0.191 -0.396  0.637  0.123
RGWpm      0.284  0.487  0.169 -0.112         0.157        -0.778  0.101
BPpm       0.270  0.442  0.259 -0.262  0.332  0.107 -0.466  0.507
UEpm      -0.110 -0.165  0.905  0.183 -0.319

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

```
> varimax(loadings(tab.pca)[,1:3])$loadings

Loadings:
          Comp.1 Comp.2 Comp.3
Matches    0.431  0.161
Titles     0.380        -0.120
Sets       0.476  0.136
Tiebreaks  0.440 -0.199  0.144
Apm        0.181 -0.393  0.340
BPFpm     -0.426  0.306
RGWpm      0.119  0.576
BPpm       0.115  0.561
UEpm              0.115  0.914
```

La matrice dei loadings (a sinistra) sembra confermare l'interpretazione qualitativa data dai piani principali, compreso il dubbio sull'assegnazione dei fattori *Tiebreaks* e *BPFpm*. Applicando una rotazione alle prime tre componenti che massimizzi la varianza dei coefficienti (a destra), si può osservare che il coefficiente di *Tiebreaks* resta alto per la prima componente, mentre diminuisce per la seconda, permettendoci così di associarlo alla prima componente. Invece, per *BPFpm* non c'è un cambiamento così significativo; dunque, poiché la seconda componente coinvolge tutti gli altri fattori relativi ai game al servizio/in risposta, per ragioni interpretative lo associamo alla seconda componente.
Ricapitolando, un'interpretazione delle nuove componenti può essere:
1. ***indice di prestazione***, con fattori associati *Matches*, *Titles*, *Sets* e *Tiebreaks* per il quale valori maggiori corrispondono a tennisti migliori;
2. ***indice di tipologia del tennista***, con fattori associati *Apm*, *BPFpm*, *RGWpm* e *BPpm* e per cui un valore negativo indica maggiore abilità nei game al servizio mentre un valore positivo indica maggiore abilità nei game in risposta;
3. ***tendenza all'errore***, con unico fattore associato *UEpm* e per cui valori maggiori corrispondono a tennisti con maggiore tendenza all'errore negli scambi.

3 Analisi dei risultati

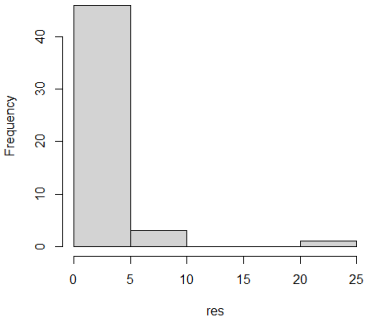
Iniziamo dando uno sguardo alla tabella con i pesi relativi ai nuovi fattori.

```
> head(tab.p[,1:3],10)
                    Comp.1     Comp.2      Comp.3
Novak Djokovic     4.902526 -0.2540766 -0.94310727
Carlos Alcaraz     4.848207  2.0749362 -0.37986319
Daniil Medvedev    4.230948  0.8943216  0.82876273
Jannik Sinner      3.578608  0.9389264  0.48364270
Andrey Rublev      3.130364  0.1843818  0.09430109
Stefanos Tsitsipas 2.205705 -1.9763507 -1.44340502
Alexander Zverev   2.594882 -0.2835612  0.14773033
Holger Rune        1.206187 -0.3070486  0.40717514
Hubert Hurkacz     2.475469 -4.2169371  0.92587159
Taylor Fritz       2.561323 -1.7576622  1.00055686
```

```
> summary(tab.p[,1:3])
     Comp.1              Comp.2              Comp.3
 Min.   :-2.9944   Min.   :-4.21694   Min.   :-2.48341
 1st Qu.:-1.6288   1st Qu.:-0.98036   1st Qu.:-0.73739
 Median :-0.2915   Median :-0.07575   Median : 0.02036
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 1.0168   3rd Qu.: 1.11449   3rd Qu.: 0.63443
 Max.   : 4.9025   Max.   : 3.28441   Max.   : 2.46962
```

3.1 Stabilità del risultato

Per valutare se ci sono osservazioni che incidono più delle altre sull'analisi, calcoliamo per ogni tennista lo scarto quadratico medio tra i suoi valori delle componenti principali ottenute dalla PCA precedente e i suoi valori previsti ottenuti dalla PCA con la tabella escludendo il tennista stesso.

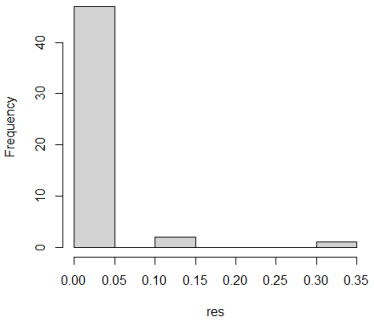



```
> tab[c(6,9,43,50),] #tennisti che più incidono nell'analisi
                        Matches Titles Sets Tiebreaks   Apm BPFpm RGWpm BPpm  UEpm
Stefanos Tsitsipas           71      1  118        26  6.70  3.64  2.11 5.79 4.775
Hubert Hurkacz               68      2  116        32 15.06  4.11  2.02 6.06 7.735
Sebastian Ofner              26      0   33         8  5.80  8.10  2.60 6.60 1.654
Botic van de Zandschulp      44      0   51         9  5.44  8.03  2.86 7.50 9.818
> res[c(6,9,43,50)]
[1]  5.330807 23.959161  6.341593  5.616384
```

A sinistra emerge che c'è un'osservazione con scarto molto alto e alcune con uno scarto significativo. A destra troviamo i tennisti che corrispondono a tali scarti e, confrontando i loro fattori d'ingresso con le quantità [**??**], possiamo osservare che: Tsitsipas ha il valore minimo della tabella in *BPFpm*; Hurkacz ha il valore massimo della tabella in *Tiebreaks* e *Apm*; Ofner ha il valore minimo della tabella per *Matches* e *UEpm*; van de Zandschulp ha il valore massimo della tabella in *UEpm*. In effetti, si può anche osservare dai piani principali [2.1] della seconda componente che Hurkacz risulti isolato dagli altri tennisti e analogamente Ofner e van de Zandschulp nei piani principali della terza componente. Questo potrebbe bastare per spiegare questi comportamenti anomali, ma per cercare di capire più nello specifico i motivi di questo fenomeno confrontiamo i coefficienti della matrice dei loadings della PCA precedente con quelli di una delle PCA anomale.

```
> loadings(princomp(tab[-50,],cor=T))

Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
Matches    0.452                0.420  0.143         0.512  0.226  0.518
Titles     0.390        -0.125 -0.494 -0.724         0.103  0.176  0.120
Sets       0.489                0.155                0.236        -0.814
Tiebreaks  0.342  0.370        -0.113  0.299 -0.688 -0.309 -0.179  0.209
Apm               0.526        -0.565  0.352  0.433  0.283
BPFpm     -0.312 -0.385  0.151 -0.350  0.210 -0.501  0.547  0.120
RGWpm      0.305 -0.463  0.184 -0.127         0.155        -0.779
BPpm       0.295 -0.414  0.242 -0.280  0.347  0.189 -0.440  0.505
UEpm              0.227  0.924        -0.268
```



A sinistra possiamo notare che l'unico cambiamento significativo è l'inversione dei segni dei coefficienti della seconda componente, il che potrebbe giustificare il comportamento anomalo. In effetti, ripetendo lo studio di stabilità precedente, cambiando il segno della seconda componente quando opportuno, si ottengono scarti molto contenuti (a destra).

Si può dunque concludere che nessuna osservazione cambia la sostanza dell'analisi a meno di un cambio di segno della seconda componente.

3.2 Esempio di applicazione

Un modo per scegliere i tennisti più adatti al *Six Kings Slam*, usando tutti i nuovi fattori ottenuti dalla PCA, può essere il seguente.

Innanzitutto, consideriamo i migliori 10 tennisti usando l'*indice di prestazione*.

Novak Djokovic	Carlos Alcaraz	Daniil Medvedev	Jannik Sinner	Andrey Rublev
4.902526	4.848207	4.230948	3.578608	3.130364
Alexander Zverev	Taylor Fritz	Hubert Hurkacz	Stefanos Tsitsipas	Adrian Mannarino
2.594882	2.561323	2.475469	2.205705	1.268771

Oss 3. Nove tennisti su dieci sono nella top 10 del ranking ATP (vedi [3] a sinistra), di cui i primi cinque nelle posizioni esatte. Manca Holger Rune (8°) rimpiazzato da Adrian Mannarino (22°).

	Matches	Titles	Sets	Tiebreaks	Apm	BPFpm	RGWpm	BPpm	UEpm
Holger Rune	65	1	105	22	4.96	5.43	2.51	6.23	7.446
Adrian Mannarino	64	3	91	15	4.59	5.80	3.00	6.28	7.484

Si può osservare come questi due tennisti hanno valori dei fattori d'ingresso molto simili e in effetti la differenza tra gli *indici di prestazione* dei due è davvero molto bassa (0.062584). Probabilmente a fare la differenza è stato il numero di tornei vinti, così, essendo un fattore importante da considerare per il principe saudita, questa variazione sembrerebbe coerente con la richiesta.

Successivamente usando l'*indice di tipologia del tennista* li suddividiamo in tre categorie: molto abili al servizio, equilibrati e molto abili in risposta. Si può fare ciò scegliendo degli opportuni range di valori oppure dividendo equamente il vettore ordinato dei pesi relativi alla seconda componente; scegliamo per semplicità la seconda possibilità, che sembra comunque ragionevole alla luce della mediana e dei quartili della seconda componente (vedi [3] a destra).

I migliori tennisti molto abili al servizio sono:

Hubert Hurkacz	Christopher Eubanks	Alexander Bublik	Ben Shelton	Tallon Griekspoor	Stefanos Tsitsipas	Nicolas Jarry
-4.2169371	-3.3769098	-2.4855874	-2.1405901	-2.0276602	-1.9763507	-1.8961180
Felix Auger-Aliassime	Taylor Fritz	Sebastian Korda	Frances Tiafoe	Alexei Popyrin	Stan Wawrinka	Aslan Karatsev
-1.8940989	-1.7576622	-1.6185502	-1.2621763	-1.0123276	-0.9949175	-0.9366917
Lorenzo Sonego	Borna Coric	Jan-Lennard Struff				
-0.8061845	-0.7662449	-0.7393959				

I migliori tennisti equilibrati sono:

Grigor Dimitrov	Karen Khachanov	Jiri Lehecka	Roman Safiullin	Holger Rune	Alexander Zverev	Novak Djokovic	Arthur Fils	Laslo Djere
-0.66252980	-0.49008174	-0.37455361	-0.31621472	-0.30704860	-0.28356115	-0.25407662	-0.23084063	0.07933689
Andrey Rublev	Max Purcell	Ugo Humbert	Casper Ruud	Tommy Paul	Daniil Medvedev	Adrian Mannarino		
0.18438179	0.39527337	0.69853415	0.79631749	0.88120449	0.89432159	0.90647345		

I migliori tennisti molto abili in risposta sono:

Jannik Sinner	Andy Murray	Tomas Martin Etcheverry	Lorenzo Musetti	Mackenzie McDonald	Yoshihito Nishioka
0.9389264	0.9577808	0.9620918	1.0254214	1.1441740	1.2015128
Alexander Shevchenko	Alejandro Davidovich Fokina	Daniel Evans	Matteo Arnaldi	Alex de Minaur	Cameron Norrie
1.2825884	1.2839066	1.5044768	1.5807492	1.7421792	1.9159382
Carlos Alcaraz	Botic van de Zandschulp	Sebastian Ofner	Sebastian Baez	Francisco Cerundolo	
2.0749362	2.1001344	2.2299768	2.7622664	3.2844071	

Infine, scegliamo per ognuna delle categorie i due tennisti meno tendenti all'errore. I tennisti in ordine crescente per *tendenza all'errore* sono:

Sebastian Ofner	Lorenzo Musetti	Sebastian Baez	Stefanos Tsitsipas	Arthur Fils
-2.483413615	-1.636942483	-1.468530566	-1.443405023	-1.393007379
Nicolas Jarry	Karen Khachanov	Daniel Evans	Frances Tiafoe	Novak Djokovic
-1.327477030	-1.281625853	-0.974174578	-0.954177585	-0.943107269
Lorenzo Sonego	Tomas Martin Etcheverry	Stan Wawrinka	Casper Ruud	Ben Shelton
-0.771402603	-0.754347888	-0.749054372	-0.702411615	-0.642657487
Aslan Karatsev	Yoshihito Nishioka	Carlos Alcaraz	Laslo Djere	Borna Coric
-0.571327984	-0.487414346	-0.379863193	-0.277491200	-0.205707497
Alexander Shevchenko	Tallon Griekspoor	Alexei Popyrin	Cameron Norrie	Christopher Eubanks
-0.154963355	-0.110710672	-0.103084279	-0.081576902	0.008254205
Max Purcell	Andrey Rublev	Alexander Zverev	Jiri Lehecka	Grigor Dimitrov
0.032456499	0.094301091	0.147730327	0.217363284	0.233899199
Roman Safiullin	Sebastian Korda	Holger Rune	Adrian Mannarino	Matteo Arnaldi
0.241284708	0.356270437	0.407175138	0.458299082	0.461389988
Alex de Minaur	Jannik Sinner	Alexander Bublik	Alejandro Davidovich Fokina	Daniil Medvedev
0.476211356	0.483642695	0.684697531	0.752154563	0.828762734
Mackenzie McDonald	Hubert Hurkacz	Tommy Paul	Taylor Fritz	Francisco Cerundolo
0.829718327	0.925871595	0.998658747	1.000556862	1.040708365
Ugo Humbert	Felix Auger-Aliassime	Andy Murray	Jan-Lennard Struff	Botic van de Zandschulp
1.430104829	1.638542067	1.721599994	1.958601734	2.469619416

Allora la scelta ricade su **Stefanos Tsitsipas** e **Hubert Hurkacz** (molto abili al servizio), **Novak Djokovic** e **Andrey Rublev** (equilibrati), **Carlos Alcaraz** e **Jannik Sinner** (molto abili in risposta). In questo modo possiamo garantire uno spettacolo di alto livello e variegato nella tipologia di gioco.

4 Conclusioni

Con la riduzione del modello ottenuta tramite la PCA, per alti valori relativi alla prima componente principale corrispondono tennisti con buone prestazioni. Questo risulterebbe sufficiente per la selezione dei sei "re" per il *Six Kings Slam* e la scelta ricadrebbe su: Djokovic, Alcaraz, Medvedev, Sinner, Rublev e Zverev, che corrispondono rispettivamente ai numeri 1, 2, 3, 4, 5 e 7 del ranking ATP. La scelta di Zverev (7°) al posto di Tsitsipas (6°) probabilmente è dovuta al fatto che il primo ha vinto più tornei rispetto al secondo (vedi [**??**]), ma essendo questo uno dei criteri principali richiesti dal principe saudita la scelta risulta coerente.
Disponendo anche della seconda e della terza componente principale, è possibile effettuare una selezione più accurata che consideri le caratteristiche tecniche dei tennisti. Ad esempio, effettuando la scelta nelle modalità descritte nel paragrafo precedente otteniamo comunque che cinque partecipanti su sei sono nella top 6 del ranking ATP, con la sola eccezione di Daniil Medvedev (3°) che è stato escluso nell'ultima fase del processo decisionale a causa dell'alto *indice di tendenza all'errore*: in effetti, il suo valore di *UEpm* pari a 7.345 è più alto del valore medio per l'intera tabella (vedi [**??**]). Però, andrebbe sottolineato che ciò è dovuto puramente alla modalità di scelta dei partecipanti: infatti, Medvedev è terzo per *indice di prestazione*.

In conclusione, avendo dimostrato nell'analisi precedente l'efficacia del modello, il principe Mohammad bin Salman e il comitato tecnico organizzatore si ritengono soddisfatti del lavoro svolto, in quanto il problema è stato ridotto da 9 a 3 componenti effettuando una massiccia riduzione dimensionale del problema.

Appendice

Script R

```
tab <- read.csv2("tabella.csv",row.names = 1)
 #analisi preliminare della tabella
head(tab,10)
str(tab)
summary(tab)
library(corrplot)
corrplot(cor(tab),"number",number.cex = 0.8)
tab.pca=princomp(tab,cor=T) #calcolo le componenti principali
summary(tab.pca)
 #visualizzo l'andamento della proprzione di varianza cumulata
pv=(tab.pca$sdev^2)/(sum(tab.pca$sdev^2))
plot(cumsum(pv), type = "b",col= "red",lwd=2,ylim=c(0,1),pch=20)
abline(0.8,0,col="blue")
 #visualizzo i piani principali relativi alle prime tre componenti
biplot(tab.pca,col=c("darkgray","red"),cex=0.6)
biplot(tab.pca,col=c("darkgray","red"),choice=c(1,3),cex=0.6)
biplot(tab.pca,col=c("darkgray","red"),choice=c(2,3),cex=0.6)
 #studio la matrice dei loadings
loadings(tab.pca)
varimax(loadings(tab.pca)[,1:3])$loadings
 #valuto la stabilità del risultato
res=rep(0,50)
for(i in 1:50){
  tab.r=data.frame(scale(tab))[-i,]
  tab.r.pca=princomp(tab.r)
  tabp=predict(tab.r.pca,newdata=data.frame(scale(tab))[i,])[1:3]
  res[i]=mean((tabp-predict(tab.pca)[i,1:3])^2)
}
hist(res)
round(res,2)
tab[c(6,9,43,50),] #tennisti che più incidono nell'analisi
res[c(6,9,43,50)]
#cosa succede alle componenti se rimuovo uno di questi tennisti?
loadings(princomp(tab[-50,],cor=T)) #cambia il verso della seconda componente
#verifico se ciò accade anche per gli altri
res=rep(0,50)
for(i in 1:50){
   tab.r=data.frame(scale(tab))[-i,]
   tab.r.pca=princomp(tab.r)
```

```
   tabp1=predict(tab.r.pca,newdata=data.frame(scale(tab))[i,])[1:3]
   tabp2=c(tabp1[1],-tabp1[2],tabp1[3])
   res[i]=min(mean((tabp1-predict(tab.pca)[i,1:3])^2),
              mean((tabp2-predict(tab.pca)[i,1:3])^2))
}
hist(res) #adesso è stabile
 #esempio di applicazione
tab.p=predict(tab.pca) #assegno i pesi relativi alle componenti principali
head(tab.p[,1:3],10)
summary(tab.p[,1:3])
tab[c(8,22),] #confronto i fattori iniziali relativi a Rune e Mannarino
rev(sort(tab.p[,1]))[1:10] #i dieci migliori tennisti per la prima componente
sort(tab.p[,2])[1:17] #tennisti molto abili al servizio
sort(tab.p[,2])[18:33] #tennisti equilibrati
sort(tab.p[,2])[34:50] #tennisti molto abili in risposta
sort(tab.p[,3]) #tennisti tendenti all'errore in ordine crescente
```