MONASH
BUSINESS
SCHOOL

MAKING AN IMPACT

Peimin Lin
Master of Business Analytics

Supervisor Stephanie Kovalchik
Data Scientist

Report for
Acme Corporation

17 June 2021

Department of Econometrics & Business Statistics

(03) 9905 2478
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

1 Introduction

What positions should players stand and get a better impact on the serve return? Are there any strategies that the players used during their tennis games? As we know, the serve return is also important in tennis, however, there are lots of tennis analysis done by data scientists before except return impact analysis because the positions of the data containing the 3D position were not easy to collect and there were not too many samples for analysis. However, in the early 2000s return serve positions' data began to be collected by some profession tracking systems like Hawkeye and since then, that data went public that more and more serve return analysis appeared(Stephanie 2021). In the project, we are going to explore a model for the return impact position of the professional male players using recently go public tracking data summaries on the ATP Tour websites of the 2D position of the ball at the time of return impact containing 84 male players, 1287 tennis matches from 2018 to 2020. Find out the relationship between the return impact and serve number, serve type, tennis court types, left-hand and right-hand user player and servers through the RETURN IMPACT shiny dashboard. According to the result, it could provide some useful training strategy to coach during the training, players could forecast the positions should stand to defence during the match.

2 Motivation

The motivation of this project is to develop a generative model for return impact positions of professional male players, discover the relationship between the return position and the typical spatial characteristics of the return impact position of men's player. Furthermore, a shiny dashboard was designed according to the data used in this project. Based on the dashboard, the project will check into the more detailed question. For example, does the serve number, serve direction, surface type make a separate influence for their return position? What about only two of those conditions? Or will all different conditions largely change their position? In the top 100 male tennis players, there are some left-handed players. Does their return position have a larger difference than the right-hand user player?

3 Overview the dataset

The data provided by Zelus Analytics includes return impact for returned points in ATP singles matches total 84 male players, 1287 tennis matches between 2018 and 2020, There are 25 variables and 126455 observations in this data set and each observation refers to a single point within a match.

The origin data can be found in Github, however, there are only some variables used in this project. The Table1 shows the main variables used in the project and their range.

The position uses (X,Y,Z) to represent, which is the length, lateral and height of the position, the center of the net use (0,0,0). However, the project discovered the 2D position of the ball at the time of return impact which only used the length and lateral variable. Figure 2 provided the visualisation of the tennis court, as players do not always hit the ball inside the tennis court that the X and Y range are over the length and lateral of the tennis court.

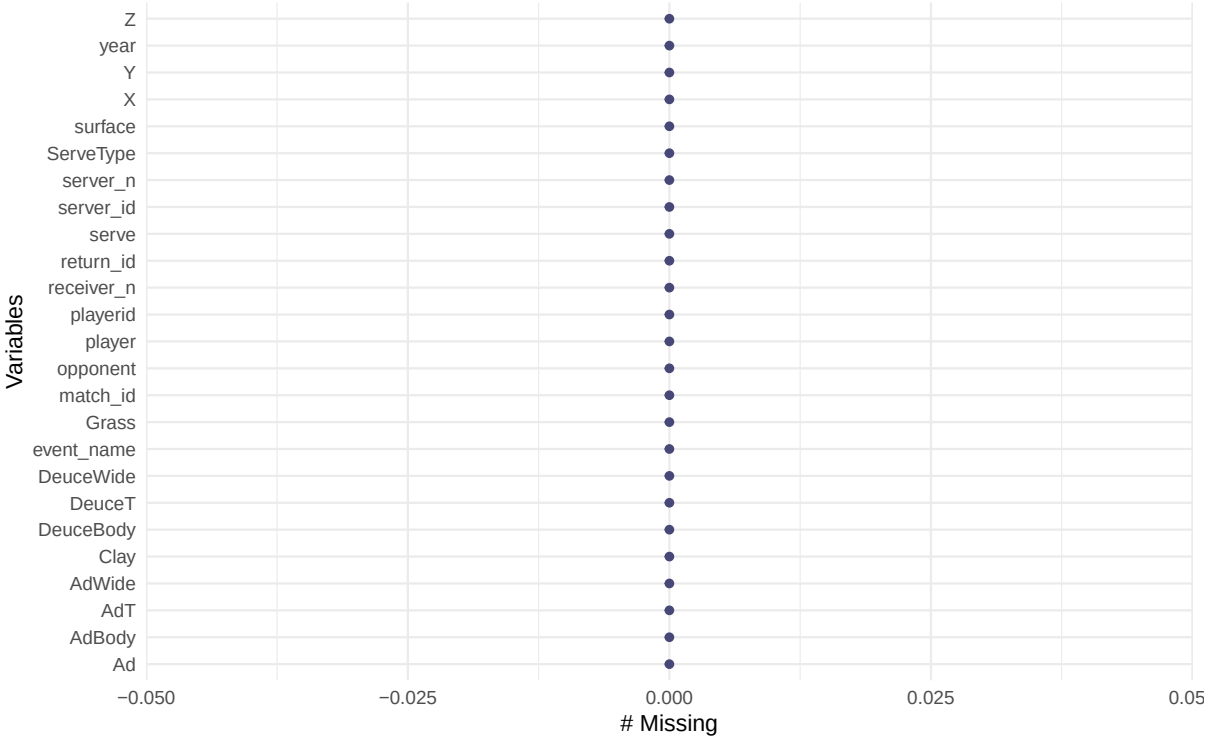

Figure 1: *Missing value check*

In tennis rules, the first serve will not lose points if there is an error, however, it will lose points on the second serve if there is a mistake(USTA n.d.). In general, there is no distinction between the first and second serves under the tennis rules. As the second serve will have a probability to lose points, players will have strategies to serve in the first and second serve. Thus, the project will have a further discussion about the serve number.

There are three types of tennis surface court now, one is hard court which is built of asphalt and concrete, one is grass which is the fastest of the tennis surface court among three and the last one is clay court, the slowest but makes the ball bounce higher. Different players have different strengths and strategies in different surface courts that this variable will also have further research.

Table 1: *Mainly Variable Used*

Variable	Description
X	Length of the Position(-23.77, 23.77)
Y	Lateral of the Position(-11.89, 11.89)
Serve	Serve Number(First Serve, Second Serve)
Player	Tennis Player Name(84 Players)
Surface	Surface Type(Hard, Clay, Grass)
ServeType	Serve Direction(AdBody, AdWide, AdT,DeuceBody, DeuceWide, DeuceT)

For the serve type, there are separate by the center line that deuce court which is serve the ball on the right side of the court and Ad court is on the left side of the court. For more precise division of the position, there are T, Wide and body, T is the serve position around the center of the court, Wide is around the edge of the court and body is in the middle of the Ad or Deuce serve position. From Figure 1, there is no missing value in the data set, so we omit the data wrangling this step and use the data directly.

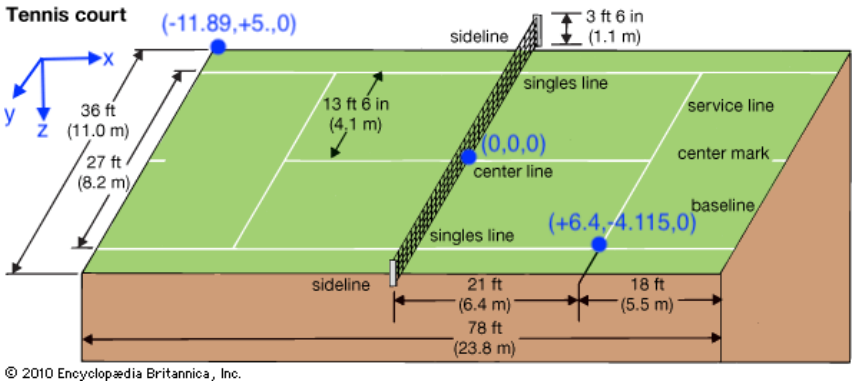

Figure 2: *Tennis Court*

4 What is the return impact

Before we start, there is a short introduction of tennis that will make you have a better understanding about the project. Serve in tennis is when a player uses a tennis racquet to hit the tennis ball and serve return if the receiver hits off of their opponent's serve. Return impact position talking in this project is when the receiving player makes contact with the ball on the serve return. Figure 3 shows the return impact position in a tennis match.

Figure 3: *Return Impact*

Table 2: *X and Y mean in difference types of surface*

surface	X	Y
Clay	-14.07975	-0.2812424
Grass	-12.89163	0.1748002
Hard	-12.82446	0.0475427

5 How variables influence player's return impact

The first panel in shiny dashboard is a marginal distribution plot used to find out how the surface, serve type and serve number influence the player's return impact position. There are three selections on the top that can change the surface, serve type in different players, for the serve number it was set as default and clarified in two colors that can better compare.

5.1 Surface

In order to make a better conclusion, table 2 shows the average length and lateral return impact position in three types of the surface. In this part, the outstanding players in three surfaces will become the examples to analyze. The French Open is the only Grand Slam tournament played on a clay court, and Rafael, who has won 10 titles, is the king(Kilit and Arslan 2018). Swat to Nadal, there is a regular range in clay court for his return impact position is (-14,-18) and the average X is -16.28605, Y is 0.7851137 which is a large difference to the average position. The serial Wimbledon winner, Roger Federer has a better performance in grass court according to his record. It can be found in the shinyapp that his return impact position in grass court is around 10 to 13. As the data did not have as many observations as hard and clay court, it will compare to the hard court for analyses.

Table 3: *X and Y mean in difference number of serve*

serve	X	Y
1	-13.35178	0.2766256
2	-12.51856	-0.3863715

Federer's performance on hard courts can be found that the range is around 9 to 14. Grass courts have the highest speed but lower bounce because of the soft grass.

Combined with the characteristics of different surfaces and the average point can draw the conclusion that a player would stand in the farthest place in a clay court match to serve the ball because the clay court is considered to be the slowest surface among three and it has higher bounce. Players stand farther and would have more time to have the reaction and serve the ball. Players will stand near the center court in the hard courts because the hard courts have medium speed and the highest bounce due to the hardest surface. Grass courts have the highest speed but lower bounce because of the soft grass, players' return impact positions are to the middle compared to the hard court and the clay court.

5.2 Serve Number

The shinyapp show the difference serve number in two colours and there is no need to select. Selected some top players for visualisation and can be found that second serve return impact positions were more forward to the center. The table 3 confirmed this view, the second serve return impact position was moved forward to the center than the first serve. It is not hard to find why there is a difference between the first and second serve, because the first serve will not lose marks unless there is a fault in the second serve. Players usually will hit the ball with as much power, skill and deception as they can to win the point. If there is a fault in the first serve, players jump to the second serve that have a risk to lose points and they will have conservative strategies that hit less hard. According to these normal strategies, players will adjust their return serve position that first serve stand farther and have more flexibility, second serve stand near the center because the ball is less hard and variability.

5.3 Serve Type

From Figure 4 can find the deuce side position will be closer to the center of the court and the Ad side is farther. As there is not a big difference between the two sides, there is an assumption that the left hand user players and right hand user players will have a large difference in the two sides. There are some ATP left hand user in the top 100 ranking also in the data as well, for example, Radial Nadal, Albert Ramos Vinolas, Adrian Mannarino, Guido Pella and Feliciano Lopes those who only

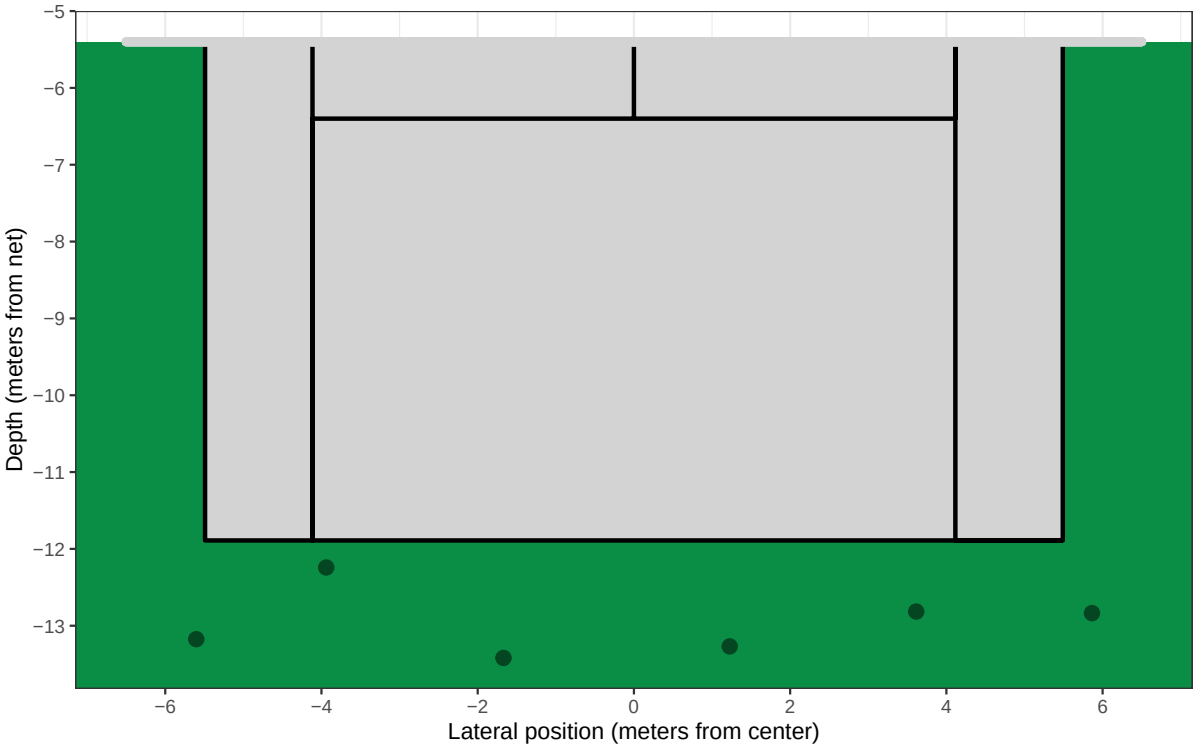

Figure 4: *The average position of different serve type*

use left hand, Denis Shapovalov, Ugo Humbert and Cameron Norrie who are left hand user also two handed backhand player. The second panel of the shiny dashboard shinyapp can be used to analyze this question as well. Select two left hand user players or select one left hand user player and one right hand user player to compare. It can be found that the left hand user will serve closer to the court center in Ad serve compared to the deuce serve.

Overview the influence to the return impact separately that player will have the farthest return impact position in clay surface and the hard court is the closest to the center. For the lose point reason, players have a closer position to the center to serve the second serve. Left hand user will serve the ball close to the center in Ad side while the right hand user player will serve the ball close to the center in the deuce side in general.

6 Model Selection

For those conclusions were all from the visualisation tool, the first and second panels of the Return Impact dashboard. What about the mixture influence of the variable to the return impact positions?

Clustering is used to reducing the dimension of the observation space, cluster analysis also called unsupervised analysis, there most common methods is K-means clustering for using to group observations into a set of K groups, K-means attempts to classify observations into mutually exclusive groups or clusters, so that observations within the same cluster try to be similar. Suppose each cluster is the center of the cluster and after multiple unsupervised calculations to find out the best result of the center.

Another common clustering algorithm is Hierarchical clustering, which is used to confirm the cluster number of a data set. Difference from k means, it creates a hierarchy of clusters and no need to specify the number of clusters up front. In addition, its results can be easily visualized using a dendrogram to confirm the number of the components(bradley, 2020). Beside those two types, model-based clustering is used where observations have a probability of belonging to each cluster. Gaussian mixture model is a probabilistic model which suppose all data points are generated from a mixture of a finite number of Gaussian distributions whose parameters are unknown. The GMM object implements the Expectation Maximization algorithm to fit the Mixture Gaussian model and it can draw confidence ellipsoids for multivariate models and calculate Bayesian information criteria to evaluate the number of clusters in the data(scikit-learn n.d.).

Summarizing those visualisation characteristics, and finding out the relationship between the 2D outcome and the multiple variables, Gaussian Mixture Model would be the best choice in this project.

6.1 Gaussian Mixture Model

$$f(x_i) = \sum_{k=1}^{G} \pi_k f_k(x_i; \mu_k, \Sigma_k)$$

Figure 5: *GMM formula*

Figure 5 where f_k is usually a multivariate normal distribution. The parameters are estimated by maximum likelihood, and choice between models is made using BIC.

$$BIC = -2log(L) + mlog(n)$$

Figure 6: *BIC formula*

Figure 6 where $log(n)$ is the maximized loglikelihood for the model and data, m is the number of free parameters to be estimated in the given model, and n is the number observations in the data.

6.2 Cluster Selection

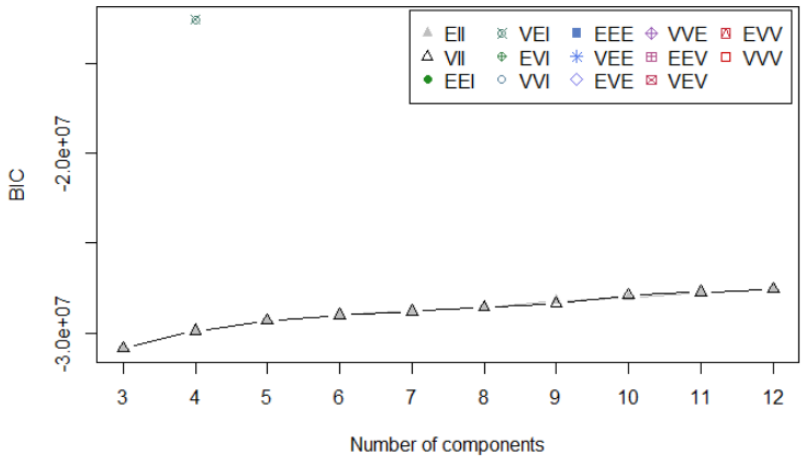

Figure 7: *BIC plot*

Although there are large models with many clusters, for better analysis, the project will set up the cluster number from 3 to 12 and calculate their BIC. Figure 7 show the trend of the BIC. It can be found the larger number of the components, the smaller BIC it has. Moving forward, the package `Rmixmod` provides a function to calculate the proportion, means variance of the cluster and find out the suitable number of components of the cluster. The function calculates 3 components to 12 components, and found the best number of the components is 11 figure **??**. However, as the plot shows the clusters have overlap figure 9 and the BIC line showed a notable growth at 8 clusters that the project will suggest components of 9 or 10 for analysis. Figure 10 shows the result of number of 9 components and figure 11 shows the variance and the mean, it can found that the variance, proportion of the number of 9 cluster is similar and not too much variance compare to the number of components 11. The last shiny panel is designed for clustering analysis, it can select multiple players in different surfaces, serve number, serve type under different number of components. Selecting the top 10 male players under ATP latest ranking, the cluster graph shows that they have the similar return impact positions under difference variable change which can provide suggestions to coach and players for the preparation of the tennis matches.

7 Conclusion

The purpose of the project is to develop a generative model for return impact positions of professional male players and explore the relationship between the return position and the typical spatial characteristics of the return impact position of men's player. The project chose Gaussian Mixture Model to

```
*** INPUT:
*****************************************
* nbCluster =  3 4 5 6 7 8 9 10 11 12
* criterion =  BIC
*****************************************
*** MIXMOD Models:
* list =  Gaussian_pk_Lk_C
* This list includes only models with
free proportions.
*****************************************
* data (limited to a 10x10 matrix) =
      X      Y
 [1,] -13.19 6.68
 [2,] -13.03 5.731
 [3,] -13.44 5.519
 [4,] -13.39 6.84
 [5,] -14.9  1.174
 [6,] -13.55 -5.544
 [7,] -13.29 6.597
 [8,] -13.24 -5.542
 [9,] -14.19 -3.489
[10,] -13.75 1.574
* ... ...
*****************************************
*** MIXMOD Strategy:
* algorithm            =  EM
* number of tries      =  1
* number of iterations =  200
* epsilon              =  0.001
*** Initialization strategy:
* algorithm            =  smallEM
* number of tries      =  10
* number of iterations =  5
* epsilon              =  0.001
* seed                 =  NULL
*****************************************


*****************************************
*** BEST MODEL OUTPUT:
*** According to the BIC criterion
*****************************************
* nbCluster   =  11
* model name  =  Gaussian_pk_Lk_C
* criterion   =  BIC(1088507.3128)
* likelihood  =  -543989.3345
*****************************************
```

Figure 8: *11 components*

analyse the relationship between the 2D outcome return impact position and the tennis surface, serve number, serve type and players, select the number of components using Bayesian information criteria (BIC) and the cluster graph in a shiny dashboard. Meanwhile, for the relationship between the return impact position and the single external factor, the player stands farthest of the court to serve the ball in clay surface court and the closest in the hard court. Players have to stand farthest in the first serve because the first serve always has a harder, faster ball that serves to hit the first serve perfectly to get the points. There are some outstanding left-hand players in the top 100 ATP ranking, they have some characteristics when they have different serve types. Right-hand user players will serve the ball more forward in deuce serve while left-hand user players are familiar in Ad serve.

The model and the discovery can be used in a coach training plan and predict the player serve return position that reduces the probability to lose points to some extent. As the data set only from 2018 to

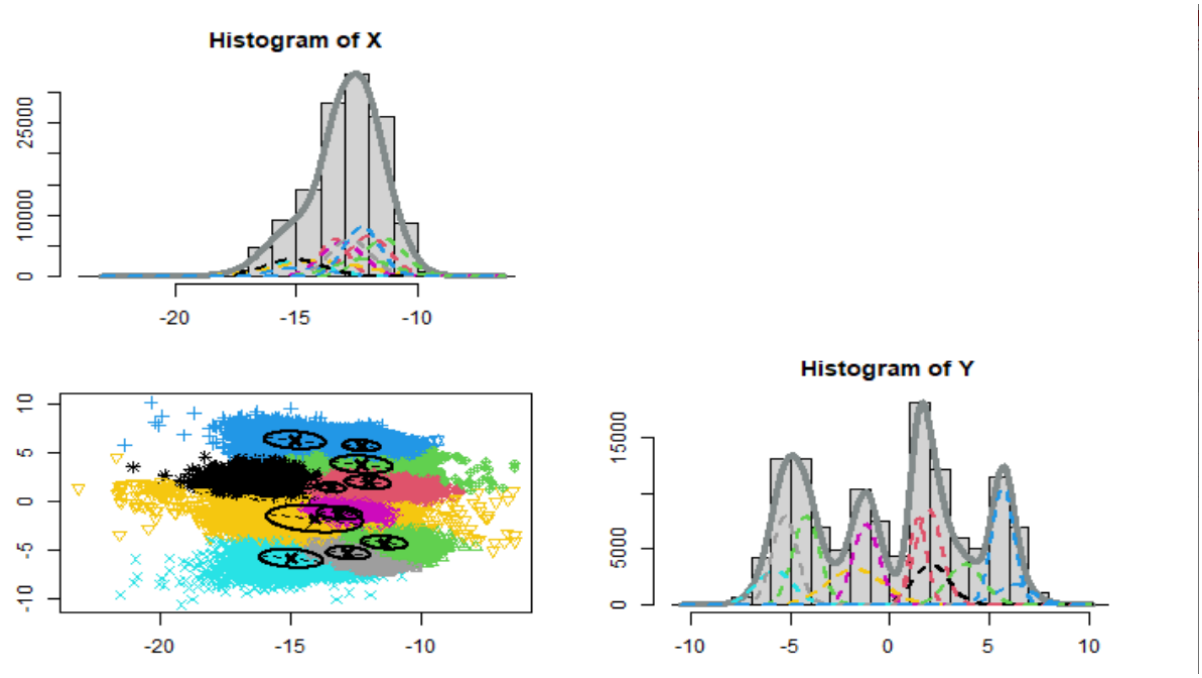

Figure 9: *Cluster plot*

2020 and the players and matches are only the top players, for the improvement of the project is expand the dataset and better to have head-to-head data set that can explore the further question, for example, does player change they return impact position when they meet difference components? There are still lots of questions yet to be discovered. Hope there will be more and more return impact position analyse in the future.

8 Acknowledgements

The dataset used in this project was provided from Zelus Analytics and written using R and `shiny`(Chang et al. 2020), the following packages were used to produce this project: `ggplot2`(Wickham 2016), `naniar`(Tierney et al. 2020), `tidyverse`(Wickham et al. 2019), `gridExtra`(Auguie 2017), `RColorBrewer`(Neuwirth 2014), `dplyr`(Wickham et al. 2020), `kableExtra`(Zhu 2020a), `shinydashboard`(Chang and Borges Ribeiro 2018), `readr`(Wickham and Hester 2020), `ggthemes`(Arnold 2019), `ggExtra`(Attali and Baker 2019),`cluster`(Maechler et al. 2021), `NbCluster`(Charrad et al. 2014),`factoextra`(Kassambara and Mundt 2020), `kableExtra`(Zhu 2020b), `golem`(Fay et al. 2021), `Rmixmod`(Langrognet et al. 2020)

```
*** INPUT:
****************************************
* nbCluster =  9
* criterion =  BIC
****************************************
*** MIXMOD Models:
* list =  Gaussian_pk_Lk_C
* This list includes only models with free
proportions.
****************************************
* data (limited to a 10x10 matrix) =
      X      Y
 [1,] -13.19 6.68
 [2,] -13.03 5.731
 [3,] -13.44 5.519
 [4,] -13.39 6.84
 [5,] -14.9  1.174
 [6,] -13.55 -5.544
 [7,] -13.29 6.597
 [8,] -13.24 -5.542
 [9,] -14.19 -3.489
[10,] -13.75 1.574
* ... ...
****************************************
*** MIXMOD Strategy:
* algorithm            =  EM
* number of tries      =  1
* number of iterations =  200
* epsilon              =  0.001
*** Initialization strategy:
* algorithm            =  smallEM
* number of tries      =  10
* number of iterations =  5
* epsilon              =  0.001
* seed                 =  NULL
****************************************


****************************************
*** BEST MODEL OUTPUT:
*** According to the BIC criterion
****************************************
* nbCluster   =  9
* model name  =  Gaussian_pk_Lk_C
* criterion   =  BIC(1099084.8517)
* likelihood  =  -549325.0945
```

Figure 10: *9 components*

```
****************************************
*** Cluster 1
* proportion =  0.0313
* means      =  -11.8471 3.3477
* variances  = |     0.4938    -0.0671 |
               |    -0.0671     0.3541 |
*** Cluster 2
* proportion =  0.1426
* means      =  -12.3430 5.6803
* variances  = |     0.5784    -0.0786 |
               |    -0.0786     0.4148 |
*** Cluster 3
* proportion =  0.1079
* means      =  -12.7936 -5.2663
* variances  = |     0.6058    -0.0824 |
               |    -0.0824     0.4344 |
*** Cluster 4
* proportion =  0.1139
* means      =  -11.3208 -4.2832
* variances  = |     0.6493    -0.0883 |
               |    -0.0883     0.4656 |
*** Cluster 5
* proportion =  0.0928
* means      =  -11.8797 1.9162
* variances  = |     0.5438    -0.0739 |
               |    -0.0739     0.3900 |
*** Cluster 6
* proportion =  0.1471
* means      =  -14.1910 3.3536
* variances  = |     4.7824    -0.6502 |
               |    -0.6502     3.4294 |
*** Cluster 7
* proportion =  0.0694
* means      =  -14.9563 -5.6759
* variances  = |     1.4618    -0.1987 |
               |    -0.1987     1.0483 |
*** Cluster 8
* proportion =  0.2005
* means      =  -13.4875 -1.3852
* variances  = |     1.8018    -0.2449 |
               |    -0.2449     1.2920 |
*** Cluster 9
* proportion =  0.0946
* means      =  -13.5347 1.5642
* variances  = |     0.3897    -0.0530 |
               |    -0.0530     0.2794 |
```

Figure 11: *variance and mean*

References

Arnold, JB (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes.

Attali, D and C Baker (2019). *ggExtra: Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements*. R package version 0.9. https://CRAN.R-project.org/package=ggExtra.

Auguie, B (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

Chang, W and B Borges Ribeiro (2018). *shinydashboard: Create Dashboards with 'Shiny'*. R package version 0.7.1. https://CRAN.R-project.org/package=shinydashboard.

Chang, W, J Cheng, J Allaire, Y Xie, and J McPherson (2020). *shiny: Web Application Framework for R*. R package version 1.5.0. https://CRAN.R-project.org/package=shiny.

Charrad, M, N Ghazzali, V Boiteau, and A Niknafs (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* **61**(6), 1–36.

Fay, C, V Guyader, S Rochette, and C Girard (2021). *golem: A Framework for Robust Shiny Applications*. R package version 0.3.1. https://CRAN.R-project.org/package=golem.

Kassambara, A and F Mundt (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.

Kilit, B and E Arslan (2018). Playing tennis matches on clay court surfaces are associated with more perceived enjoyment response but less perceived exertion compared to hard courts. *Acta Gymnica* **48**(4), 147–152.

Langrognet, F, R Lebret, C Poli, S Iovleff, B Auder, and S Iovleff (2020). *Rmixmod: Classification with Mixture Modelling*. R package version 2.1.5. https://CRAN.R-project.org/package=Rmixmod.

Maechler, M, P Rousseeuw, A Struyf, M Hubert, and K Hornik (2021). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.2 — For new features, see the 'Changelog' file (in the package source). https://CRAN.R-project.org/package=cluster.

Neuwirth, E (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer.

scikit-learn (n.d.). *Gaussian mixture models¶*. https://scikit-learn.org/stable/modules/mixture.html.

Stephanie, K (2021). *Introducing Return Impact Maps*. http://on-the-t.com/2021/01/19/Return-Impact-Maps/.

Tierney, N, D Cook, M McBain, and C Fay (2020). *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. R package version 0.6.0. https://CRAN.R-project.org/package=naniar.

USTA (n.d.). *Tennis Serving Rules: USTA*. https://www.usta.com/en/home/improve/tips-and-instruction/national/tennis-serving-rules.html.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo, and H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.

Wickham, H, R François, L Henry, and K Müller (2020). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr.

Wickham, H and J Hester (2020). *readr: Read Rectangular Text Data*. R package version 1.4.0. https://CRAN.R-project.org/package=readr.

Zhu, H (2020a). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra.

Zhu, H (2020b). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra.