MONASH BUSINESS SCHOOL

# MAKING AN IMPACT: A model of return impact in professional tennis

**Marie Curie**
Nobel Prize, PhD

**Pierre Curie**
Nobel Prize, PhD

Report for
Acme Corporation

**8 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 is lots of analysis about the tennis before and return impact analysis were not really common, mainly because the positions of the data containing the 3D position is not easy to collect and there is not too much sample for analysis. In the project, we are going to explore a model for the return impact position of the profession mal 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,

## 2 Project Goals

The serve return is the shot the receiver hits off of their opponent's serve. The position use (x,y) to represent, the center of the net use (0,0),Figure 1 provided the visualisation of the tennis court, the (x,y) is the length and lateral position. This project will develop a generative model for the return impact position of professional male players. Furthermore, the project will identify key contextual variables that may influence return impact, including but not limited to: * Serve number * Serve direction * Surface * Receiver * Server Moreover, there is a shiny dashboard designed for the project visualisation, there is a section will show the user guide about the shiny dashboard.

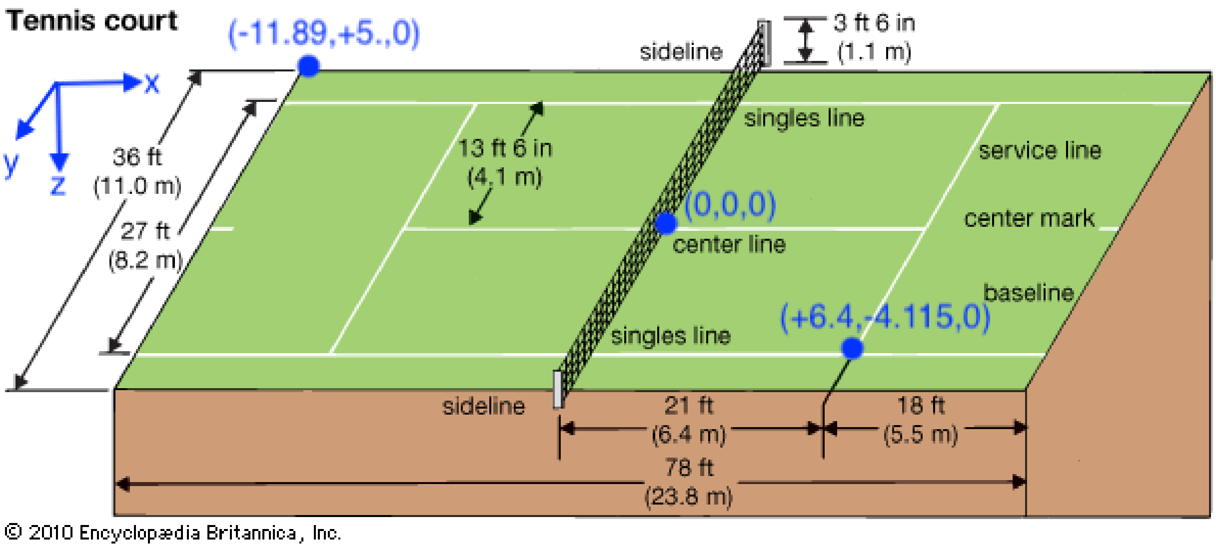



**Figure 1:** *Tennis Court*

## 3 Overview the dataset

The data set Figure2 includes return impact for returned points in ATP singles matches for events 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. From Figure3, there is no missing value in the data set, so we omit the data wrragling this step and use the data directly.

| | match_id | X | Y | Z | serve | player | opponent | playerid | event_name | year | surface | Ad |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2018/352/MS001 | -13.186 | 6.680 | 1.344 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |
| 2 | 2018/352/MS001 | -13.032 | 5.731 | 1.505 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |
| 3 | 2018/352/MS001 | -13.444 | 5.519 | 1.274 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |
| 4 | 2018/352/MS001 | -13.389 | 6.840 | 1.352 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |
| 5 | 2018/352/MS001 | -14.905 | 1.174 | 1.185 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |
| 6 | 2018/352/MS001 | -13.551 | -5.544 | 1.888 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 1 |
| 7 | 2018/352/MS001 | -13.291 | 6.597 | 1.129 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |
| 8 | 2018/352/MS001 | -13.239 | -5.542 | 1.106 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 1 |
| 9 | 2018/352/MS001 | -14.188 | -3.489 | 1.093 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 1 |
| 10 | 2018/352/MS001 | -13.746 | 1.574 | 1.275 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |
| 11 | 2018/352/MS001 | -12.880 | 6.195 | 1.395 | 1 | N. Djokovic | K. Khachanov | D643 | paris | 2018 | Hard | 0 |
| 12 | 2018/352/MS001 | -14.007 | -0.618 | 1.217 | 1 | K. Khachanov | N. Djokovic | KE29 | paris | 2018 | Hard | 0 |

| | Clay | Grass | server_n | receiver_n | ServeType | AdT | AdBody | AdWide | DeuceT | DeuceBody | DeuceWide | server_id | return_id |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 52 | 43 | Wide | 0 | 0 | 0 | 0 | 0 | 1 | 147 | 45 |
| 2 | 0 | 0 | 52 | 43 | Wide | 0 | 0 | 0 | 0 | 0 | 1 | 147 | 45 |
| 3 | 0 | 0 | 52 | 43 | Wide | 0 | 0 | 0 | 0 | 0 | 1 | 147 | 45 |
| 4 | 0 | 0 | 52 | 43 | Wide | 0 | 0 | 0 | 0 | 0 | 1 | 147 | 45 |
| 5 | 0 | 0 | 52 | 43 | T | 0 | 0 | 0 | 1 | 0 | 0 | 147 | 45 |
| 6 | 0 | 0 | 52 | 43 | Wide | 0 | 0 | 1 | 0 | 0 | 0 | 147 | 45 |
| 7 | 0 | 0 | 52 | 43 | Wide | 0 | 0 | 0 | 0 | 0 | 1 | 147 | 45 |
| 8 | 0 | 0 | 52 | 43 | Wide | 0 | 0 | 1 | 0 | 0 | 0 | 147 | 45 |
| 9 | 0 | 0 | 52 | 43 | Body | 0 | 1 | 0 | 0 | 0 | 0 | 147 | 45 |
| 10 | 0 | 0 | 52 | 43 | T | 0 | 0 | 0 | 1 | 0 | 0 | 147 | 45 |
| 11 | 0 | 0 | 43 | 52 | Wide | 0 | 0 | 0 | 0 | 0 | 1 | 108 | 61 |
| 12 | 0 | 0 | 52 | 43 | T | 0 | 0 | 0 | 1 | 0 | 0 | 147 | 45 |

**Figure 2:** *Data set overview*

```
##    match_id                 X                  Y                    Z
##  Length:126455      Min.   :-23.111   Min.   :-10.587000   Min.   :-0.004
##  Class :character   1st Qu.:-13.902   1st Qu.: -3.951000   1st Qu.: 1.191
##  Mode  :character   Median :-12.841   Median :  0.782000   Median : 1.328
##                     Mean   :-13.012   Mean   :  0.006236   Mean   : 1.344
##                     3rd Qu.:-11.867   3rd Qu.:  2.903000   3rd Qu.: 1.493
##                     Max.   : -6.423   Max.   : 10.212000   Max.   : 3.902
##      serve          player             opponent             playerid
##  Min.   :1.000   Length:126455      Length:126455      Length:126455
##  1st Qu.:1.000   Class :character   Class :character   Class :character
##  Median :1.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.408
##  3rd Qu.:2.000
##  Max.   :2.000
##   event_name             year        surface                 Ad
##  Length:126455      Min.   :2018   Length:126455      Min.   :0.0000
##  Class :character   1st Qu.:2019   Class :character   1st Qu.:0.0000
##  Mode  :character   Median :2019   Mode  :character   Median :0.0000
##                     Mean   :2019                      Mean   :0.4229
##                     3rd Qu.:2020                      3rd Qu.:1.0000
##                     Max.   :2020                      Max.   :1.0000
```

```
##       Clay             Grass              server_n        receiver_n
##  Min.   :0.0000   Min.   :0.00000   Min.   : 1.00   Min.   :10.00
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:16.00   1st Qu.:21.00
##  Median :0.0000   Median :0.00000   Median :29.00   Median :32.00
##  Mean   :0.1465   Mean   :0.05388   Mean   :28.18   Mean   :32.78
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:37.00   3rd Qu.:38.00
##  Max.   :1.0000   Max.   :1.00000   Max.   :66.00   Max.   :66.00
##   ServeType              AdT              AdBody          AdWide
##  Length:126455      Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :0.0000   Median :0.000   Median :0.0000
##                     Mean   :0.1102   Mean   :0.138   Mean   :0.1747
##                     3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:0.0000
##                     Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##      DeuceT          DeuceBody          DeuceWide        server_id
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :  1.00
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 50.00
##  Median :0.0000   Median :0.00000   Median :0.0000   Median : 98.00
##  Mean   :0.3154   Mean   :0.09277   Mean   :0.1689   Mean   : 99.94
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:151.00
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :205.00
##    return_id
##  Min.   : 1.00
##  1st Qu.:21.00
##  Median :39.00
##  Mean   :41.05
##  3rd Qu.:63.00
##  Max.   :84.00
```

# 4 How variables influence player's return impact(Implementation)

## 4.1 Model selection

Started from the basic models to find out the relationship of the return impact. * Logistic Regression

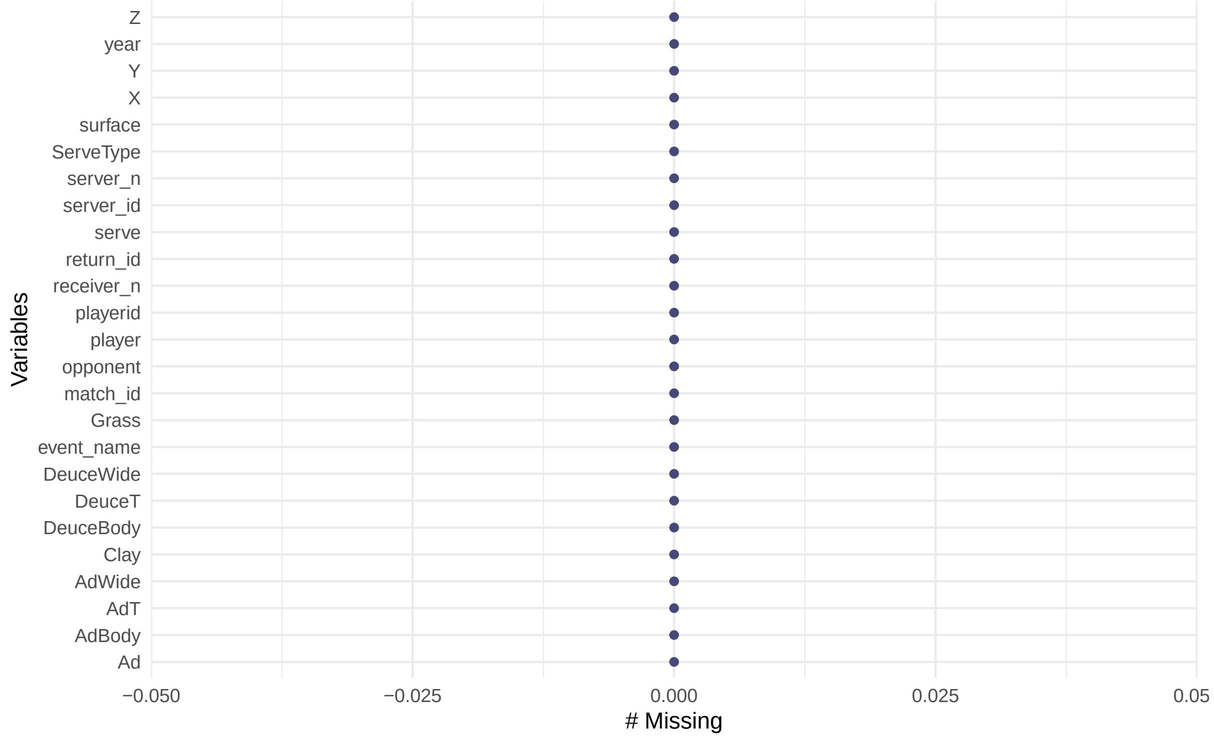



**Figure 3:** *Check missing value*

- Decision Tree
- Random Forests
- Gradient Boosting
- Gaussian Mixture Model

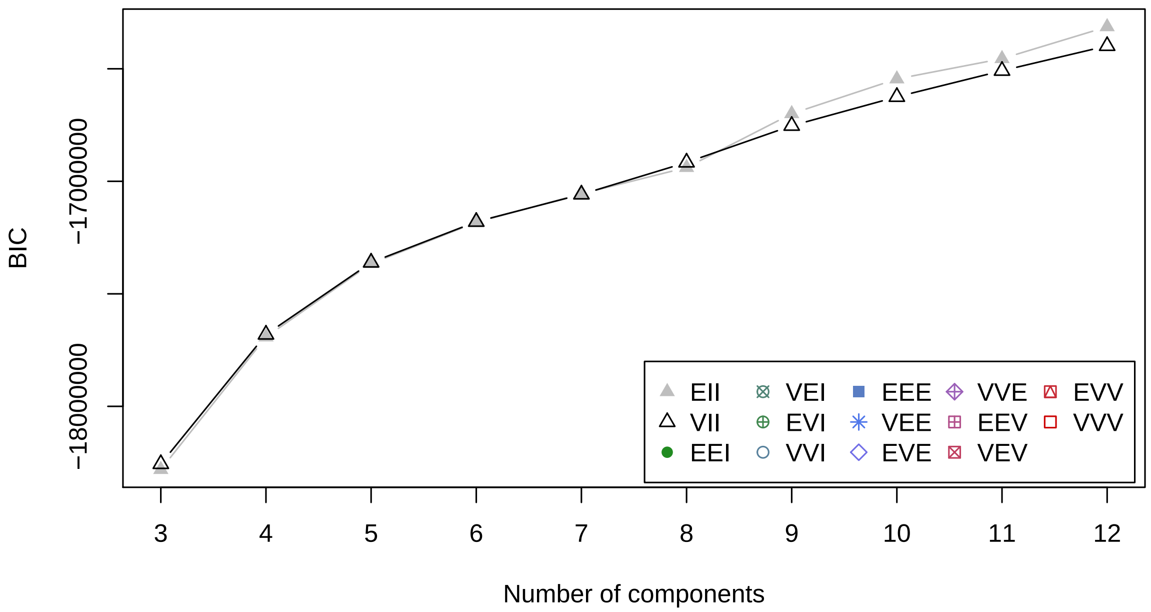
## 4.2 Cluster Selection

The number of cluster components chose for the analysis was using the `Mclust` package that calculate their BIC and Figure **??** show the trend of the BIC.

## 4.3 serve one

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EII (spherical, equal volume) model with 12 components:
##
```

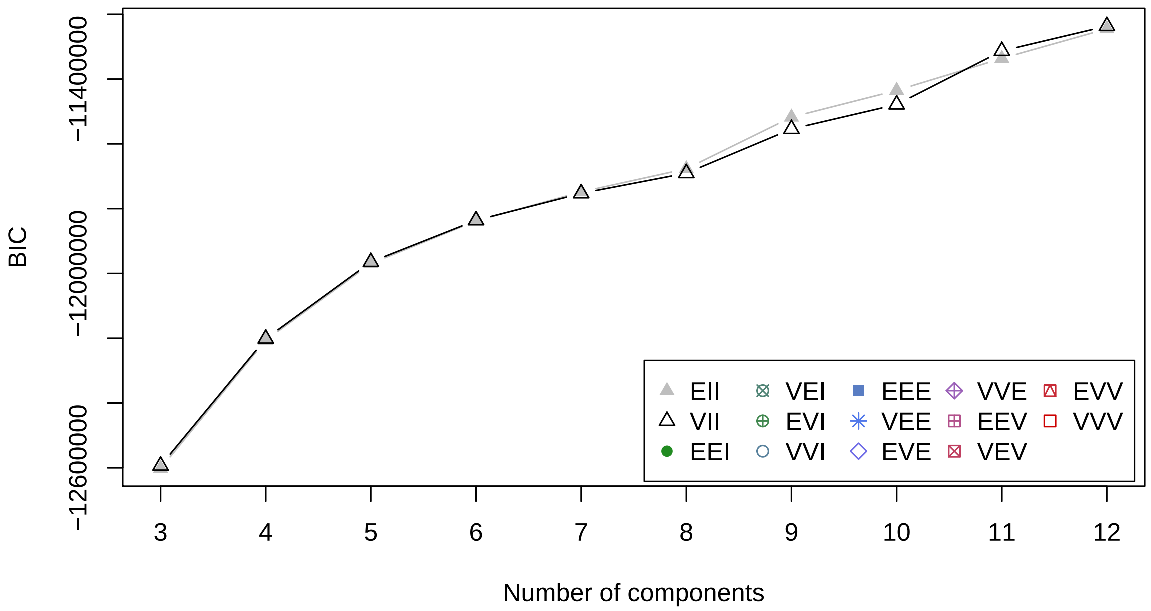
```
##  log-likelihood     n  df       BIC       ICL
##        -8154922 74883 312 -16313347 -16315565
##
## Clustering table:
##     1     2     3     4     5     6     7     8     9    10    11    12
##  4949  4781  5238 12176  7012  4843  6042  6785  7124  5452  5233  5248
```




## 4.4 serve two

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VII (spherical, varying volume) model with 12 components:
##
##  log-likelihood     n  df       BIC       ICL
##        -5616451 51572 323 -11236406 -11237669
##
## Clustering table:
##     1     2     3     4     5     6     7     8     9    10    11    12
```

```
## 3995 3702 5181 3580 5262 3816 3000 3246 3564 8060 4199 3967
```




Compare to two serve of the difference number of cluster components' BIC, the number of 9 cluster perform well in the model. Thus, it will use 9 cluster for the rest analysis. In the shiny dashboard, there is a panel can run the function below to check the change of difference components cluster by selecting serve number, serve type, player, surface type.

```
## ****************************************
## *** 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
```

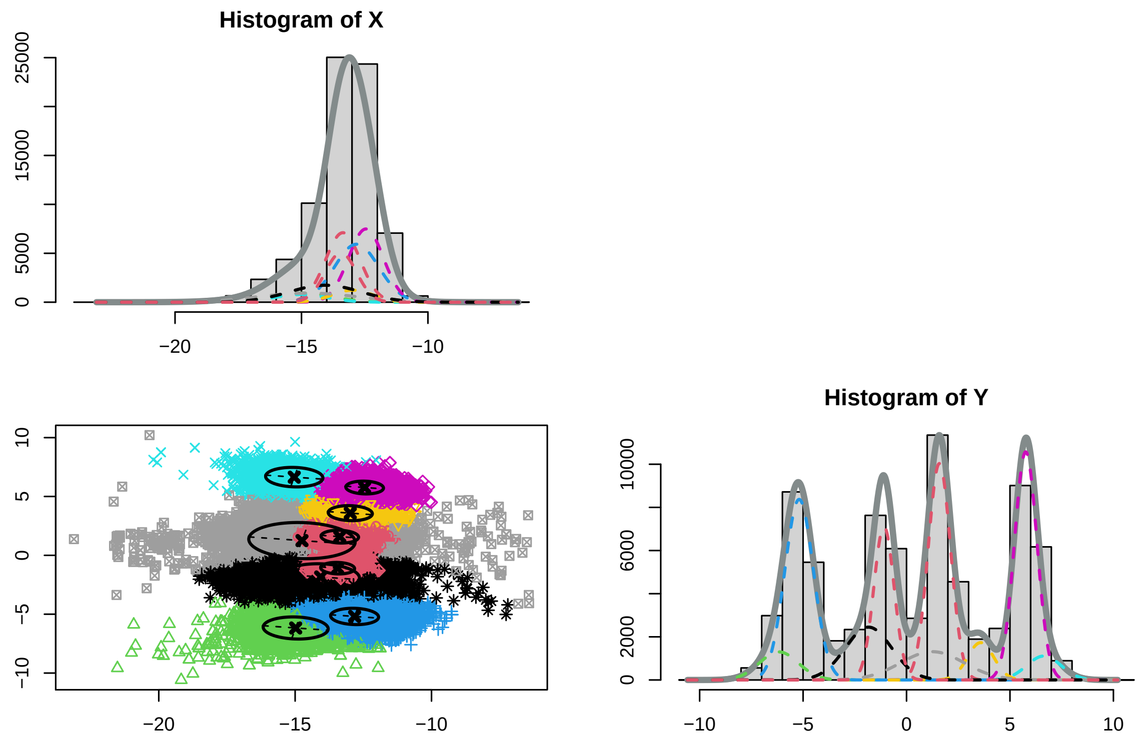
```
##  [2,] -13.03 5.731
##  [3,] -13.44 5.519
##  [4,] -13.39 6.84
##  [5,] -14.9  1.174
##  [6,] -13.55 -5.544
##  [7,] -13.29 6.597
##  [8,] -13.24 -5.542
##  [9,] -14.19 -3.489
## [10,] -13.75 1.574
## * ... ...
## ****************************************
## *** MIXMOD Strategy:
## * algorithm            =  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(594167.0499)
## * likelihood  =  -296875.8868
```

```
## ****************************************
## *** Cluster 1
## * proportion =  0.1911
## * means      =  -13.3630 1.5874
## * variances  = |     0.4737    -0.0298 |
##                |    -0.0298     0.2946 |
## *** Cluster 2
## * proportion =  0.0429
## * means      =  -14.9798 -6.1702
## * variances  = |     1.4187    -0.0893 |
##                |    -0.0893     0.8824 |
## *** Cluster 3
## * proportion =  0.2035
## * means      =  -12.8196 -5.1903
## * variances  = |     0.7719    -0.0486 |
##                |    -0.0486     0.4802 |
## *** Cluster 4
## * proportion =  0.0323
## * means      =  -15.0320 6.6422
## * variances  = |     1.1014    -0.0693 |
##                |    -0.0693     0.6851 |
## *** Cluster 5
## * proportion =  0.2019
## * means      =  -12.4514 5.7605
## * variances  = |     0.4761    -0.0300 |
##                |    -0.0300     0.2962 |
## *** Cluster 6
## * proportion =  0.0387
## * means      =  -12.9781 3.5837
## * variances  = |     0.6530    -0.0411 |
##                |    -0.0411     0.4062 |
## *** Cluster 7
## * proportion =  0.0708
```

```
## * means       =  -14.7497 1.2513
## * variances  = |     3.8013    -0.2393 |
##                |    -0.2393     2.3645 |
## *** Cluster 8
## * proportion =  0.0964
## * means       =  -14.0630 -1.8374
## * variances  = |     2.0216    -0.1272 |
##                |    -0.1272     1.2575 |
## *** Cluster 9
## * proportion =  0.1223
## * means       =  -13.4413 -1.0902
## * variances  = |     0.3865    -0.0243 |
##                |    -0.0243     0.2404 |
## ***************************************
```

```
## [1] 1
```

```
## [1] 2
```



Under the result used the GMM model, there is a further discussion of the return impact base on the player, the match intensity increase and the rest of the variables.

### 4.5 The strategy of top 3 players' return impact positions

Are the top players have large difference of the return impact positions? Or similar.

### 4.6 Any ajust strategy in the Promotion event especially the final round?

Will player stand near or far away from the court during the semi-final ground? Or final ground?

### 4.7 How are surface type influence players'performance?

As grass and clay surface type have some slightly difference and will they influence the return impact positions?

### 4.8 ATP Lefties In The Top 100 Rankings' return impact

Are the player will have similar return impact positions because they are left hand users?

### 4.9 (will they change stategy when were facing familar player)

Compare head-to-head history result, find out the player have larger win proportion and compare their each game return impact positions.

## 5 Dashboard User Guide

## 6 Conclusion