Football Players Clustering Project

January 22, 2019

1 Introduction

In this project, we intend to use the DTI scanning result on 27 tracts and 4 measurements for 195 players to cluster/classify 3 groups of players. The 3 groups stands for players who had head injury in the body contact game, players at the same spot in the body contact game but without having head injury, and player at the same spot in the noncontact game without having head injury.

The observation length along each tract are different across subjects, and thus the approaches to cluster/classify the players is to use the density functions or quantile function of these brain signal observations obtained per tract per measure per player to describe the brain activity signals, and follow that, dimension reduction tools from functional data analysis can be employed. Due to the natural constraint of density space, densities do not live in a vector space and thus, commonly used Hilbert space based methods of functional data analysis are not applicable. Therefore for the density function approach, we consider the log quantile density transformation (Petersen and Muller, 2016) to map the density function into a linear space using a continuous and invertible function, and then the functional data analysis techniques such as functional PCA can be properly implemented.

2 Data Preprocessing

1. After deleting the missing information, we have 94 pairs of players;
2. Focusing on the player injury caused by football game, we have 55 pairs of players;
3. Observing that there are duplicate players in the player's information table, we deleted all these duplicates, resulting in 47 pairs of players;
4. Merging 195 brain signals observation with football player's information table, we finally have 88 football players, including 28 players in group 1, 31 players in group 2, 29 players in group 3.

	length.sd	FA.min.sd	FA.max.sd	MD.min.sd	MD.max.sd	Da.min.sd	Da.max.sd	Dr.min.sd	Dr.max.sd
ar_l	666.2872	0.02148651	0.08777763	67.59674	3.6537918	100.98961	81.58805	91.38152	14.16697
ar_r	621.1160	0.02457896	0.07830346	67.08769	4.0613704	111.18482	88.70072	84.77857	18.86482
atr_l	541.4741	0.03715570	0.07325321	61.97450	5.2326643	88.10774	151.96600	82.28029	28.03287
atr_r	584.6707	0.03560174	0.07061179	59.04786	7.4336780	87.91863	153.82069	77.80097	26.58533
cgc_l	323.8905	0.03551219	0.05379091	67.13404	112.1586105	64.73748	180.98126	64.95049	131.85176
cgc_r	318.5215	0.02975264	0.05860940	64.99973	104.7325189	61.97992	150.91631	62.74892	125.06812
cgh_l	489.3647	0.02193005	0.08289240	82.00048	1.6274283	118.48207	113.37685	81.67307	14.52776
cgh_r	400.1337	0.01800204	0.08690106	88.61231	1.1286383	107.59221	116.14052	83.64839	14.17011
cst_l	899.5872	0.03670792	0.05599088	36.84343	2.7503629	93.28314	109.65490	77.61346	29.39867
cst_r	746.2032	0.03265296	0.05247509	43.40065	2.8655882	88.23742	101.58766	76.51024	27.47039
fma	949.2516	0.02997498	0.04670490	40.08557	0.6684907	115.21985	139.23979	87.40511	30.27645
fmi	724.7536	0.04514789	0.06068417	62.15016	1.4020793	132.28615	192.95356	109.73035	41.51449
ifo_l	929.7968	0.03599624	0.08408619	64.36125	5.6964992	121.94595	153.70695	99.66216	41.07715
ifo_r	934.5341	0.03428371	0.08615086	64.07492	2.9019792	120.41752	163.61920	100.69971	44.52732
ilf_l	752.2661	0.03792216	0.08753303	75.70351	26.4740387	107.83327	138.01425	107.45052	62.52132
ilf_r	707.2656	0.03118449	0.09272422	79.65293	14.4605665	109.58894	137.02341	113.33056	57.04777
mcp	986.4200	0.03642621	0.05760962	25.32270	0.8957110	93.84337	108.29724	71.68348	33.62092
ml_l	362.2589	0.02908403	0.06335862	32.86359	1.1647987	78.25298	96.08510	65.74730	21.67918
ml_r	354.8482	0.02812207	0.06350547	42.43338	1.5424148	76.17337	97.74573	70.64354	22.25302
ptr_l	802.3845	0.03225131	0.08648466	57.30387	1.9472568	91.27830	187.65460	101.45546	23.23825
ptr_r	772.3008	0.03362415	0.09107109	55.54160	5.3636728	97.19267	186.40613	104.98872	27.29220
slf_l	463.4023	0.05120869	0.04471693	73.49844	122.2469520	90.35890	80.83110	53.92113	146.44843
slf_r	436.6017	0.05027455	0.05277012	70.50403	121.0779498	75.73148	92.74780	62.70473	144.51885
str_l	521.7381	0.02829842	0.04389450	70.31178	20.9113542	79.78587	117.43866	57.40454	33.57278
str_r	491.9707	0.02872917	0.04769652	72.93658	20.9709008	82.80224	114.61462	55.12984	28.79467
unc_l	689.0016	0.03082693	0.06883138	83.61191	18.0491657	112.94530	112.40859	82.30031	33.58358
unc_r	666.4117	0.03104136	0.05297688	88.85549	18.3650580	133.05486	149.71131	65.34587	34.42641

Figure 1: Standard deviation for length, and range of observations across 195 subjects per tract per measure.

3 Data Summary

Since there are some players who are not having the corresponding group 2 or group 3 control information, we have three following ways to deal with the missing data (focusing on football players):

1. Keep all the data from each group separately (also delete the duplicated football players inside group 1 and group 2 separately)
2. For the first two groups, only keep the data that have both group 1 and group 2 information
3. For all the three groups, only keep the data that has no missing values on both of group 2 and group 3 information

The following table shows the resulting number of observations that also have the corresponding scanning data.

Table 1: Available observation

group	method 1	method 2	method 3
group 1	39	29	28
group 2	34	32	31
group 3	34	NA	29

The following figure shows the frequency of FPCs we have for each of the four methods in the 2 cluster analysis (left) and 3 cluster analysis (right):

```
###################### or
#only 2 clusters

#den .95:
#fPCss
# 2  3  4  5
#32 57 16  3

#q .95
#fPCss
# 2  3  4
#10 73 25

#fPCss .95
#qdt
# 4  5  6  7  8
# 3 22 34 44  5

#qdtg .95:
#fPCss
#  4  5  6  7  8
# 7 20 32 41  8

#######################
#den .99:
#fPCss
#3  4  5  6
#18 57 28  5

#q .99
#fPCss
#3  4  5  6  7
#1 24 48 34  1

#fPCss .99
#qdt
# 7  8  9 10 11
# 3 22 38 44  1

#qdtg .99:
#fPCss
# 7  8  9 10 11
# 4 21 44 38  1
```

```
#den .95:
#fPCss
# 2  3  4  5
#26 59 20  3

#q .95
#fPCss
# 2  3  4
#15 73 20

#fPCss .95
#qdt
# 4  5  6  7  8
# 3 19 36 43  7

#qdtg .95:
#fPCss
# 4  5  6  7  8
# 9 18 31 39 11

#######################
#den .99:
#fPCss
#3  4  5  6
#16 58 29  5

#q .99
#fPCss
# 4  5  6  7
#29 48 30  1

#fPCss .99
#qdt
# 7  8  9 10 11
# 2 18 35 50  3

#qdtg .99:
#fPCss
# 7  8  9 10 11
# 5 18 37 44  4
```

4 Models

Here we try for 3 models for the brain signals D_{ijk} of football player i at tract j using measurement k (i=1, ..., 88, j=1, ..., 27, k=1, ...,4):
1. Density function:
Normalize D_{ijk}, and extract density function $f(t)$ on 200 equally spaced $t \in [0.0025, 0.9975]$
2. Quantile function:
Without normalizing D_{ijk}, calculate the quantile function $q(t)$ on 200 equally spaced $t \in [0.0025, 0.9975]$
3. Log quantile function:
Normalize D_{ijk}, calculate the quantile function $q(t)$ on 200 equally spaced $t \in [0.0025, 0.9975]$, and calculate the log quantile transformed density as $-log(q(t))$

Following that, perform the FPCA on these functional predictors and obtain principal scores using 95% pve; use kmeans clustering method on the principal scores to make 3 or 2 clusters.

5 result

5.1 Cluster result for 3 groups

Table 2: Minimum 5 misrate with corresponding tract for 4 measurements in 4 models (3 clusters)

Measure	misrate				Tract			
	(d)	(q)	(qdt)	(qdtg)	(d)	(q)	(qdt)	(qdtg)
FA	0.51	0.53	0.55	0.55	cgc_r	cgc_l	cgc_r	ptr_l
FA	0.53	0.56	0.57	0.55	cst_r	fma	fmi	cst_r
FA	0.57	0.57	0.58	0.55	ifo_r	ar_r	cst_l	fmi
FA	0.57	0.58	0.58	0.57	slf_l	slf_r	fma	cst_l
FA	0.58	0.59	0.58	0.57	cgh_l	fmi	ilf_l	cgc_r
MD	0.53	0.58	0.58	0.55	unc_r	ar_l	unc_l	cgc_l
MD	0.56	0.59	0.59	0.57	mcp	cgc_r	ilf_l	atr_r
MD	0.57	0.60	0.60	0.58	ar_l	cgh_l	ml_l	cgh_l
MD	0.57	0.59	0.60	0.58	slf_l	ifo_l	cgh_l	unc_l
MD	0.58	0.59	0.60	0.59	ptr_l	mcp	cgh_r	ar_l
Da	0.53	0.55	0.52	0.56	atr_l	ml_l	fma	ifo_l
Da	0.57	0.57	0.56	0.56	ar_r	cgc_r	cst_r	ml_l
Da	0.57	0.57	0.56	0.56	fmi	str_l	unc_l	fma
Da	0.58	0.57	0.57	0.56	cgh_r	unc_l	cgc_r	ptr_l
Da	0.58	0.58	0.57	0.57	cst_r	fma	ilf_r	ilf_r
Dr	0.52	0.58	0.55	0.57	ar_r	atr_r	fma	cgc_r
Dr	0.57	0.58	0.56	0.59	str_r	fmi	ar_l	fma
Dr	0.58	0.58	0.56	0.59	atr_l	ifo_l	cst_l	ar_l
Dr	0.58	0.58	0.56	0.59	slf_l	mcp	cst_r	ar_r
Dr	0.59	0.58	0.56	0.59	mcp	slf_r	slf_l	atr_r

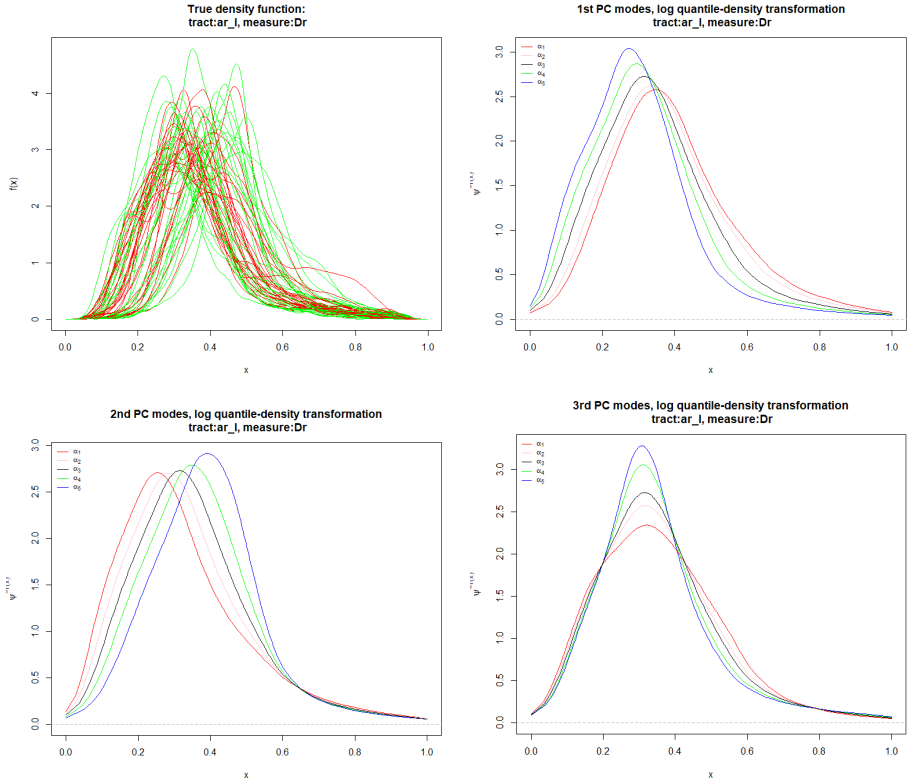
5.2 Cluster result for the first 2 groups

Table 3: Minimum 5 misrate with corresponding tract for 4 measurements in 4 models (2 clusters)

Measure	misrate				Tract			
	(d)	(q)	(qdt)	(qdtg)	(d)	(q)	(qdt)	(qdtg)
FA	0.41	0.36	0.37	0.34	cgc_l	cgc_l	ar_r	cst_l
FA	0.41	0.41	0.37	0.39	cgh_l	atr_r	cst_l	cst_r
FA	0.41	0.42	0.37	0.41	ptr_l	str_r	cst_r	mcp
FA	0.42	0.42	0.39	0.42	ar_r	unc_r	cgc_r	ar_l
FA	0.42	0.43	0.39	0.42	atr_l	ml_r	ptr_l	ar_r
MD	0.39	0.41	0.39	0.37	cst_l	ar_r	unc_r	cgc_l
MD	0.41	0.41	0.42	0.41	slf_l	atr_r	atr_r	atr_r
MD	0.41	0.42	0.42	0.41	unc_r	cgc_r	cst_l	cst_l
MD	0.42	0.43	0.42	0.42	atr_l	ml_r	ptr_r	cgc_r
MD	0.42	0.44	0.42	0.42	cst_r	str_r	slf_l	cgh_l
Da	0.36	0.42	0.37	0.39	str_r	cgc_r	ptr_r	cgc_r
Da	0.39	0.39	0.37	0.39	atr_l	str_r	unc_r	cgh_r
Da	0.39	0.40	0.38	0.39	fmi	ml_l	ml_r	ptr_l
Da	0.39	0.40	0.41	0.42	unc_r	ml_r	ml_l	fma
Da	0.41	0.41	0.42	0.42	cgh_r	cgc_l	ilf_r	str_r
Dr	0.36	0.41	0.36	0.41	mcp	atr_r	ar_l	atr_r
Dr	0.37	0.41	0.37	0.41	atr_l	cgc_r	fma	cst_r
Dr	0.37	0.42	0.39	0.41	unc_r	ar_r	cst_r	unc_l
Dr	0.39	0.42	0.41	0.42	ptr_r	cgc_l	ar_r	cgc_r
Dr	0.41	0.42	0.41	0.42	slf_l	slf_l	cst_l	fma

5.3 Visualize the PC modes

To visualize the PC modes using the ordinary fPCA directly on the density function VS back-transformed PC modes using fPCA on the log quantile transformed density function, we use the tract and measure which minimized the 2 cluster misrate using log quantile function, i.e. tract "ar_l" measure Dr.

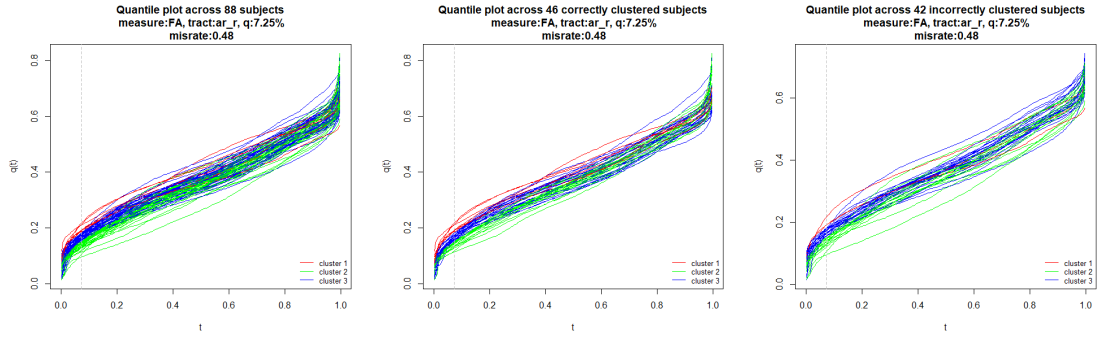


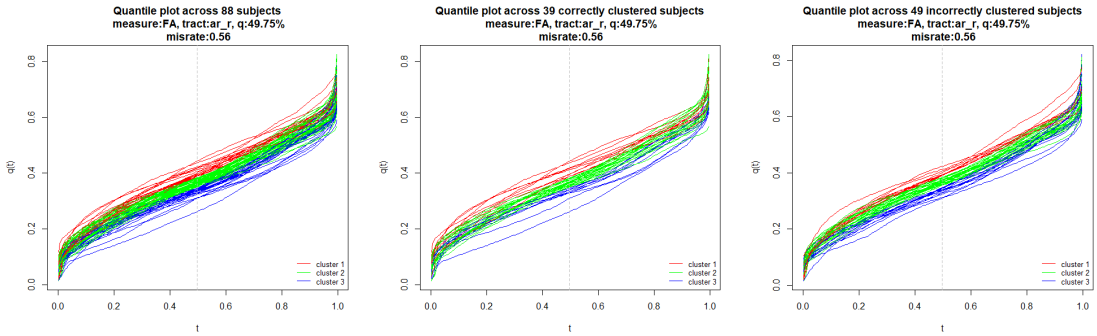
Here, fPCA on the transformed density function shows the natural shift in the mode of density function on the horizontal direction in PC1 and PC2 respectively, which is also reflected in the true density's variation; The 3rd PC reflects the vertical shift in the density function.

5.4 Clustering using quantile

5.4.1 3 groups

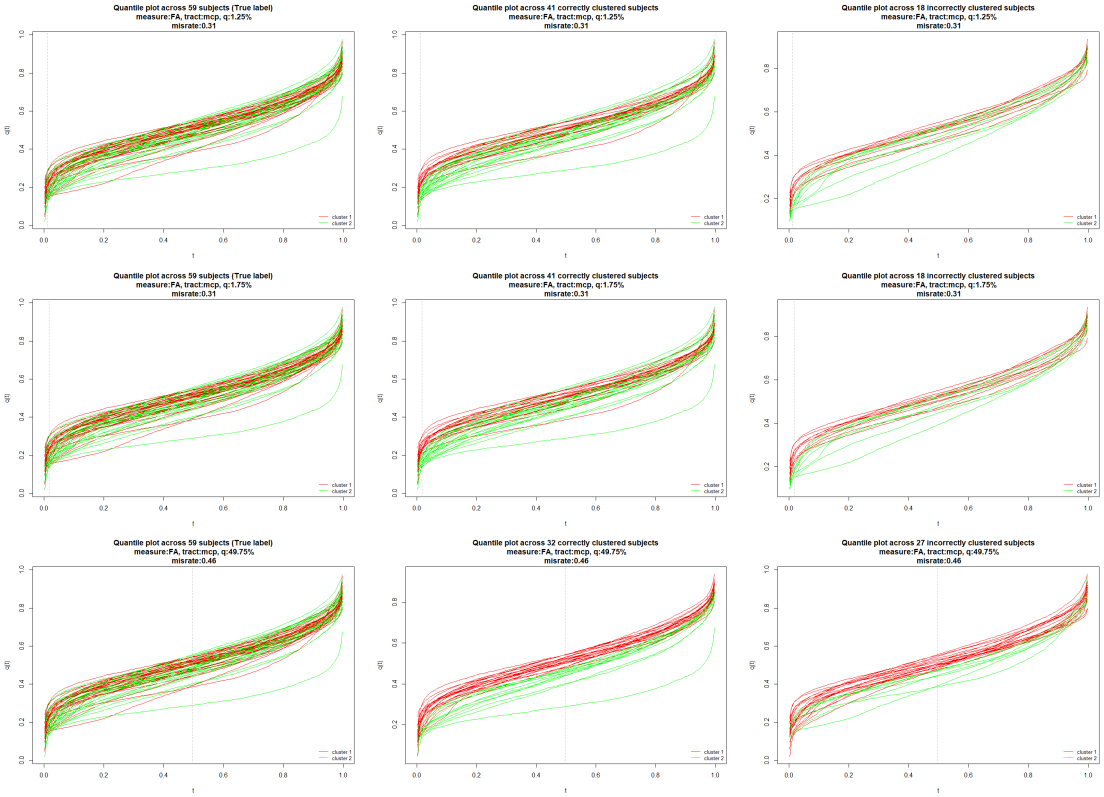
Tract "ar_r" with measure FA gives the smallest misrate 0.48.





5.4.2 2 groups

There is one tract and one measurement with two quantiles which both achieve the smallest misrate 0.31. Tract "mcp", measure FA with quantile 1.25% and 1.75%.



For 3 clusters performance, we can see that 50% quantile can separate the quantile functions for football players better while with higher cluster errors, compared to selected quantiles. Also, selected quantiles which have the best cluster results always are the tails of quantile, that is, either these selected quantiles are very close to 5% or they are close to 99% can help better differentiate the groups.

For 2 clusters performance, we can see that the finding remains the same.

5.5 Classification using Random forest

Here we use random forest to classify the football players into 3 groups or 2 groups. We use 70% data as training data, and use the rest 30% data as validation data. We use ntree=1000, and 40 randomly selected predictors, and minimum misrate in the validation set for each model are given as follows:

Table 4: Minimum prediction misrate for each method

cluster	density	quantile	log quantile transformation
3	0.40	0.48	0.48
2	0.24	0.18	0.29

Since we did not tune parameters for each method separately, the results may not be optimized yet.