STAT602 Final Project

Negassi Tesfay

4/25/2020

Kinematic Features Final Project (STAT 602 2020)

The goal of this project to predict the following thee charactersics based the kinematic features of hand written text:

-What phrase was written.

-Was Cursive or Print used.

-Who was the writer of the sample.

Data Collection:

-40 Writers wrote 6 phrases

-Each phrase were written in Cursive and Print

-This process was repeated 3 times.

A total of 40 x 6 x 2 x 3 = 140 lines were collected.

Data Processing:

-MoVAlyzeR system were used to process kinematic features of the text.

Each line of phrase was broken into segments from 8 to 75 segments.

```
Maximum number of  segments= 75

,   Minimum number of  segments= 8
```

Data Description:

variable	class	first_values
Group	integer	CUR, CUR, CUR, CUR, CUR, CUR
Subject	integer	0, 0, 0, 0, 0, 0
Condition	integer	L1, L1, L1, L1, L1, L1
Trial	integer	1, 1, 1, 1, 1, 1
Segment	integer	1, 2, 3, 4, 5, 6
Direction	integer	1, -1, 1, -1, 1, -1
Duration	double	0.0894, 0.2604, 0.2688, 0.6222, 0.2233, 0.212
VerticalSize	double	0.0826, -0.5291, 0.4285, -0.395, 0.2724, -0.2334
PeakVerticalVelocity	double	1.6658, -3.3324, 2.515, -1.5824, 2.1229, -1.8526
PeakVerticalAcceleration	double	53.6843, -40.1977, 56.8252, -23.3085, 28.77, -33.3139
HorizontalSize	double	-0.0946, -0.0362, 0.1273, 0.1741, 0.2388, 0.0854
StraightnessError	double	0.0604, 0.1225, 0.1451, 0.2099, 0.0533, 0.1007

variable	class	first_values
Slant	double	2.3458, -1.6903, 1.3306, -1.5045, 0.8858, -1.3882
LoopSurface	double	0, 0, 0, 0.001, 0, 4e-04
RelativeInitialSlant	double	-0.0996, -0.9133, -0.7585, 1.3047, -0.3695, -0.4244
RelativeTimeToPeakVerticalVelocity	double	0.494, 0.496, 0.2349, 0.3777, 0.5168, 0.3949
RelativePenDownDuration	double	1, 1, 1, 0.2903, 1, 1
RelativeDurationofPrimary	integer	0, 0, 0, 0, 0, 0
RelativeSizeofPrimary	integer	0, 0, 0, 0, 0, 0
AbsoluteSize	double	0.1256, 0.5303, 0.447, 0.4316, 0.3623, 0.2486
AverageAbsoluteVelocity	double	1.5377, 2.9476, 2.4795, 1.2614, 1.7396, 1.4434
Roadlength	double	0.1375, 0.7674, 0.6664, 0.7848, 0.3884, 0.3061
AbsoluteyJerk	double	580.494, 236.992, 370.418, 239.508, 214.757, 221.835
NormalizedyJerk	double	7.8102, 10.9292, 21.9435, 119.806, 9.8757, 13.0666
AverageNormalizedyJerkPerTrial	double	173.998, 173.998, 173.998, 173.998, 173.998, 173.998
AbsoluteJerk	double	617.42, 338.503, 471.81, 751.2, 301.218, 284.057
NormalizedJerk	double	8.307, 15.6105, 27.9499, 375.762, 13.8517, 16.7316
AverageNormalizedJerkPerTrial	double	293.266, 293.266, 293.266, 293.266, 293.266, 293.266
NumberOfPeakAccelerationPoints	integer	1, 1, 4, 5, 3, 0
AveragePenPressure	double	338.222, 489.308, 629.037, 99.7097, 300.045, 290.773

```
103311  Rows and  30  Cols
```

Summary of the data

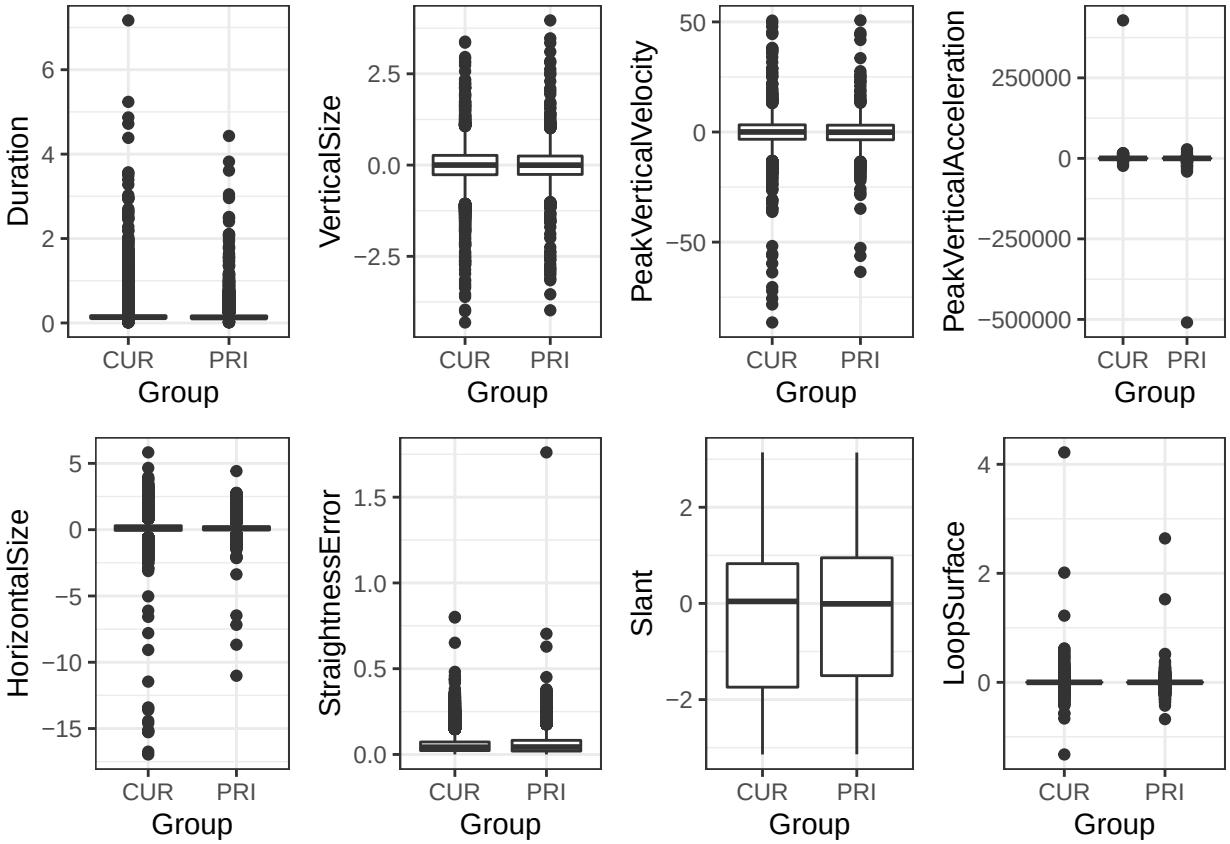
Min, Max, Mean,and Median were used as summary for printig convenience.

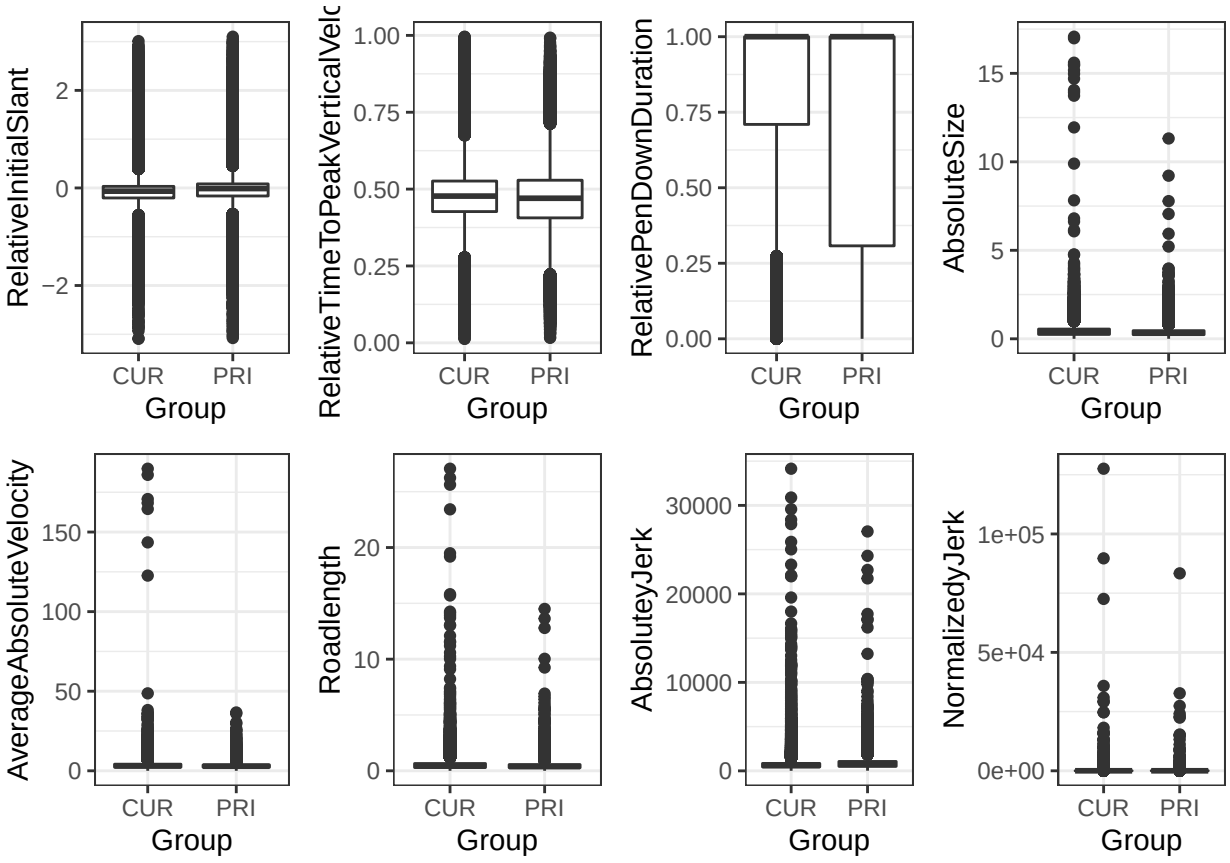
var	max	mean	median	min
AbsoluteJerk	176682.0000	1258.5789054	972.5550	10.5969
AbsoluteSize	17.0697	0.4167459	0.3405	0.0001
AbsoluteyJerk	34146.7000	829.1822324	659.9660	6.7213
AverageAbsoluteVelocity	189.7840	3.3014756	2.9619	0.0000
AverageNormalizedJerkPerTrial	84524.3000	73.9371056	29.1036	11.1791
AverageNormalizedyJerkPerTrial	20909.9000	34.5145549	17.2804	6.8866
AveragePenPressure	1023.0000	381.7036266	399.9000	0.0000
Direction	1.0000	-0.0018294	-1.0000	-1.0000
Duration	7.1690	0.1567955	0.1322	0.0115
HorizontalSize	5.8305	0.1359957	0.1023	-16.9556
LoopSurface	4.2189	0.0007889	0.0000	-1.3231
NormalizedJerk	749816.0000	73.3119985	12.1939	0.0387
NormalizedyJerk	127543.0000	34.2401376	8.8107	0.0042
NumberOfPeakAccelerationPoints	76.0000	1.5626410	1.0000	0.0000
PeakVerticalAcceleration	428441.0000	-9.9065737	3.4832	-509504.0000
PeakVerticalVelocity	50.6937	-0.2259509	-0.0612	-86.4733
RelativeDurationofPrimary	0.0000	0.0000000	0.0000	0.0000
RelativeInitialSlant	3.1002	-0.0355494	-0.0377	-3.0922
RelativePenDownDuration	1.0000	0.7354474	1.0000	0.0000
RelativeSizeofPrimary	0.0000	0.0000000	0.0000	0.0000
RelativeTimeToPeakVerticalVelocity	0.9957	0.4762272	0.4740	0.0131
Roadlength	27.0589	0.5047119	0.4030	0.0000
Slant	3.1416	-0.2584962	0.0178	-3.1410
StraightnessError	1.7612	0.0586022	0.0428	0.0000
VerticalSize	3.9638	-0.0046163	-0.0017	-4.3111

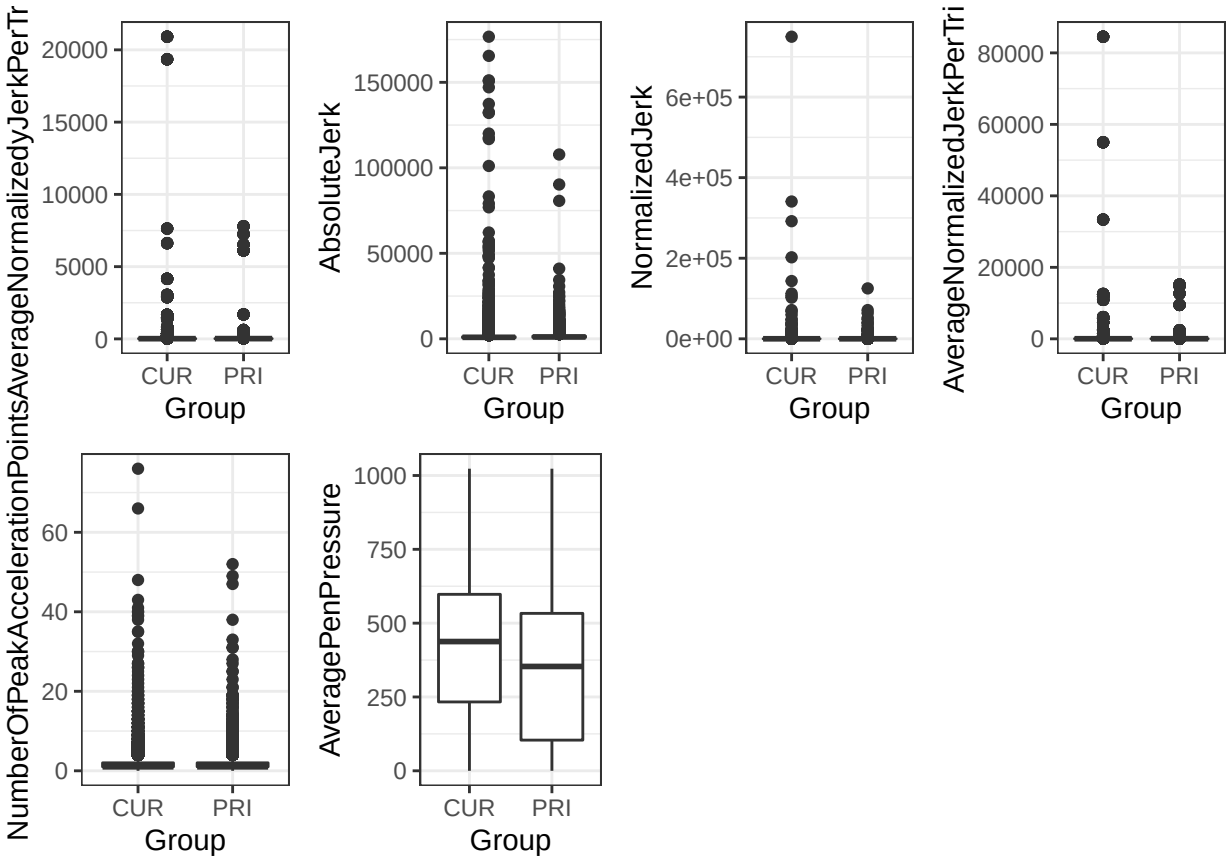
From the summary, we can see that two of the predctors are constant(all zeros). They need to be removed.

Explore Data

Box plot for the kinematic fetures based on Group. Some the variable made clear separation.The same thing can be done for Condition and Subject variables. (Change the varaible G at the beginign of the below chunk)




RelativeInitialSlant
2
0
−2
CUR
PRI
Group
RelativeTimeToPeakVerticalVelc
1.00
0.75
0.50
0.25
0.00
CUR
PRI
Group
RelativePenDownDuration
1.00
0.75
0.50
0.25
0.00
CUR
PRI
Group
AbsoluteSize
15
10
5
0
CUR
PRI
Group
AverageAbsoluteVelocity
150
100
50
0
CUR
PRI
Group
Roadlength
20
10
0
CUR
PRI
Group
AbsoluteyJerk
30000
20000
10000
0
CUR
PRI
Group
NormalizedyJerk
1e+05
5e+04
0e+00
CUR
PRI
Group


NumberOfPeakAccelerationPointsAverageNormalizedJerkPerTr
20000
15000
10000
5000
0
CUR
PRI
Group
AbsoluteJerk
150000
100000
50000
0
CUR
PRI
Group
NormalizedJerk
6e+05
4e+05
2e+05
0e+00
CUR
PRI
Group
AverageNormalizedJerkPerTri
80000
60000
40000
20000
0
CUR
PRI
Group
NumberOfPeakAccelerationPointsAverageNormalizedJerkPerTr
60
40
20
0
CUR
PRI
Group
AveragePenPressure
1000
750
500
250
0
CUR
PRI
Group

Prepare data for modeling

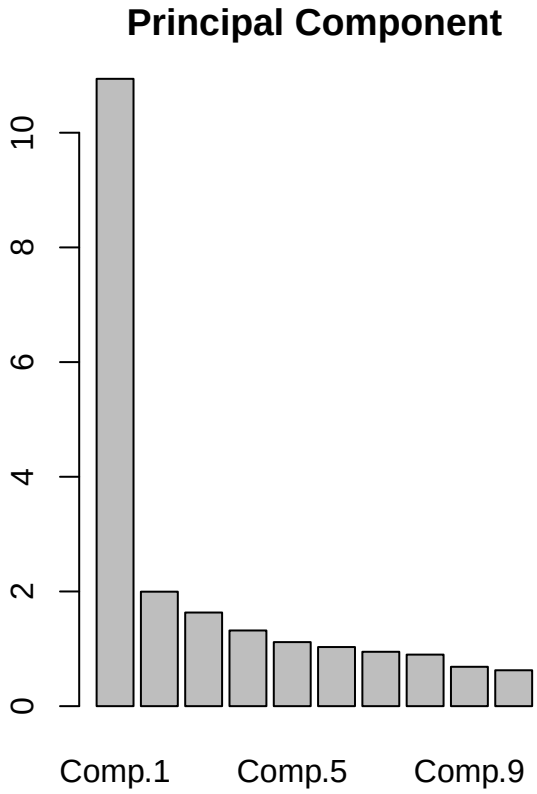
-Average the kinematic features per line.(Take a mean of all the segment-features in line). And also separate the Class variables and Feature varaibles. And we will set aside the third trial for testing purpose

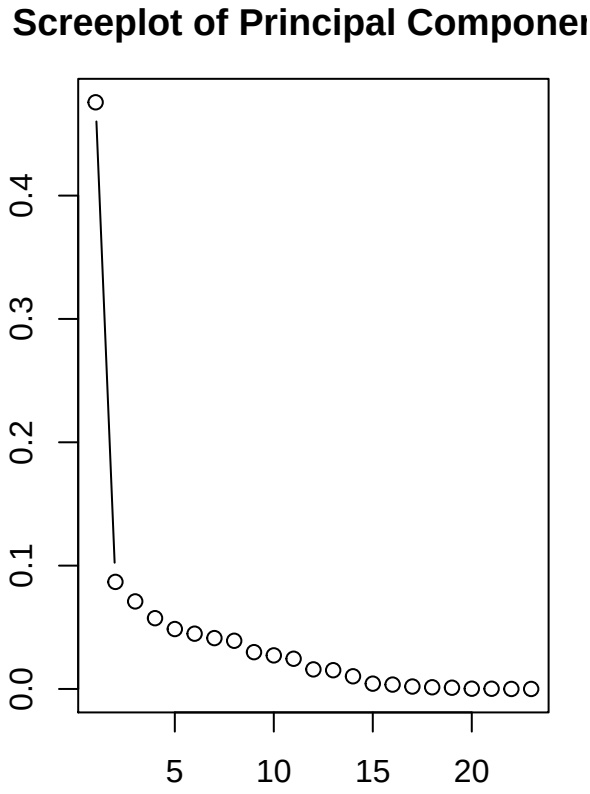
Principal component Analysis:

Principal component anlysis was used to reduce the dimension of the features and remove possible correlations betweeen predictors. and about more than 92% of variablity was captured by reducing to 10 components.

Table 3: The first 11 components captured about 92% of variance

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
0.47565	0.5624513	0.6334348	0.6908212	0.7393965	0.7842168	0.8254416	0.8645162	0.8943593	0.9215819

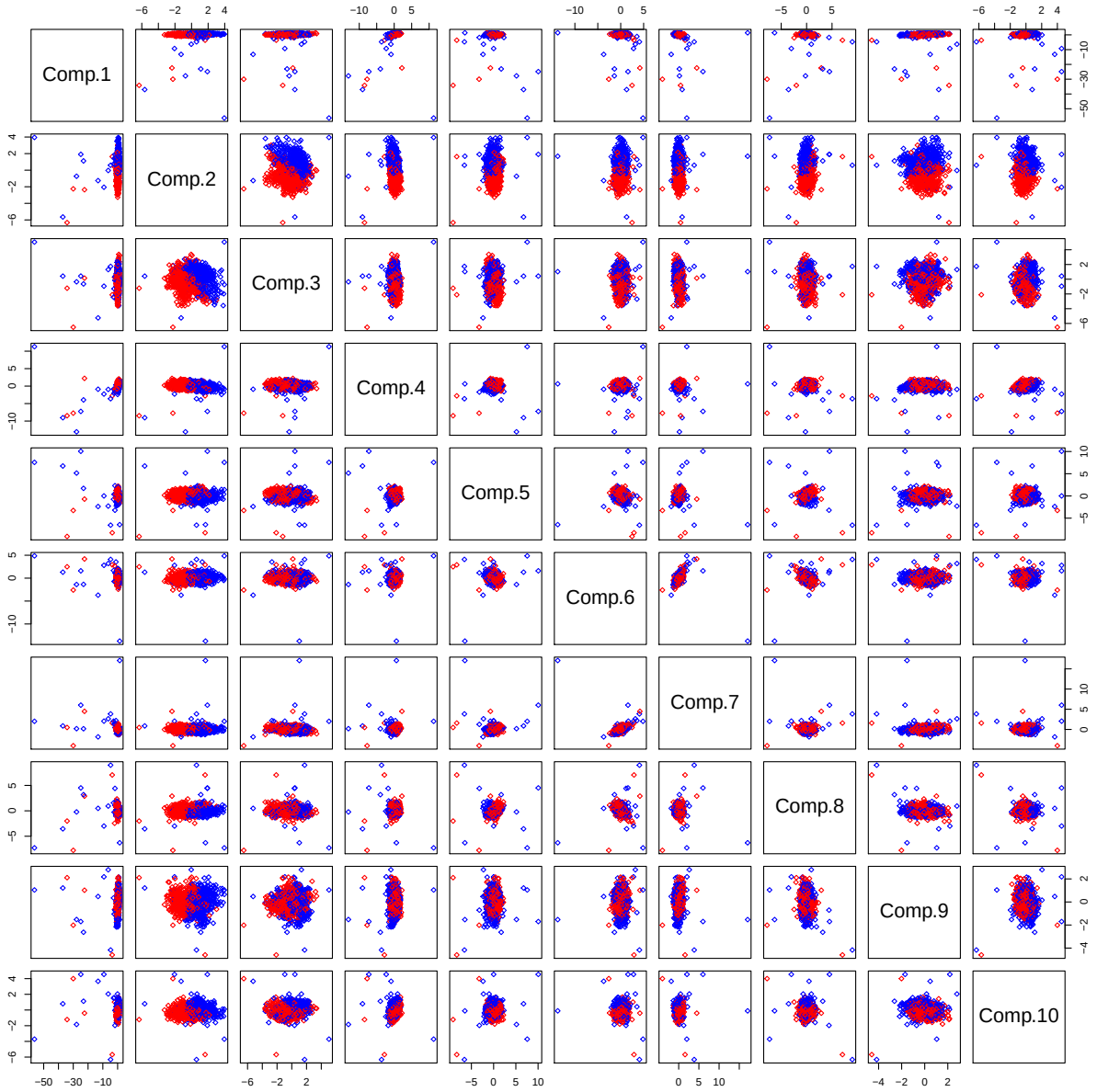




The components can be seen graphically

Explore our data afer transformation:

A matrix of scatterplots between each pair of predictors after transformation, random sample 0f 70% was use to improve over plotting.



Model building:

To build the a model for classificaiton of the handwriting, Linear Discriminant Analysis and K-nearst means models were used. Among all the classifying models, like logistic regression , quadratic regression , LDA and Knn were used for the following reasons.Logistic regression is handy only for binary classification. thats only when the the response variable is has two classes. since this classifcation has resposnse that has more than 2 classes, it is removed from the selection. QDa is good when we have enough observations that outnumber the parameters it uses. for LDA the number of paramenters is always $q - 1$, where q is the number of predictors. but for QDA, the number of parameters $k(q)(q + 1)/2$ and k is the number of classes. The number $k(q)(q+1)/2$ came from the fact that varianc-covariance marix of of each class of the predictors are not equal. For LDA we take only the diagonal values in the matrix. but for QDA we have to take all the upper and lower triangle values and each multipied by the classes becacuse QDA takes separate matrix for each class. in our case q=23, and k=(2,6,40,480).

For group = 2x23x24/2 = 552 parameters and 1440/2=720 observations

For Condition = 6x23x24/2=1656 parameters , 1440/6 = 240 observations

And for subject and JOint, parameter increase and obsevation decrease

Omitting Logistic regression and QDA from the choice, for the above reason, I will use LDA and QDA

Linear Discrimininant Analysis (LDA)

LDA algorithm is based on Bayes theorem and classification of an observation is done in following two steps.

-Identify the distribution of each of each class in in the input variable.

-Flip the distribution using Bayes theorem.

In LDA algorithm, the distribution is assumed to be Gaussian and exact distribution is plotted by calculating the mean and variance from the historical data.

datascienceplus.com/how-to-perform-logistic-regression-lda-qda-in-r/

-Three dataset will be used.

-Full data for prdiction

-Train data for Accuracy test

-Test data for Accuracy test

```
Group Error= 10.85 %
```

```
Condition Error= 65.82 %
```

```
Subject Error= 61.43 %
```

For joint Accuracy,the number of observation are not flexible. If sample random for train and test, the chance to enclude all the levels in a dataset is rare. Instead, Trial 1, and 2 were selected for train, and the third trial for test.

```
Joint Error= 86.46 %
```

Prediction of unlabeled data.

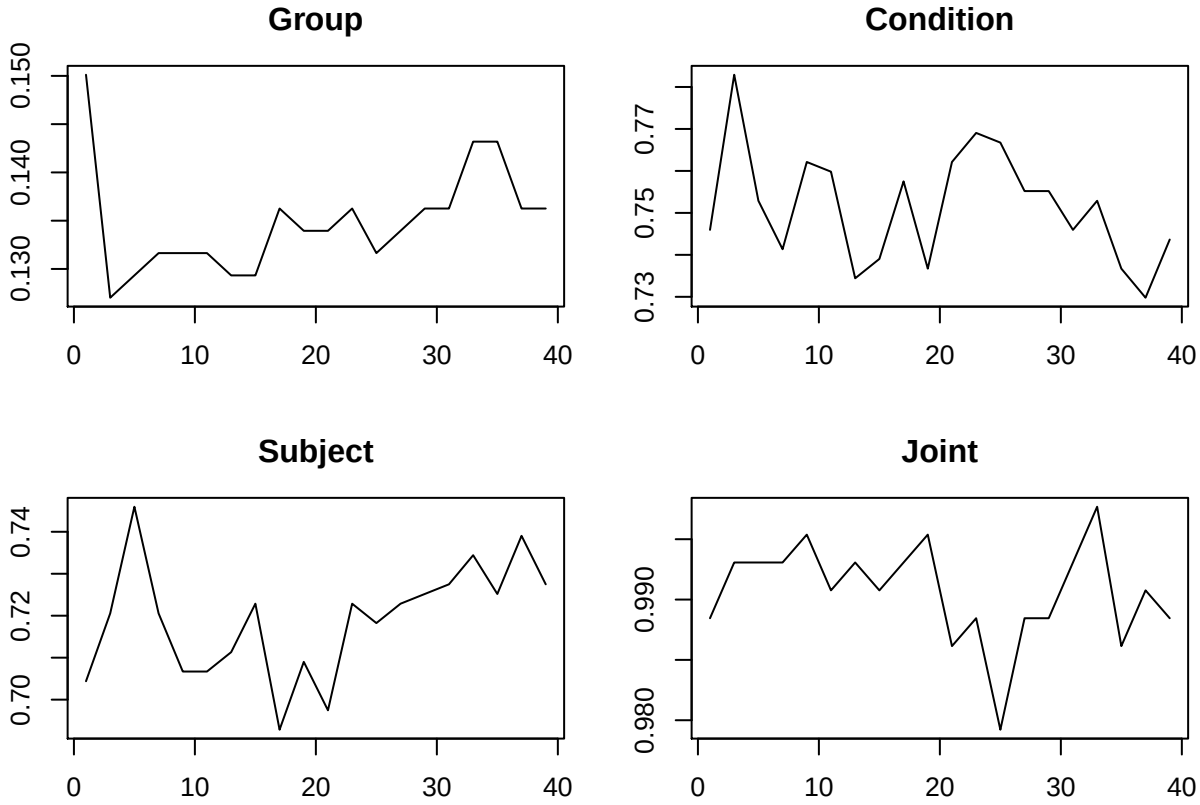
To predict the data, we have to transform our unlabeled data to the same structure as the labeled data. that is, same number of predictors, the same dimention reduction and use the the model for predction.

Joint	Group	Subject	Condition
CUR:00Q:L1	CUR	0	L1
CUR:0:L2	CUR	0	L4

K nearest neighbours. Knn

Unlike LDA which assumes that the classes are on gausian distribution, Knn is a non-parametric method which uses an algorithm of classifying an observation to the majority in its neighbours. the number of neighbors are represented by k. the choice of k is where it gives the minimum error. I will iterate 100 odd numbers of neighbours to find the miminmum error. Odd was choosen to break ties in case it happens.

Selection of K



For Joint due to lack of obsevations, we have only three observations, we will use k=1 and we wll use the trial 1 and 2 as training and trial as 3.

Table 5: Errors for different K

Class	k	err_grp	err_con	err_subj	err_joint
Group	3	0.1270208	0.7829099	0.7205543	0.9930716
Condition	37	0.1362587	0.7297921	0.7390300	0.9907621
Subject	17	0.1362587	0.7575058	0.6928406	0.9930716
Joint	25	0.1316397	0.7667436	0.7182448	0.9792148

```
Condition Error= 83.54 %
```

Prediction

The following prediction was made from the model.

Joint	Group	Subject	Condition
CUR:00X:L1	CUR	7	L1
CUR:0:L1	CUR	0	L4