CSS 490

Pattern Classification of Cars by Region

Iteration 3

Danny Ly
Kevin Ulrich
Robert Stewart

Revision History

Iteration 0 — April 5, 2017

- Developed project proposal, based around the MOBA Dota.
- Descriptions of the research topic, the problem, and the area of application
- Began looking into getting data, specifically via Valve's API
- Created measurement of success

Iteration 1 — April 28, 2017

- Changed project topic from Dota to automobile classification
- Rewrote and extended research topic, area of application, and measurement of success, focusing on automobiles
- Decided on a different method of data acquisition: scraping automobile characteristic data from ultimatespecs.com
- Wrote technical approach, with data control, data collection, assumptions, and noise sources being described within.

Iteration 2 — May 12, 2017

- Decided on features and categories that the automobiles should have and represent, respectively.
- Decided on and created initial plots: histograms, 2D scatter plots, and probability density plots.
- Created analysis for all initial plots.
- Discussed normality, central limit theorem and its potential application to our dataset, and the independence of the data.

Iteration 3 — May 26, 2017

- Performed PCA on the data, using SVD.
- Created plots of 3D original feature space, loading vectors, scree plots, correlation matrix, covariance matrix, 3D PC scatter plots.
- Using PCA eliminated highly correlated, low importance features to reduce dimensionality.
- Performance analysis using the resulting scree plots and loading vectors.
- Plotted our resulting chosen features in the new 3D space.

Table of Contents

Project Proposal

Research Topic

Team Dota is researching patterns in vehicle features and how they are associated with car brands. Many car brands have developed patterns in having specific features in their vehicles, for example, Ferrari's always limits their top speed to 200 mph.

We want to know given a specific vector of features, which category (car brand, e.g. BMW, Honda, etc.) does it belong to.

By just looking at a vehicle, it might be difficult to know at a glance what brand it is. But using pattern recognition, a computer might be able to tell what brand a car is given some basic features.

Given a set of features that define a car, what brand does the car belong to. We have decided to have three classes: Volkswagen, Ford, and Toyota. The features we will be using to identify a car are length, width, height, max horsepower, fuel tank capacity, curb weight, and top speed.

Vehicles

In this proposal, we will use the term car interchangeably with vehicle or automobile. When we say car, vehicle, or automobile, we mean anything with four wheels driven by a human, including trucks, vans, hatchbacks, etc.

For this project our classification categories will be Volkswagen, Ford, and Toyota. We choose these automobile brands because of their vast similarities in features and particular differences.

Area of Application

Some establishments that could benefit from this pattern recognition includes car companies, the police, and the Department of Motor Vehicles (DMV). Car purchasers will also be able to use our model to assist in making a purchase.

Car companies can use our model when designing a vehicle. When designing a car, companies may want to see if their design is categorized as another company's brand. Companies would want to do this to either avoid producing vehicles similar to another company's brand or to design a vehicle similar to a brand that is popular. Companies may also want to insure that their design conforms to their brand's pattern and will be able to use our model to classify a car before they build it and see if it fits their current brand recognition. Or perhaps if they want to break out of the mold of their previous models, they could make sure that a new model does NOT classify as their brand.

The police, when trying to find criminals' vehicles, could potentially figure out the brand of a car they are trying to find based on some of the features that we will classify based on.

Frankenstein automobiles, cars that are created from used car parts mixed with various panels to replicate high-end sports car are being put on the market and sold to individuals who cannot tell the difference between real and fake models, at the same high price as authentic models. Using the pattern classifiers we develop, a user will have the ability to distinguish whether they are getting into buying something of authenticity or a replicated model.

The DMV could classify a vehicle that they are licensing to make sure it is the brand that it looks to be, and is not a counterfeit.

Vehicle purchasers can use our model when choosing a car to purchase. When choosing a vehicle, purchasers often want to buy a vehicle with specific features. Given a purchaser's specified set of desired features, our model will categorize the set of features to a specific car brand. The purchaser will then have a better idea of what car brands to look for in their purchase.

Data Acquisition

We will gather data by scraping it from ultimatespecs.com, a domain of all car brands and their sub generations of models for each model list their respective manufacture specification(s).

Through the process of scraping from this website, we will grab all data pertaining not only to a particular car model, but as an entire entity of that car brand. Data we will be looking at as discussed above: length, width, height, max horsepower, fuel tank capacity, curb weight, and top speed.

Measurement of Success

The measurement of success of this project will be based on the classification of a car brand into the three brands that we are observing, given a feature vector.

Technical Approach

Detailed Process

We are studying how cars relate to manufacturers. So we want to find out given a set of features (a feature vector), and if we can find a pattern of what feature sets map to what manufacturers.

We are collecting data for each car model straight from the manufacturer listing themselves. There will be no experimentation necessary to get our data. Therefore, there are no unstable elements to this project as we see it. Because all data we've collected through ultimatespecs.com are from manufacturer specifications for all models there are no unstable elements.

We will have to get rid of some data (specific cars) that lack one or more of the features we need to form a feature vector, as an incomplete feature vector is not classifiable.

Questions

Team Dota is attempting the answer the following questions:

1. Given parameters that describe a certain model of a car that is either Ford, Toyota, or Volkswagen, without knowing which brand it is we want to see if we can figure out which of the three brands it belongs to.
2. Given parameters that describe a certain model of car that does NOT belong to Ford, Toyota or Volkswagen, which of the three brands is it most similar to?

Data

Features that we are interested in (variables): Max horsepower, top speed, width, height, length, curb weight, and fuel tank capacity return the automobile brand it belongs too.

Most observations of cars have the variables that we are interested in. Cars that do not have all of the variables we are interested in will not be included in our process. The variables we are interested in vary depending on the type of car, meaning that there are potential patterns in the variables for specific car brands, as car brands often are associated with specific types of cars.

Data Control

For our process, there is no random variables. Data for the variables we are interested in is provided by the vehicle manufacturers themselves, meaning that the data is consistent across all models. This means that we can fully control all variables.

Constant variables: none

Dependent variables: none (feature variable top speed is possibly dependent on other feature variables such as weight and horsepower.)

Data Collection Process

Through the process of scraping from this website, we will grab all data pertaining not only to a particular car model, but as an entire entity of that car brand. Data we will be looking at as discussed above: length, width, height, max horsepower, fuel tank capacity, curb weight, and top speed.

Assumptions

The assumption that the variables we are interested in are constant is due to the vehicle manufactures providing the data themselves. The assumption is made that the car manufacturers have provided accurate data on their car models.

Noise Sources

A car's year may be a potential source of noise. Due to trends, changing of standards of vehicles, and technological innovation, car companies may change patterns in the car models they produce over time. This would cause unwanted complexity and increase variability of features within a car brand.

The type of vehicle that we are observing can also produce noise. The differences in our variables between a van and a sports car, for instance, can be fairly large. This might mean that we will find stratification in our data within our classes, and also across classes.

Technical Content

Features and Categories

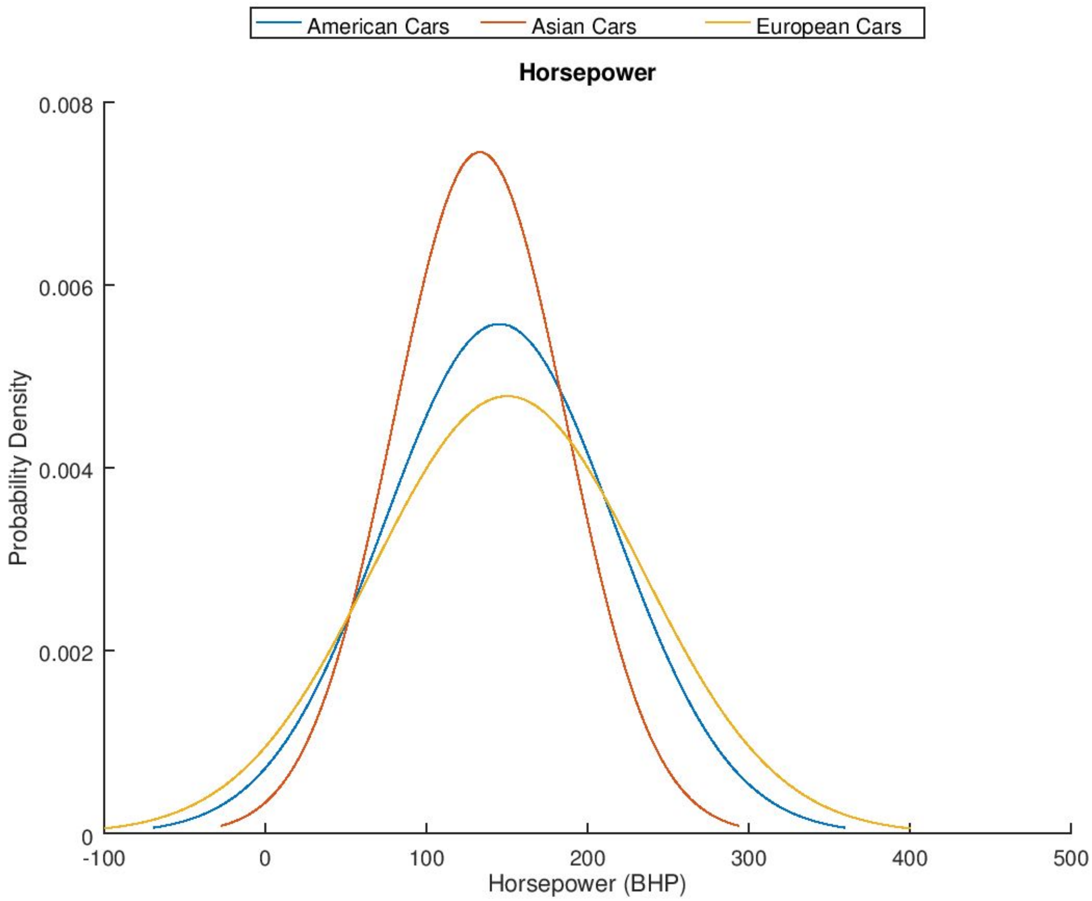
From the car observation data we have collected, we decided to limit the features used to maximum horsepower, maximum torque, fuel tank capacity, top speed, length, width, height, and curb weight. Every car observation has these features and they will help classify car observations into one of the three categories: American car, Asian cars, and European cars.

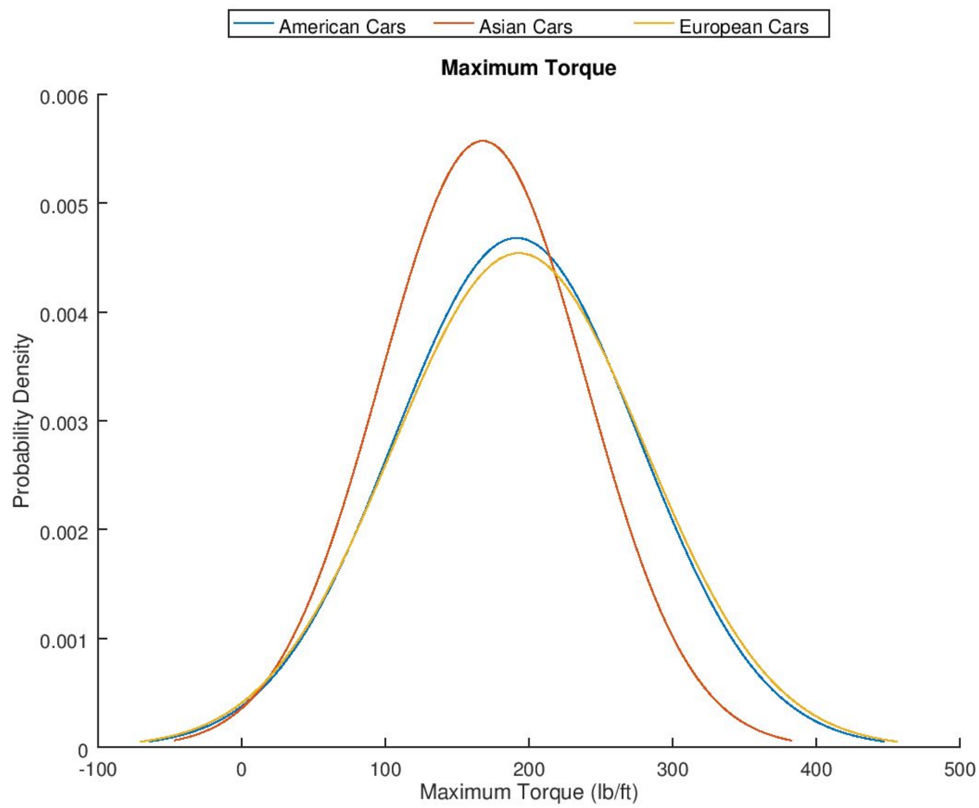
Determination of Plots

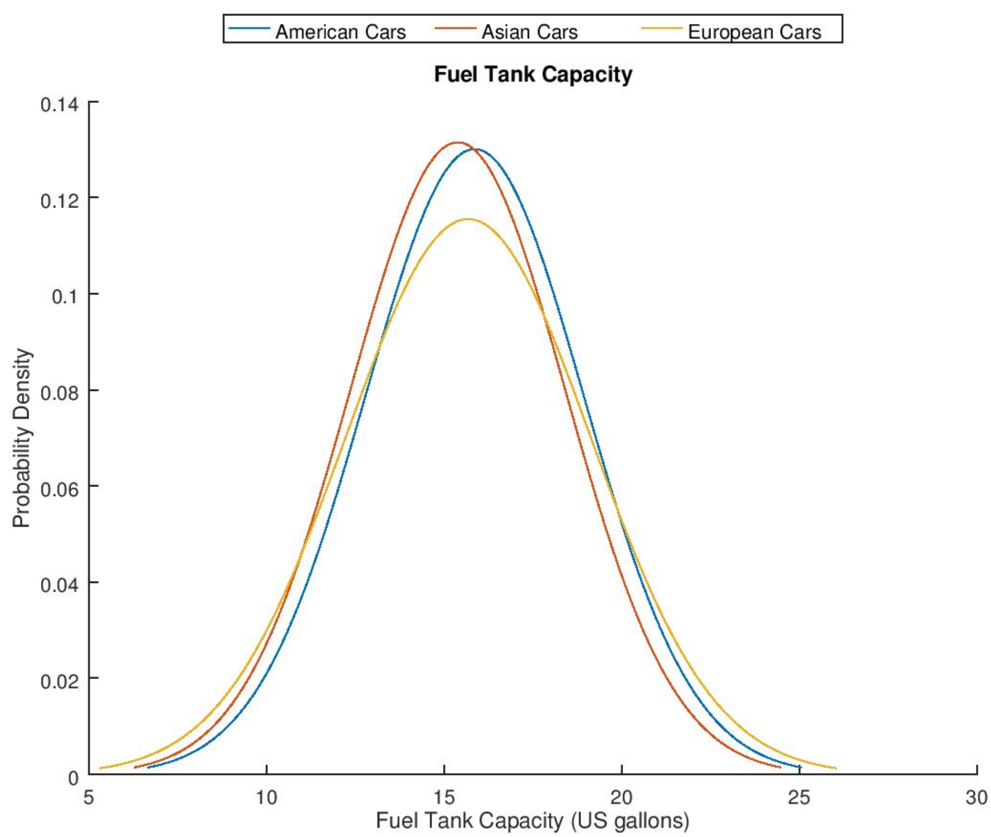
We decided that a 2D scatter plot summary, a probability density plot summary, and a histogram summary would be all be useful in analyzing our collected data.

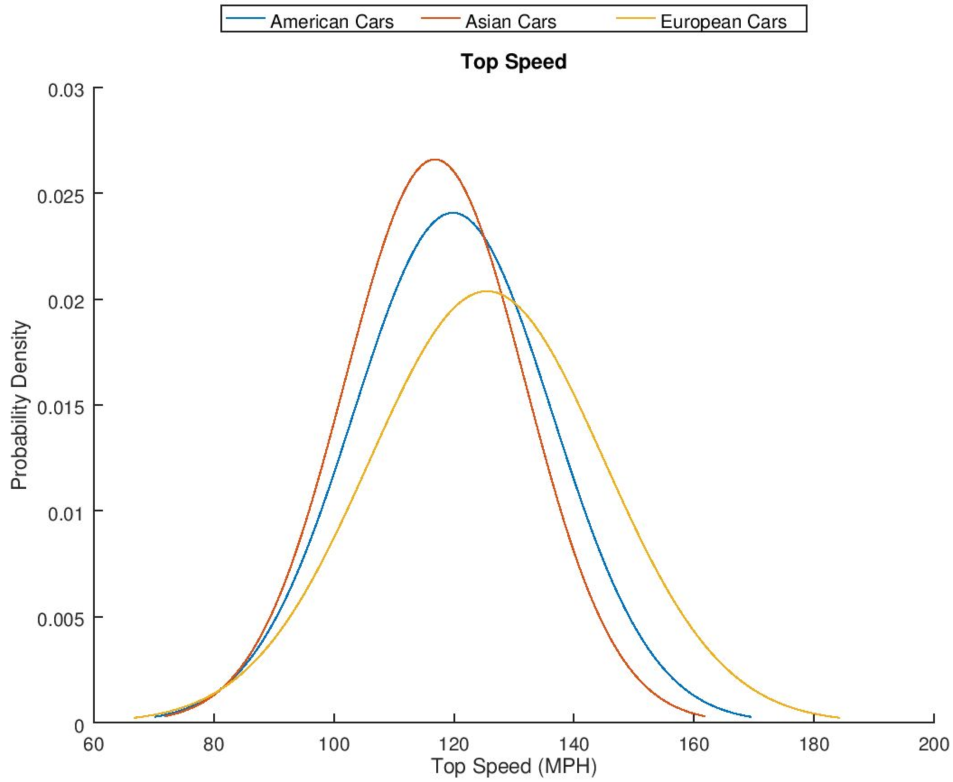
- We decided a probability density plot summary would be useful for informing the decision boundaries between the three different classes (regions). It's also a good way to tell, when combined with a dot diagram, whether or not normality can be assumed for a dataset.
- We decided a histogram summary would useful for analyzing the spread of features across intervals as histograms capture frequency of observations. With the frequency of observations captured, features frequency can be closely analyzed in the bins that they are grouped in. For example, using a histogram would help in determining the normality of a feature as a normalized histogram should be in a bell shaped curve. Comparing the histograms between features will also help visualize patterns. For example, from the histograms that capture the frequency of car height, width, and length for american cars, it is clear that there are two groupings of sizes. This makes sense as car sizes can generally be grouped into larger SUV's/trucks and smaller compact cars.
- A 2D scatter plot summary is useful to be able to see correlation among the different classes. This will be useful for PCA, where we are interested in looking at inconsequential or highly correlated features, the latter being evident in scatter plots.

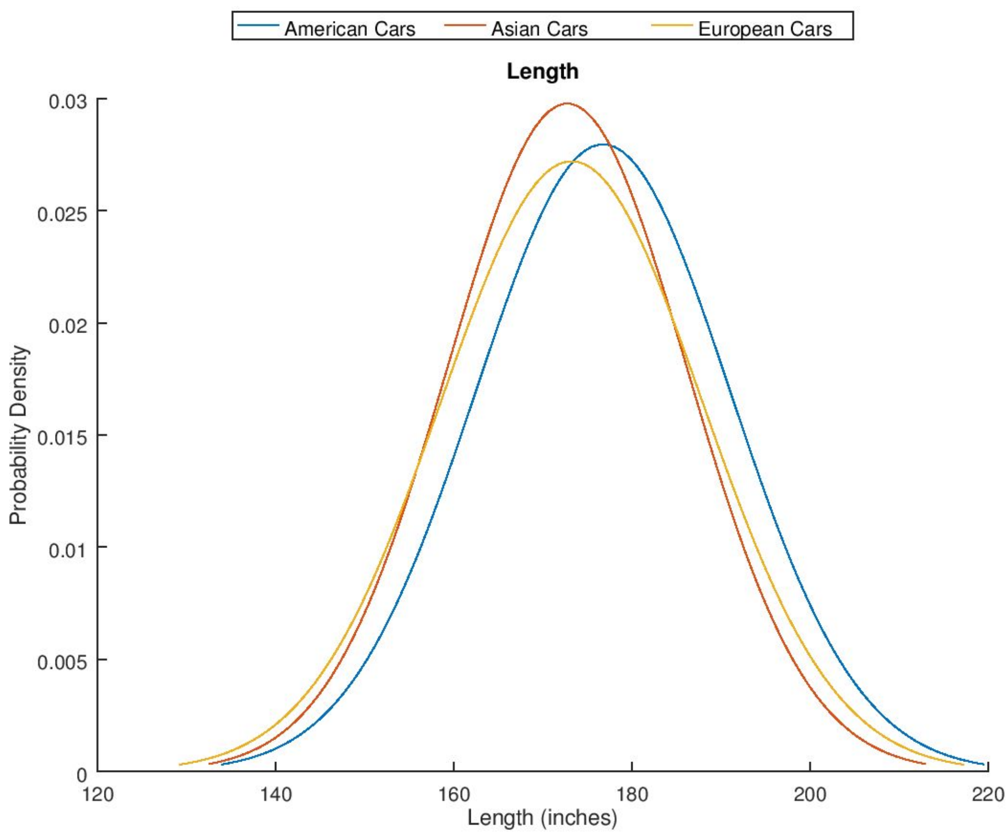
Probability Density Plots

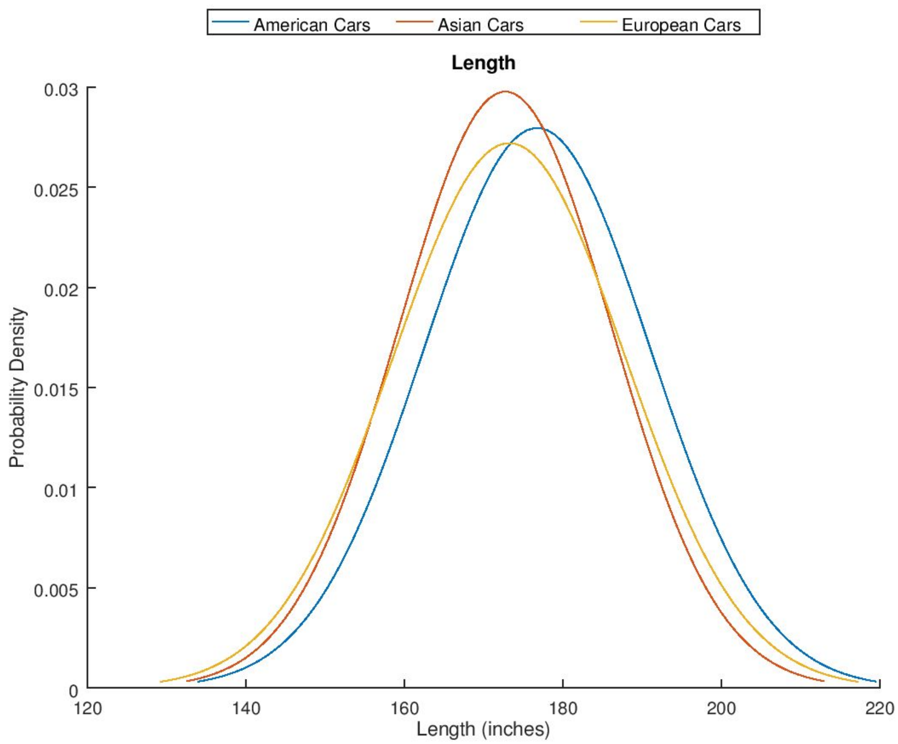


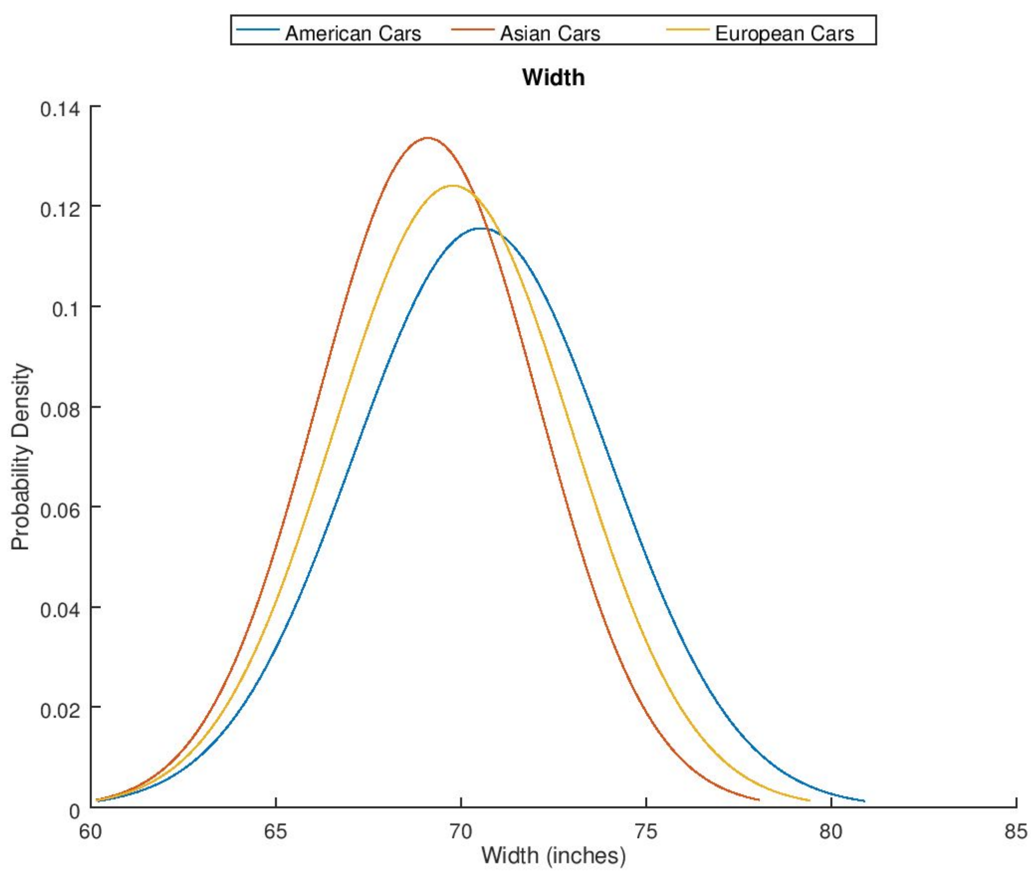

American Cars
Asian Cars
European Cars
Maximum Torque
0.006
0.005
0.004
0.003
0.002
0.001
0
Probability Density
-100
0
100
200
300
400
500
Maximum Torque (lb/ft)

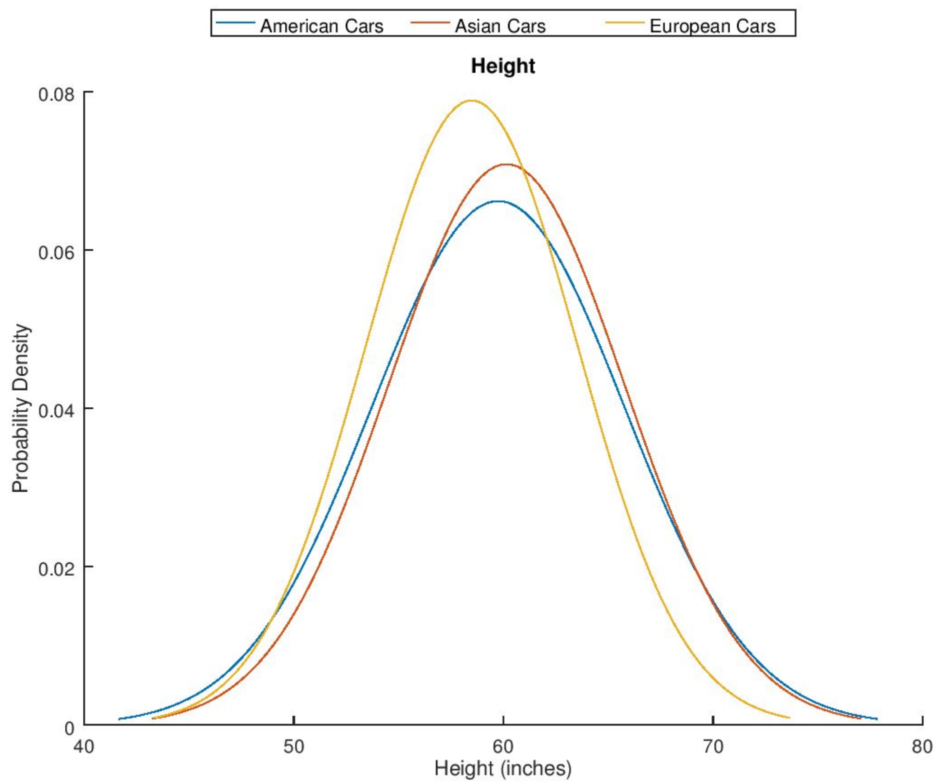

American Cars
Asian Cars
European Cars
Fuel Tank Capacity
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Probability Density
5
10
15
20
25
30
Fuel Tank Capacity (US gallons)

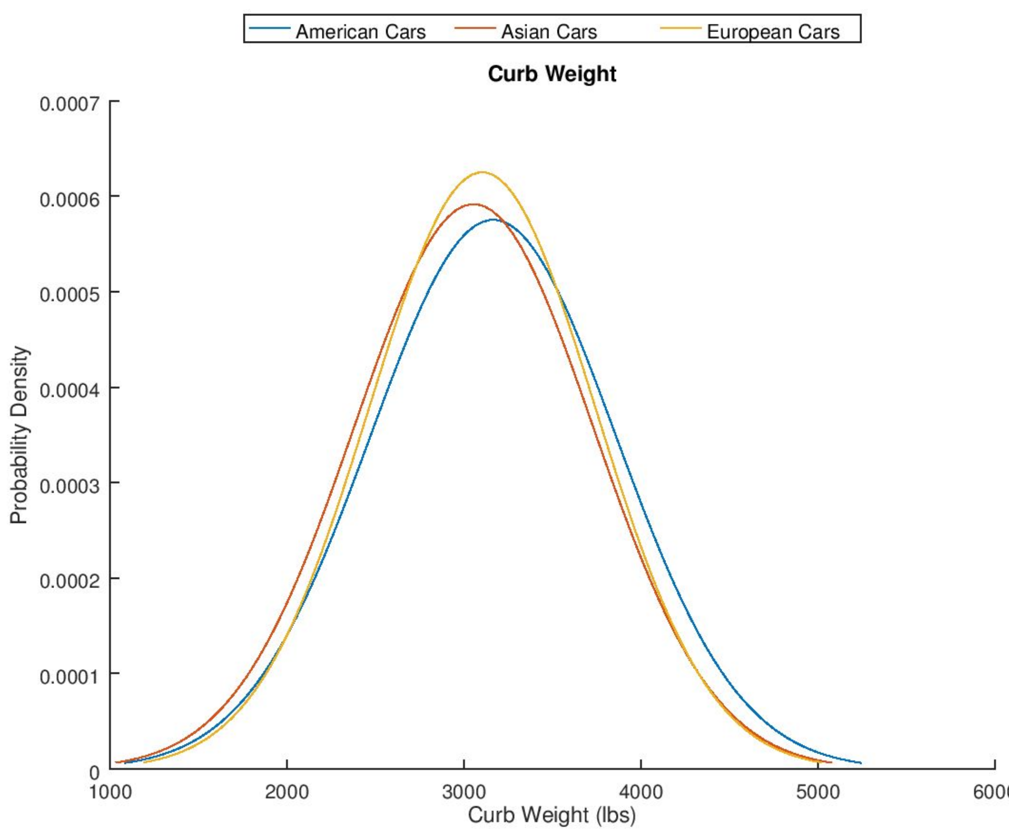

American Cars
Asian Cars
European Cars
Top Speed
Probability Density
0.03
0.025
0.02
0.015
0.01
0.005
0
60
80
100
120
140
160
180
200
Top Speed (MPH)


American Cars
Asian Cars
European Cars
Length
Probability Density
0.03
0.025
0.02
0.015
0.01
0.005
0
120
140
160
180
200
220
Length (inches)


American Cars
Asian Cars
European Cars
Length
Probability Density
0.03
0.025
0.02
0.015
0.01
0.005
0
120
140
160
180
200
220
Length (inches)

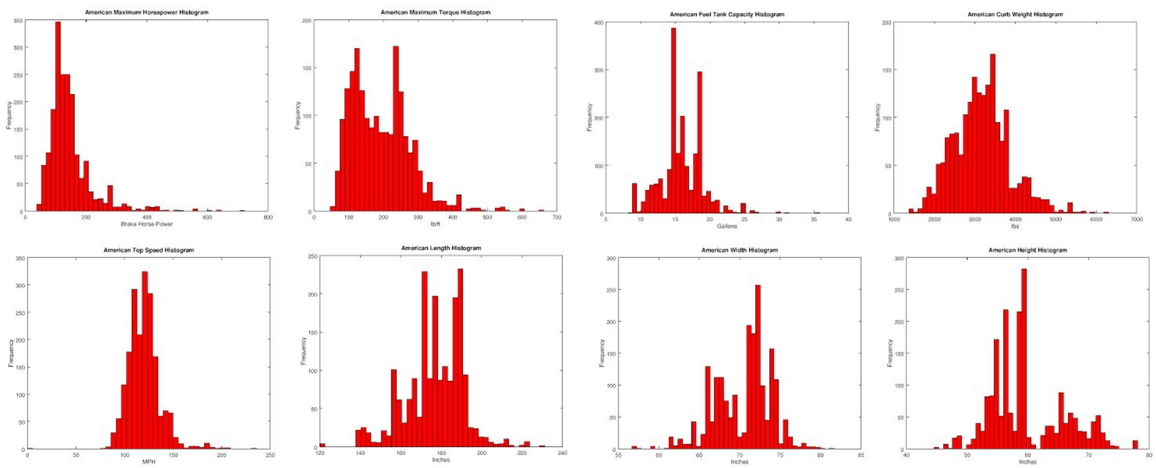

American Cars
Asian Cars
European Cars
Width
Probability Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
60
65
70
75
80
85
Width (inches)


American Cars
Asian Cars
European Cars
Height
Probability Density
0.08
0.06
0.04
0.02
0
40
50
60
70
80
Height (inches)

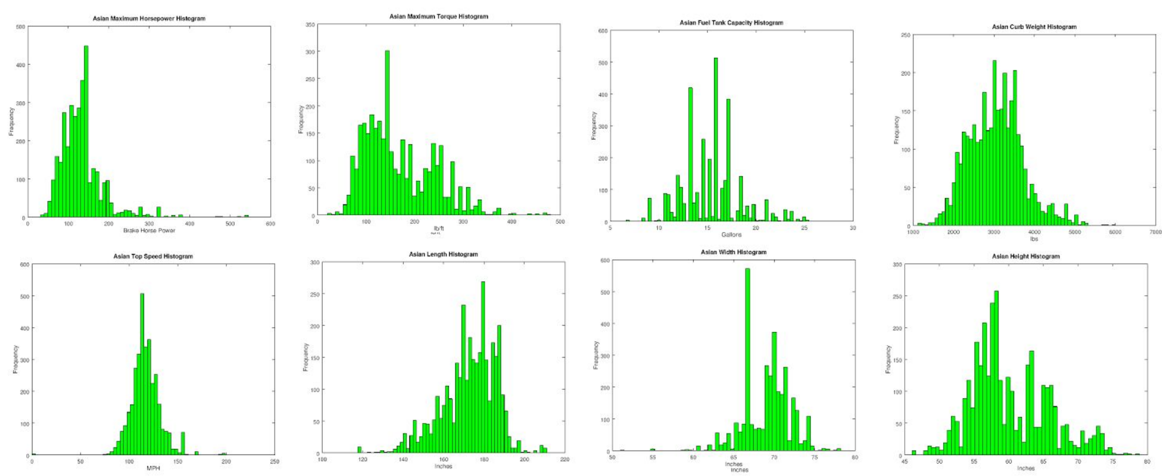

American Cars
Asian Cars
European Cars
Curb Weight
Probability Density
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0
1000
2000
3000
4000
5000
6000
Curb Weight (lbs)

Histograms

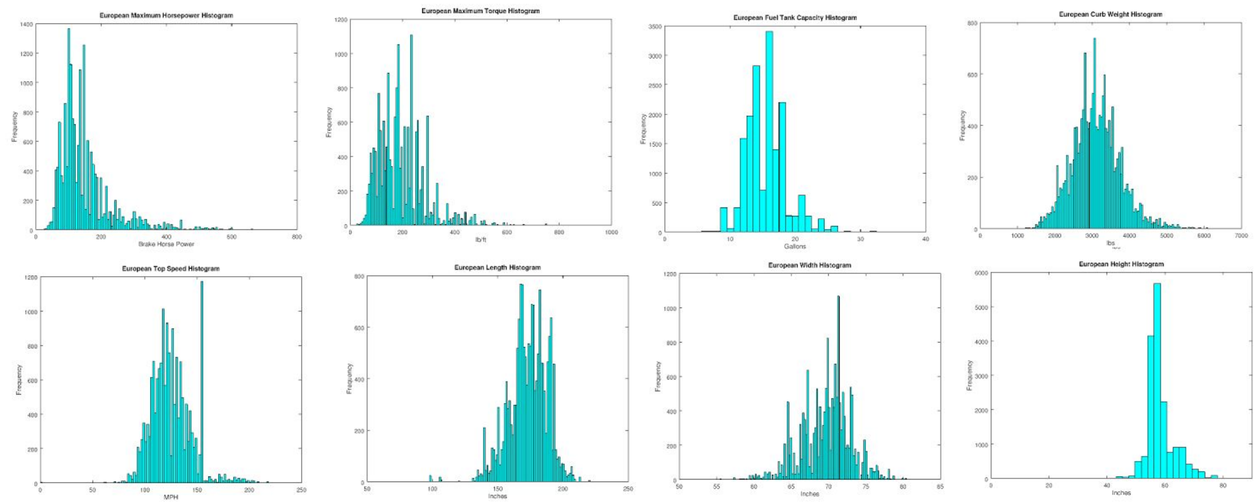
American Feature Histograms



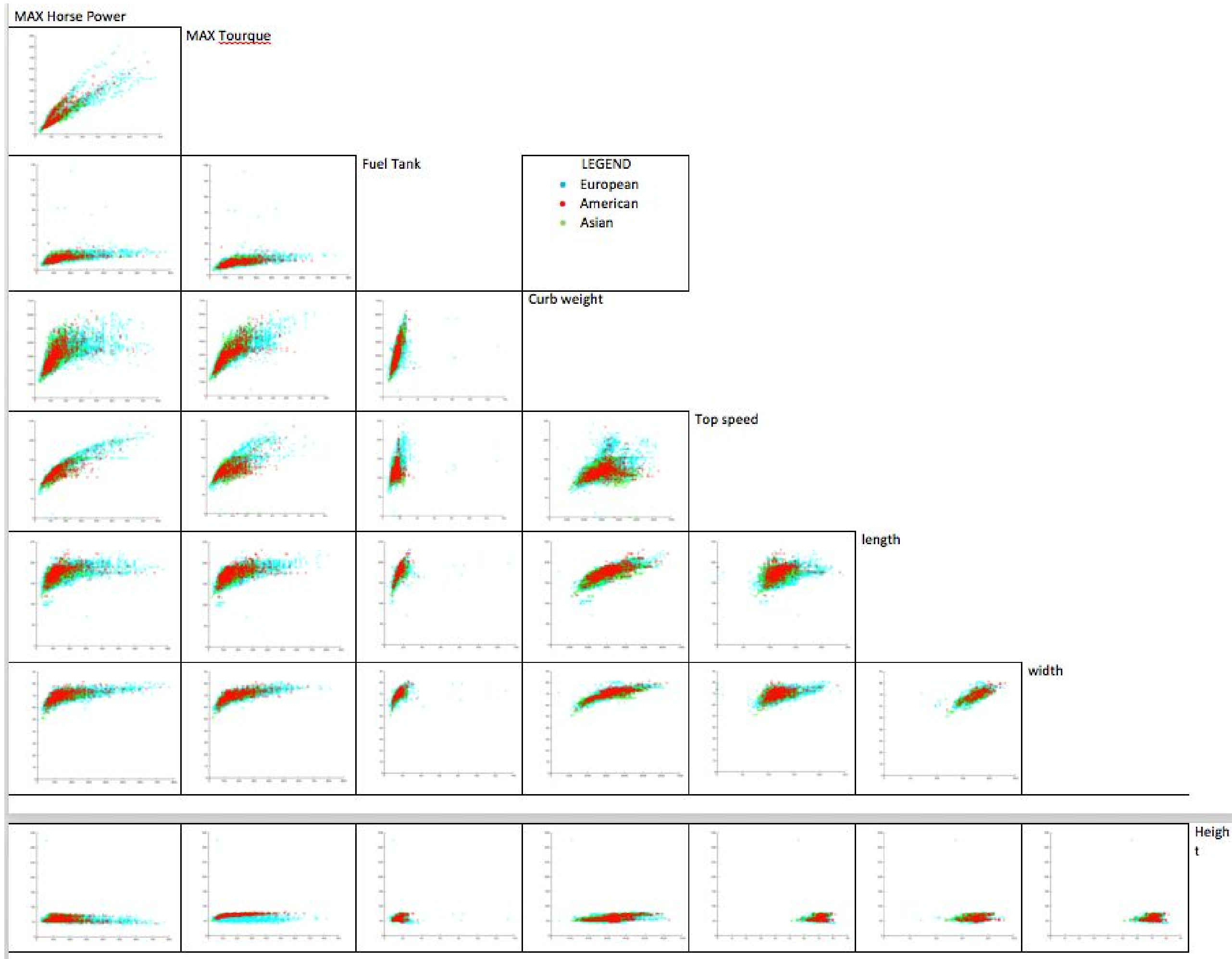
Asian Feature Histograms



Asian Feature Histograms



Scatter Plots



Scatter Plot Matrix Analysis

- Horsepower to Torque
    - Looking at this data, it is clear that for any car in any region that horsepower has a strong correlation to torque of a vehicle.
- Max Horsepower to Fuel Tank Capacity
    - Looking at the data, at first glance we would think fuel tank has some correlation to max horsepower, however taking a deeper interpretation, if a fuel tank's weight, meaning if a fuel tank capacity is smaller, that should positively impact the abilities of a vehicle's max power or max horsepower allowing it to go faster because of the reduce in fuel weight.
- Max Horsepower to Curb Weight

    - To interpret this data, for our data we can see that there has been a medium correlation to max horsepower, if we think about fuel tank capacity, we can see that the data on the scatter plot for the max horsepower has stronger than the correlation between the fuel tank capacity, that is because curb weight is the total weight of the vehicle, whereas the fuel tank capacity is only accounting for the weight of the fuel tank.
- Horsepower to Top Speed
    - For this interpretation, our initial hypothesis was that horsepower does not affect a vehicle's ability of its top speed, but interpreting more of this data we can see that the data does tell us if our vehicle were to be heavier, without its correct max horsepower the vehicle would not reach its top speed. Therefore when looking at our scatter plot we can see a high correlation between horsepower and top speed, because in the event that a vehicle does have a high top speed, that does mean it has a higher top speed because more horsepower means a better ability to attain higher speeds.
- Horsepower and Length
    - Initially, our hypothesis was that there was no way that a vehicle's length could affect its horsepower or have correlation, looking at our data we can see our hypothesis is seldomly correct, and does prove our initial hypothesis, in order to understand the data, we need to understand that a longer vehicle, will in fact have more material to its mass, which is where a lot of a vehicle's weight is contributed too. Therefore we we can see the minor correlation in our scatter plot.
- Horsepower to Width
    - Rationally understanding a vehicle and how width our affect its horsepower, we do not see much correlation, the only rational reason we see this deviation from our initial idea of these two continuous variables being minimally correlated is, the aerodynamics of a vehicle, when a vehicle is traveling the width of its body can negatively affect its ability to perform(horsepower).
- Horsepower to Height
    - For our initial hypothesis we claimed the height of a vehicle will have a strong correlation to a vehicle's max horsepower, because the taller a vehicle is the less it is to its center of gravity, thus not allowing that vehicle to fully perform. However for this scatter plot data, we can see for all regions this to be untrue. There is no correlation between horsepower and height, to reflect on our initial though, we realize horsepower is a function of its ability to push the vehicle in a direction, the height shouldn't affect that ability, therefore we understand this data in the scatter plot.
- Torque to Fuel Tank Capacity
    - Just as we see the scatter plot between horsepower to fuel tank, these two variables show very similar behaviors, we believe mainly because of the size or capacity of a fuel tank, because torque is the measurement of how much force a vehicle will produce at initial throttle, as we can see for all regions in the event that a vehicle does possess a larger fuel capacity tank, there will be a correlation

between the two variables. Something to note, looking closely we can see that the correlation between fuel tank and torque are stronger with american cars than Asian or European. Mainly because american cars possess iron fuel tanks and typically have higher torque than asian and european cars.

- Torque and Curb Weight
  - For torque and curb weight, we can see a strong correlation between these two continuous variables, that is because the heavier a vehicle is the less torque it will be able to produce. It's important to note that the european vehicles have the most correlation between torque and its weight. This may be because european vehicles are not as heavy as asian and american vehicles, it's also important fact to note, most european vehicles do have an forced induction component , such as turbos, which give them better torque.
- Torque to Top Speed
  - Before analyzing this data it's important to note that torque is engaged during the initial burst of driving a vehicle, after initial throttle torque does not have much influence on the experience of driving. Therefore with this knowledge we can see that there isn't such a strong correlation between these two variables, mainly because torque only plays a role in top speed during the beginning of driving.
- Torque to Length
  - Looking at our data, just as max horsepower to length the correlation between the two are very similar, we can agree with this data, because just as the explanation of how max horsepower has a minor impact on length, torque also doesn't hold much correlation to length.
- Torque to width
  - Torque to width is just similar to max horsepower to width, the correlation between the two are not strong, a vehicle's width does not impact the amount of torque it will produce.
- Torque to height
  - Just as with width , and the data from max horsepower there is no correlation between a vehicle's ability to produce torque to its height, the minimal correlation we can see through the scatter plot data could be from initial aerodynamic force reducing the ability to perform fully is what we think.
- Fuel tank to curb weight
  - This correlation is very strong, just as the rationalization between horsepower to torque is, this is because fuel tank capacity is apart of the total curb weight of a vehicle. Therefore we can see the strong correlation, because curb weight is apart of the fuel tanks weight. For Example a lighter fuel tank will contribute to an overall lighter curb weight, we can especially see this rational for all three regions of car manufacturers.
- Fuel tank to top speed
  - We see a strong correlation here, why? Because looking at the fuel tank we can rationalize that with a larger fuel tank the vehicle will not have a higher top speed. This is because of the added extra weight to the vehicle which if we see the

correction from top speed to curb weight there is a strong correlation between them.

- Fuel Tank to length
  - With fuel tank to length we can see the correlation in the scatter plot data to be seldomly correlated, this is explained because as the vehicle is longer or shorter the vehicle is this is depend minimally but still depend on the size of the gas tank.
- Fuel Tank Width
  - With fuel tank to width we can see the correlation in the scatter plot data to be seldomly correlated, this is explained because as the vehicle is wider the vehicle is this is depend minimally but still depend on the size of the gas tank.
- Fuel Tank Height
  - It's important to note, that there is zero correlation between fuel tank and height that is mainly because, if we think about it the fuel tank height doesn't play any role in the height of a vehicle's height
- Curb Weight to Top Speed
  - As seen in the scatter plot data, we can summarize that there is a lower correlation between curb weight and top speed, this can be explained top speed doesn't have much effect on rub weight of the vehicle and if there was it would be because there was more weight on the vehicle, via more effective for max torque.

Analysis

Normality

Looking at the histograms, the data is in almost all cases normal. In a few cases, the maximum does not belong with the majority of the data, or the data is bivariate. In general though, we can assume normality. We were unable to plot a PDF curve with a dot diagram underneath due to the sheer number of data points for all the regions. Since we work out of Octave (Danny works with Matlab, and he did the scatter plots), the plotting is very slow with more than 1000 data points, and the European cars data set has over 16,000.

Central Limit Theorem

To infer whether or not the central limit theorem applies to the data we have, we calculate a sample mean for each region, and then the population mean for each region.

For American cars, we have a population of 1921, so we took 500 as a statistically large sample. The sample mean is:

147.62 191.89 15.735 3157.4 120.58 176.19

And the population mean is:

145.18 191.48 15.856 3165.2 119.82 176.76

For Asian cars, we have a population of 3342, so we took 1000 as a statistically large sample. The sample mean is:

133.21 167.66 15.319 3046.3 116.89 172.79

And the population mean is:

133.18 167.92 15.382 3054.7 116.83 172.73

For European cars, we have a population of 16935, so we took 6000 as a statistically large sample. The sample mean is:

149.4 192.08 15.655 3103.4 125.47 173.24

And the population mean is:

150.21 193.03 15.675 3103.4 125.5 173.23

These all are normal according to the central limit theorem, since the variance between the sample mean for each is within a small finite range of the population mean.

Independence

Independence cannot be assumed across all features analyzed. Looking at correlations between features using the 2D scatter plot summary, there appears to be strong correlation between some features. For example fuel tank capacity and curb weight seem to be strongly correlated as they are linearly fit. This is exemplified in Table 1 as the correlation between the fuel tank capacity and curb weight for American cars is over 0.8. This makes intuitive sense because the heavier a car weighs, the more energy it will need to move and therefore heavier cars will be fitted with larger fuel capacities. There are strong correlations between other features such as top speed and maximum horsepower. Principle component analysis will be used to limit the feature set by removing dependent features.

Table 1: Correlation Matrix for all cars

	maximum horsepower	maximum torque	fuel tank capacity	Curb weight	Top Speed	Length	Width	height
maximum horsepower	1	0.78982977	0.5612432	0.60973	0.82386638	0.484228	0.55137	-0.05121
maximum torque	0.789829768	1	0.6097648	0.802452	0.633136911	0.600892	0.69436	0.205579
fuel tank capacity	0.561243151	0.60976479	1	0.777376	0.380161677	0.747517	0.64925	0.314841
Curb weight	0.60973023	0.80245186	0.7773757	1	0.39746656	0.750277	0.81524	0.555784
Top Speed	0.82386638	0.63313691	0.3801617	0.397467	1	0.454056	0.48341	-0.2994
Length	0.484227886	0.60089186	0.7475171	0.750277	0.45405559	1	0.76564	0.228399
Width	0.551370699	0.69436282	0.6492456	0.815241	0.483409996	0.765645	1	0.419747
height	-0.051207931	0.20557893	0.3148405	0.555784	-0.299401598	0.228399	0.41975	1

Table 2: Covariance Matrix for all cars

	maximum horsepower	maximum torque	fuel tank capacity	Curb weight	Top Speed	Length	Width	height
maximum horsepower	4917.982482	4561.36556	125.65444	28861.97	1021.431318	490.7045	133.081	-20.0623
maximum torque	4561.36556	6781.65644	160.31107	160.3111	921.7739387	715.0575	196.804	94.57931
fuel tank capacity	125.6544373	160.311073	10.192168	1675.171	21.4565811	34.48507	7.13383	5.615308
Curb weight	28861.97102	44604.8053	1675.1705	455606.1	4743.012463	7318.013	1893.92	2095.801
Top Speed	1021.431318	921.773939	21.456581	4743.012	312.5487743	115.9964	29.4141	-29.5708
Length	490.7044941	715.057491	34.485073	7318.013	115.9964444	208.8111	38.0789	18.43826
Width	133.081441	196.80434	7.133828	1893.918	29.41405156	38.07889	11.8457	8.070813
height	-20.06228304	94.5793073	5.6153076	2095.801	-29.57077131	18.43826	8.07081	31.2104

Principal Component Analysis

Correlation

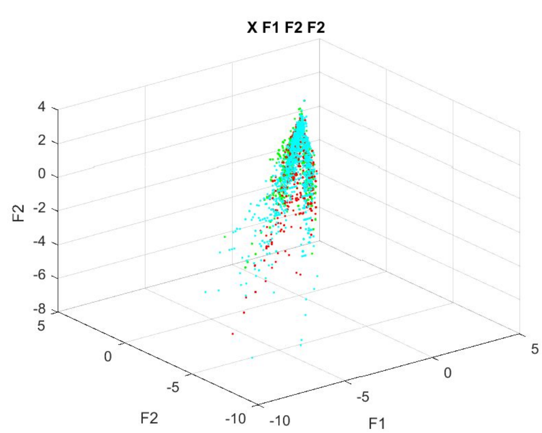
The correlation matrix puts quantitative data to how strongly related two features are. For PCA, features that are strongly related should be reduced. This makes sense when trying to limit your features because if two features are strongly related you can reasonably estimate one feature value from the other, and therefore only one is needed.

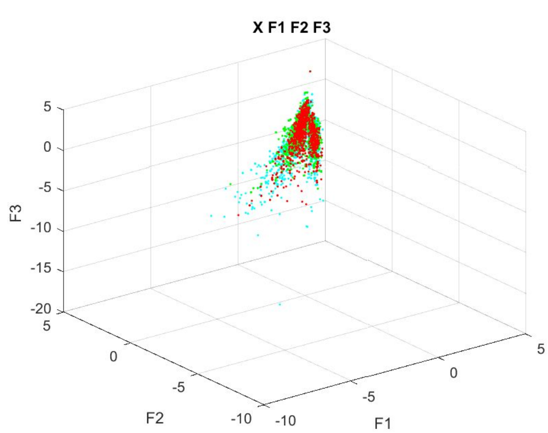
Plots

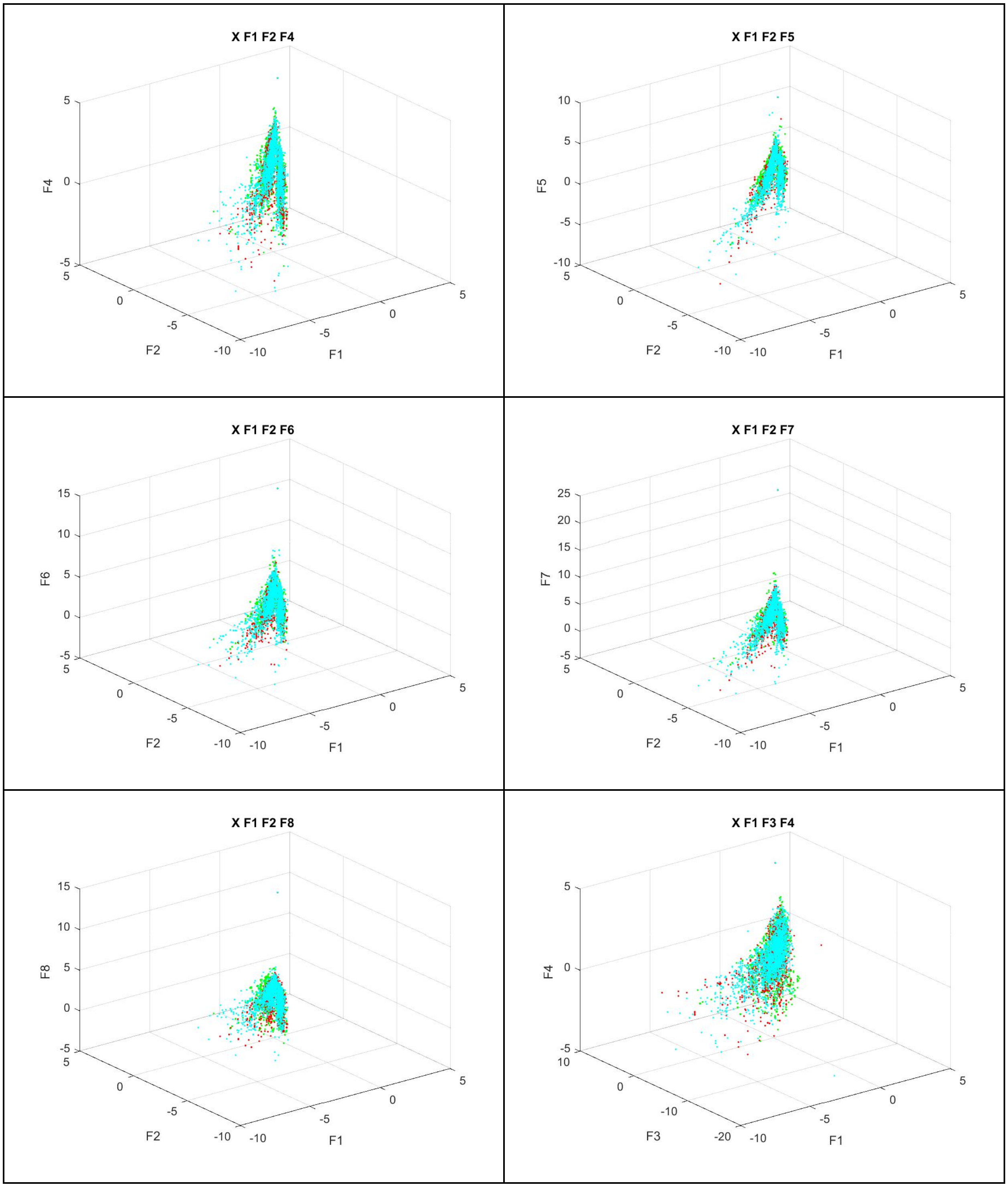
Legend	
American Cars	
European Cars	
Asian Cars	

Original Feature 3D Scatter Plots

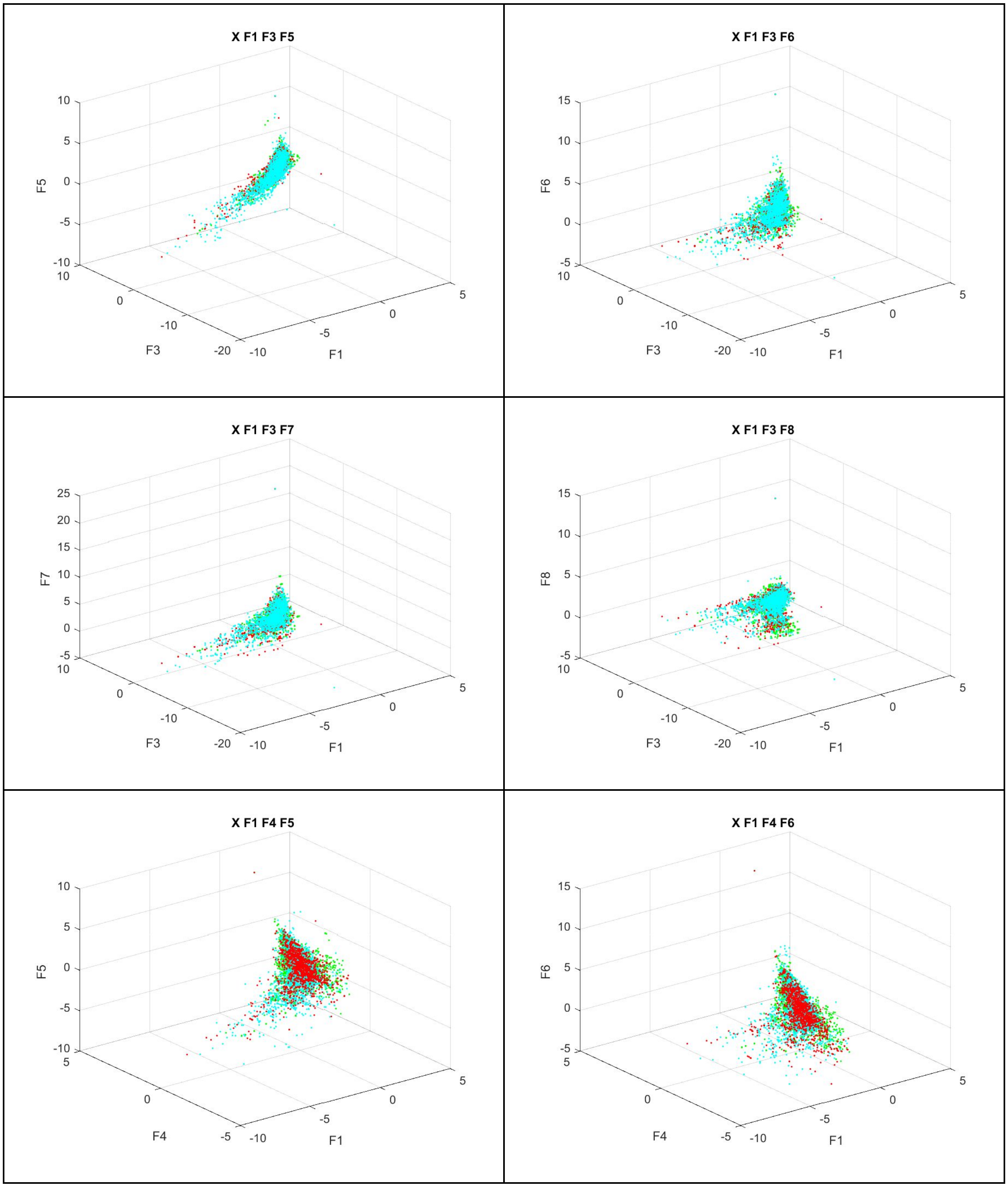
Each unique combination of three features is plotted below. This is used for verifying the three features kept after PCA. The scatter plot using the three kept features should produce clustering that can be used for classifying observations.

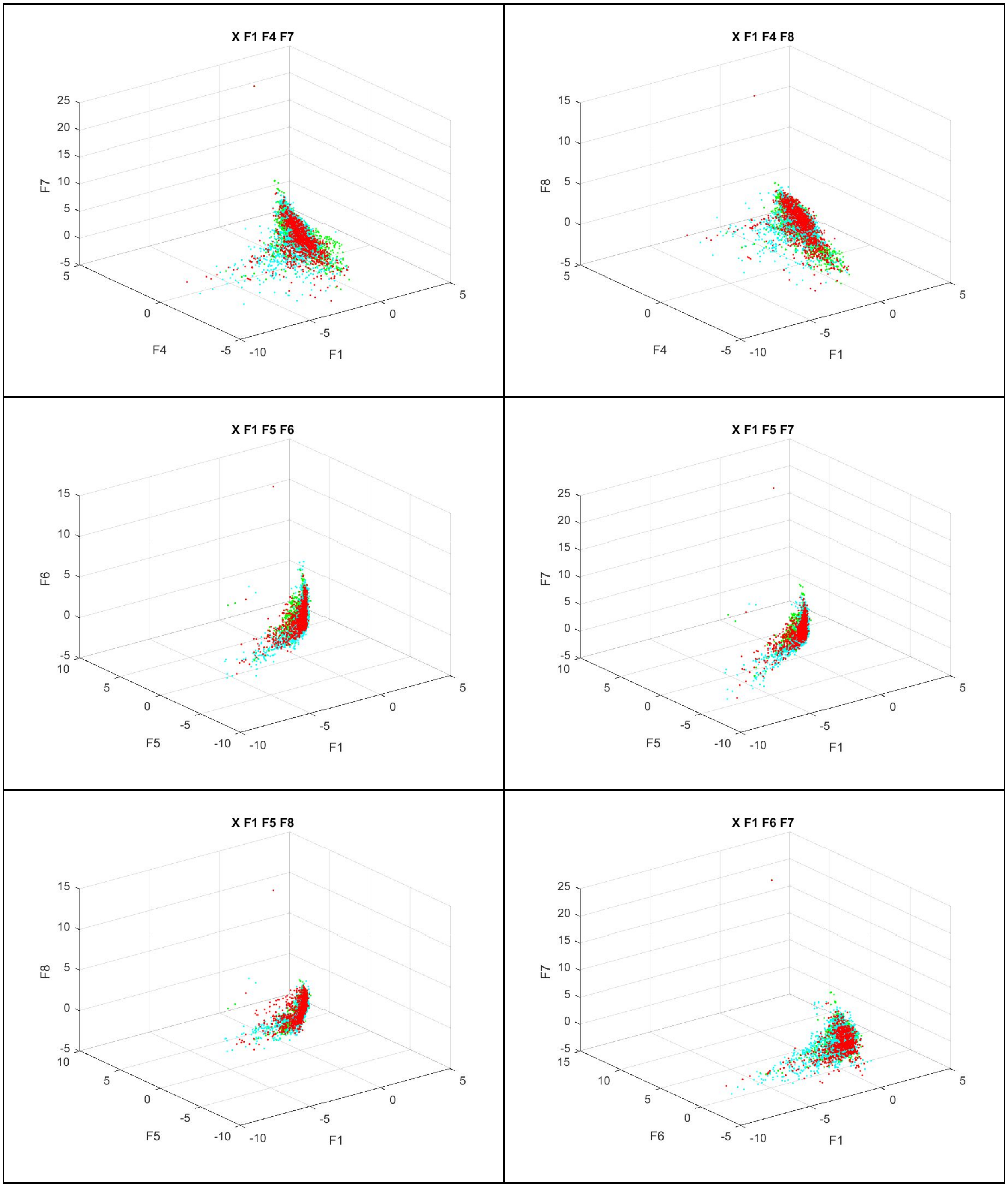


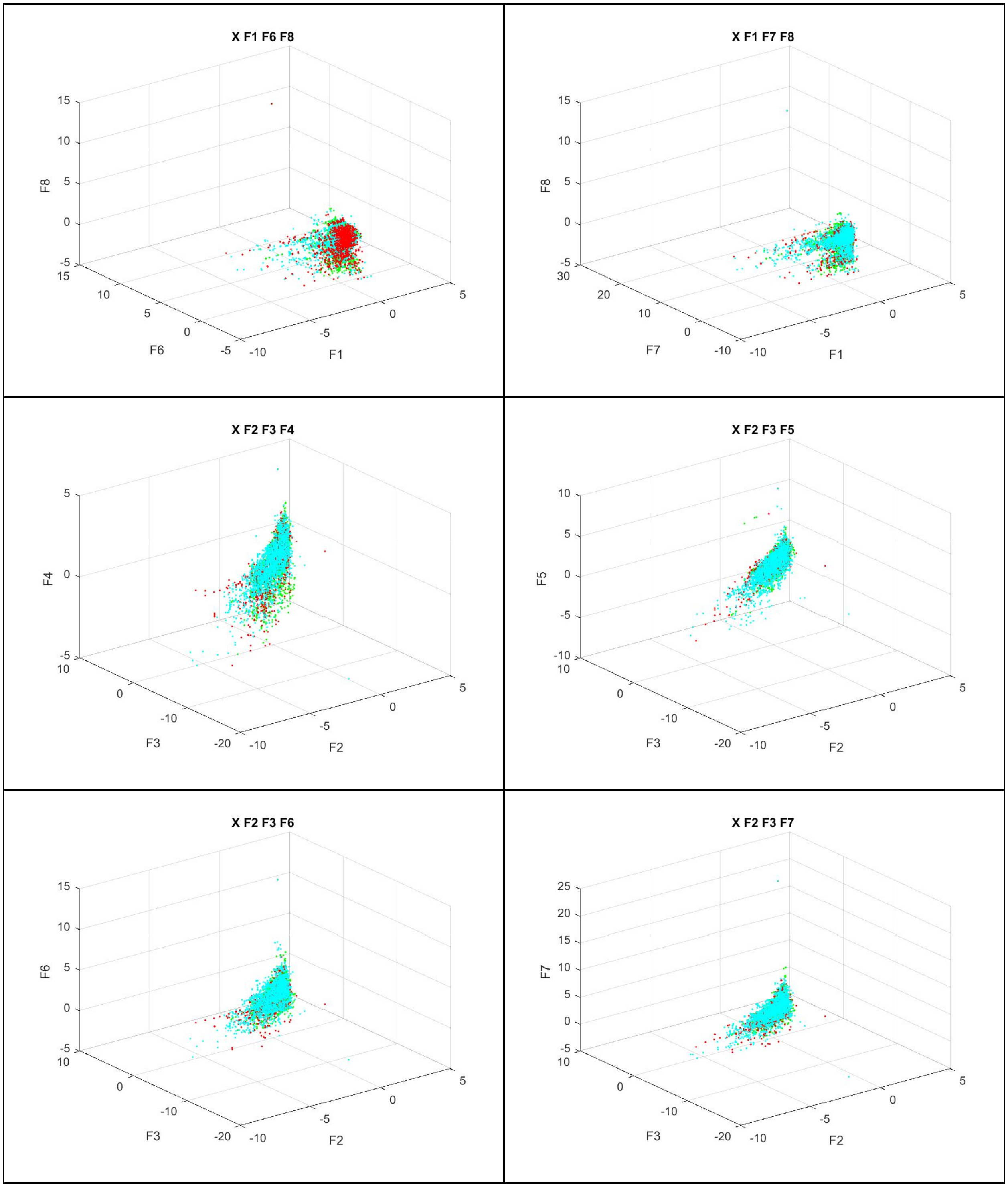


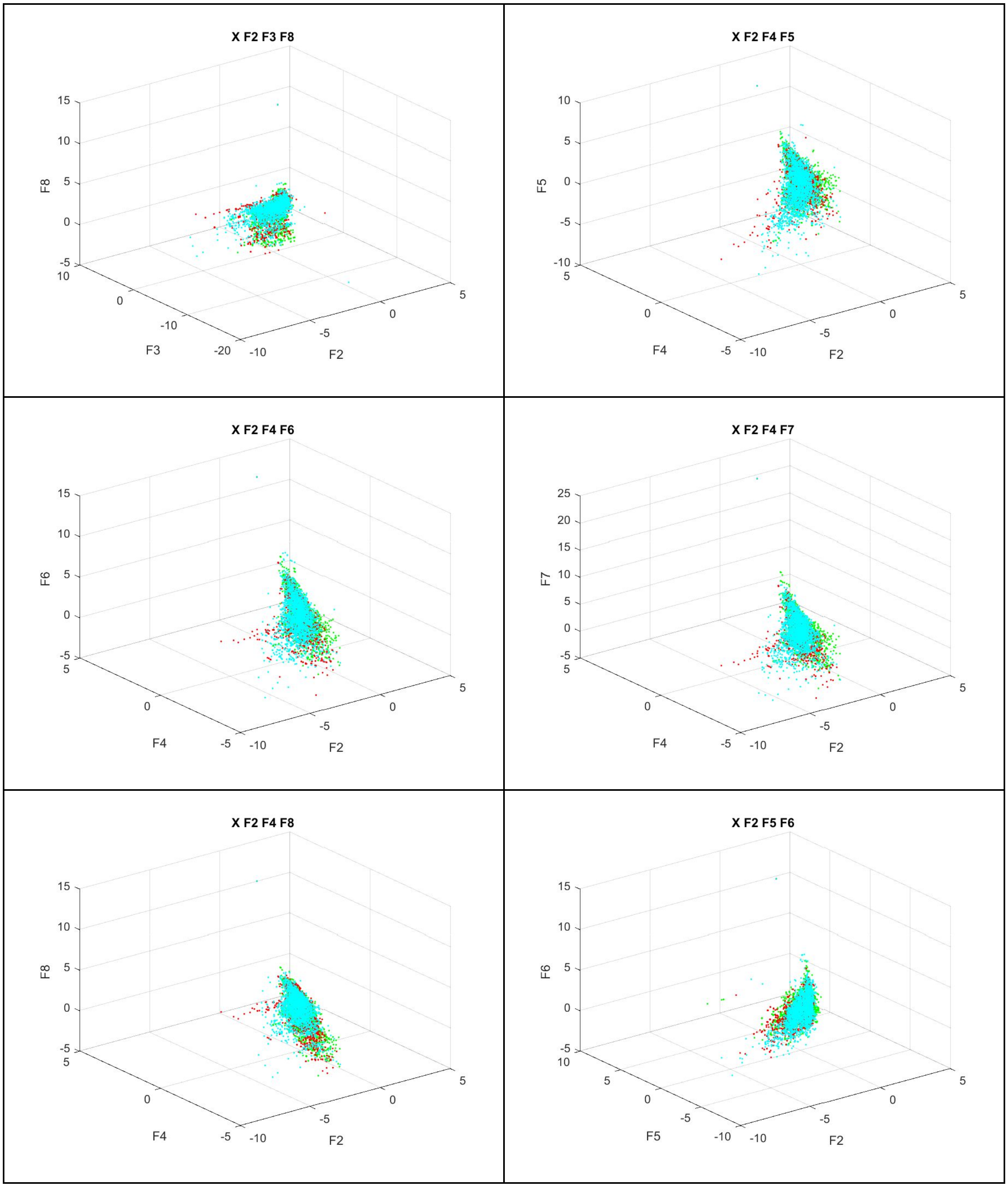


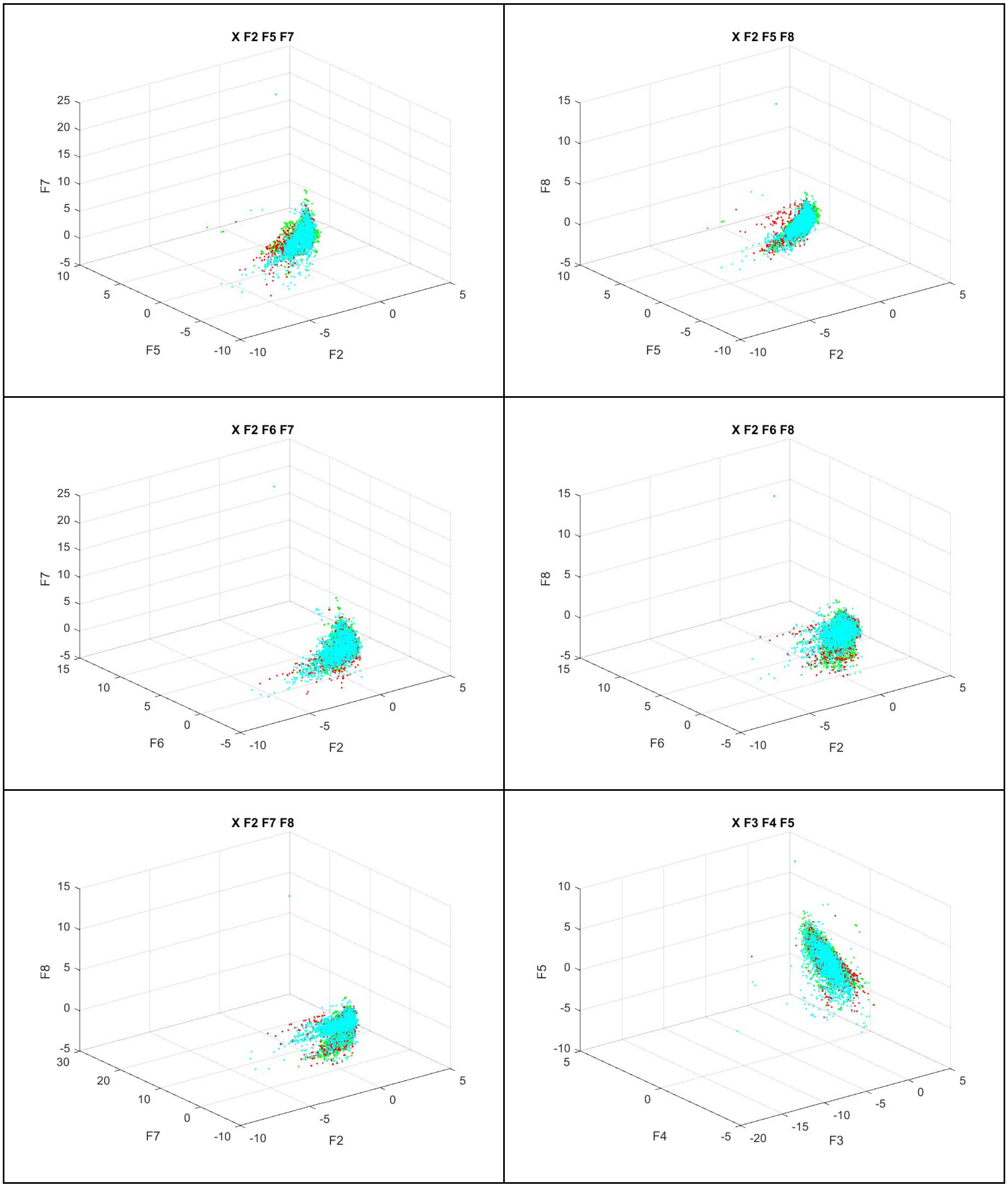
X F1 F2 F4
F4
F2
F1
X F1 F2 F5
F5
F2
F1
X F1 F2 F6
F6
F2
F1
X F1 F2 F7
F7
F2
F1
X F1 F2 F8
F8
F2
F1
X F1 F3 F4
F4
F3
F1

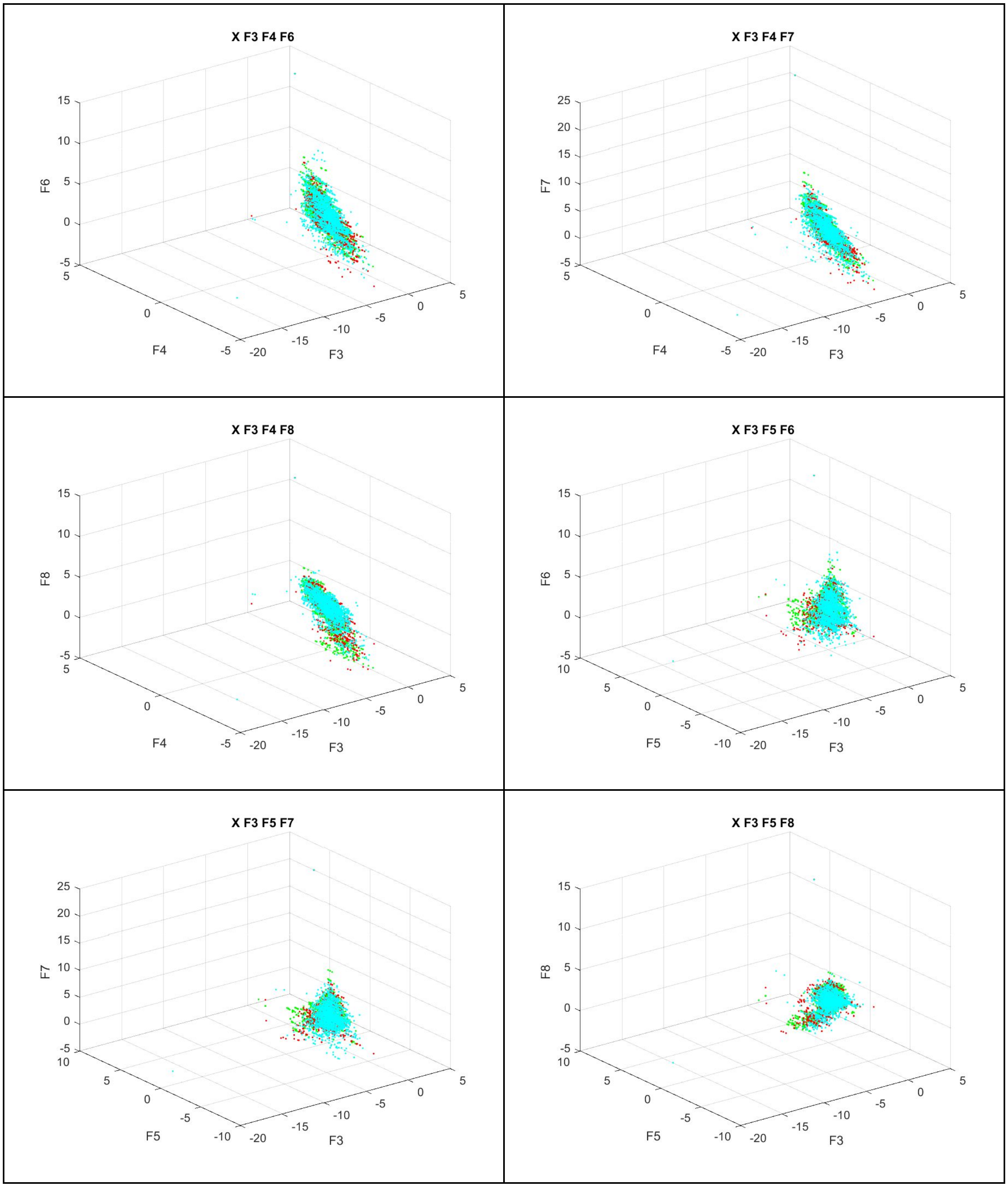

X F1 F3 F5
F5
F3
F1
X F1 F3 F6
F6
F3
F1
X F1 F3 F7
F7
F3
F1
X F1 F3 F8
F8
F3
F1
X F1 F4 F5
F5
F4
F1
X F1 F4 F6
F6
F4
F1

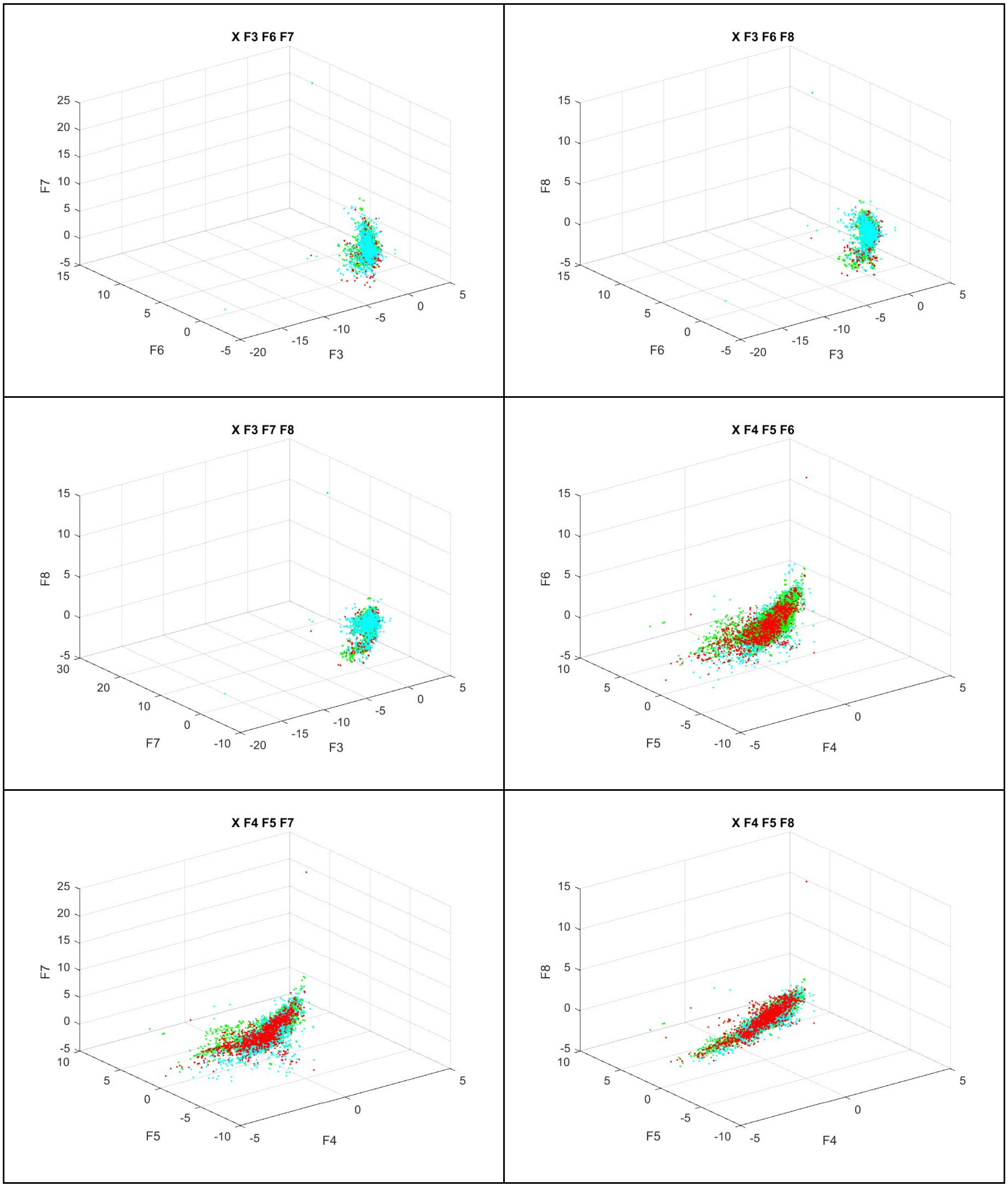

X F1 F4 F7
F7
F4
F1
X F1 F4 F8
F8
F4
F1
X F1 F5 F6
F6
F5
F1
X F1 F5 F7
F7
F5
F1
X F1 F5 F8
F8
F5
F1
X F1 F6 F7
F7
F6
F1

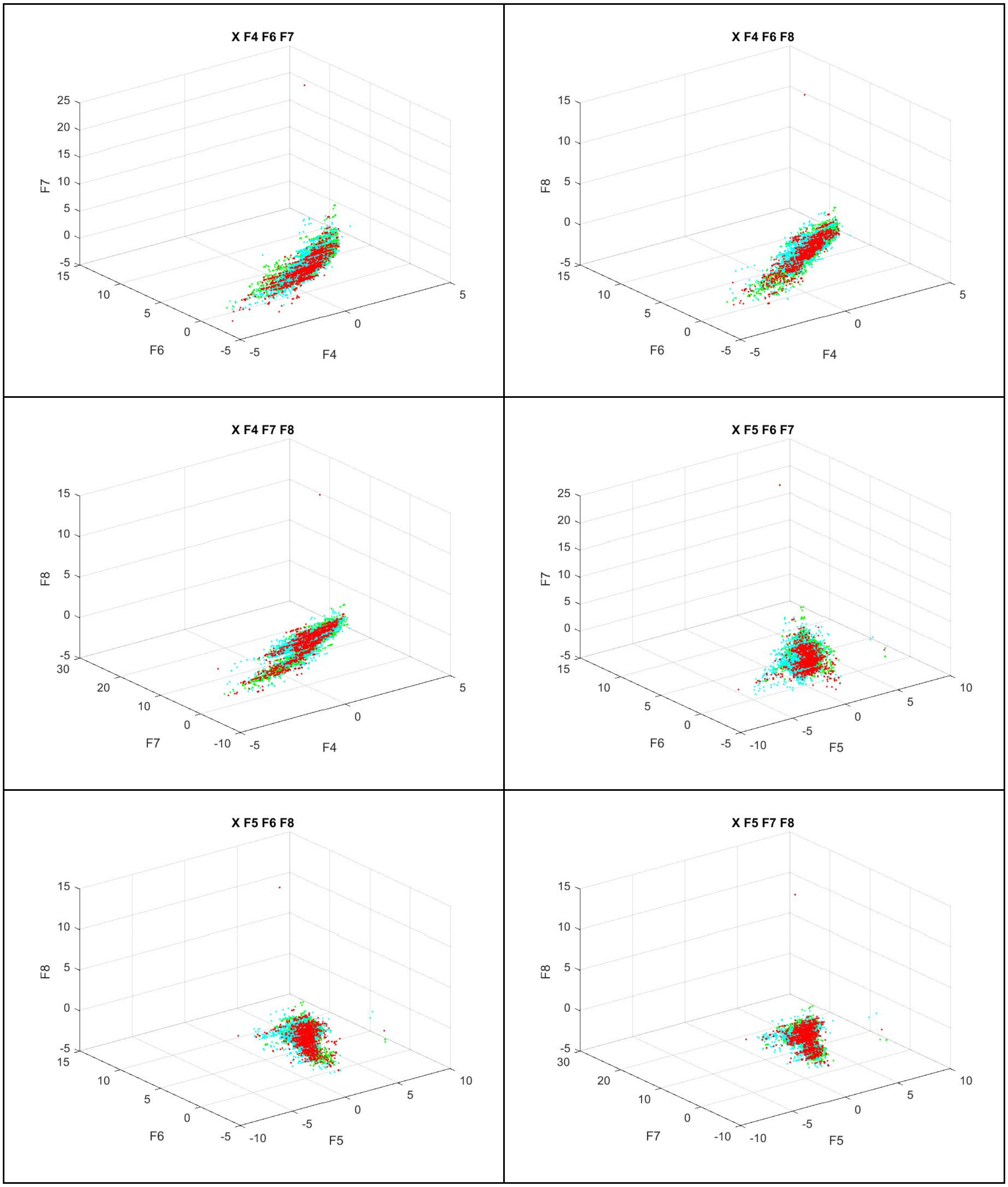

X F1 F6 F8
F8
F6
F1
X F1 F7 F8
F8
F7
F1
X F2 F3 F4
F4
F3
F2
X F2 F3 F5
F5
F3
F2
X F2 F3 F6
F6
F3
F2
X F2 F3 F7
F7
F3
F2

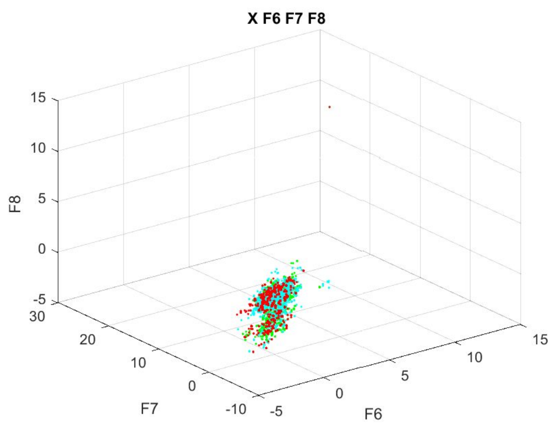

X F2 F3 F8
F8
15
10
5
0
-5
F3
10
0
-10
-20
F2
-10
-5
0
5
X F2 F4 F5
F5
10
5
0
-5
-10
F4
5
0
-5
F2
-10
-5
0
5
X F2 F4 F6
F6
15
10
5
0
-5
F4
5
0
-5
F2
-10
-5
0
5
X F2 F4 F7
F7
25
20
15
10
5
0
-5
F4
5
0
-5
F2
-10
-5
0
5
X F2 F4 F8
F8
15
10
5
0
-5
F4
5
0
-5
F2
-10
-5
0
5
X F2 F5 F6
F6
15
10
5
0
-5
F5
10
5
0
-5
-10
F2
-10
-5
0
5


X F2 F5 F7
F7
F5
F2
X F2 F5 F8
F8
F5
F2
X F2 F6 F7
F7
F6
F2
X F2 F6 F8
F8
F6
F2
X F2 F7 F8
F8
F7
F2
X F3 F4 F5
F5
F4
F3

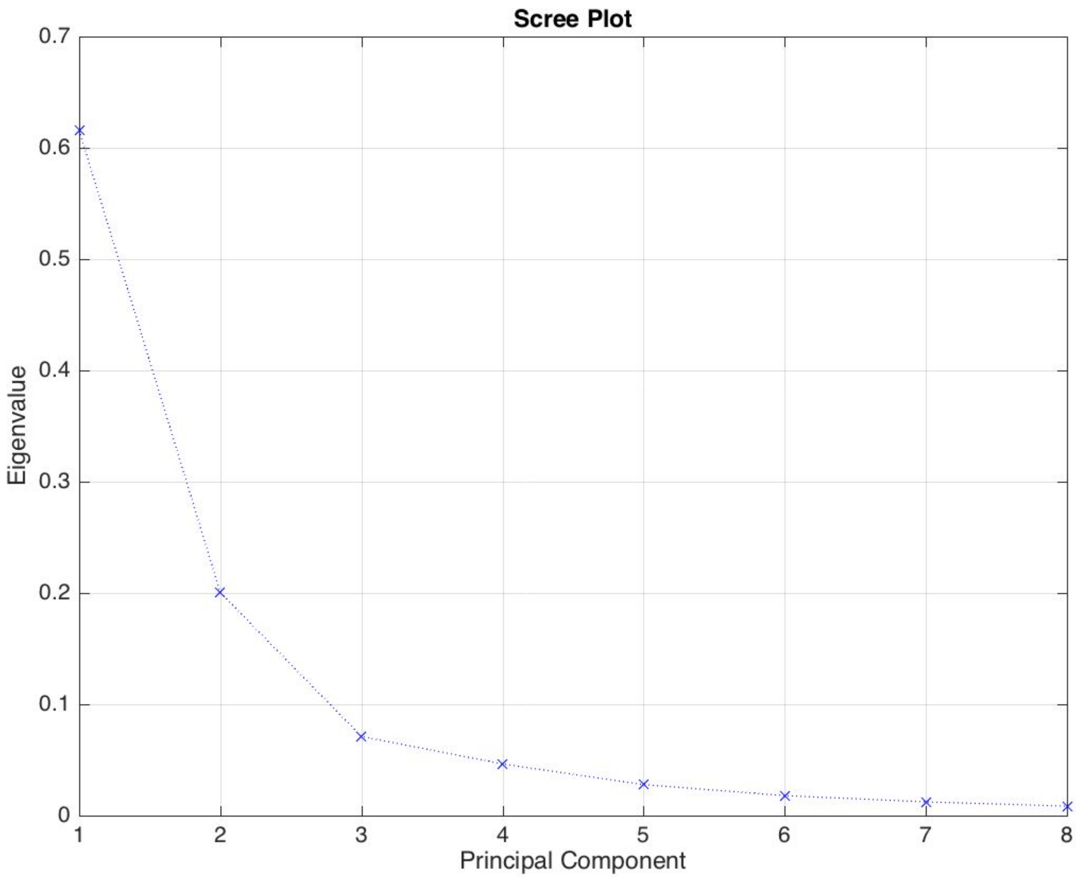

X F3 F4 F6
F6
F4
F3
X F3 F4 F7
F7
F4
F3
X F3 F4 F8
F8
F4
F3
X F3 F5 F6
F6
F5
F3
X F3 F5 F7
F7
F5
F3
X F3 F5 F8
F8
F5
F3


X F3 F6 F7
F7
F6
F3
X F3 F6 F8
F8
F6
F3
X F3 F7 F8
F8
F7
F3
X F4 F5 F6
F6
F5
F4
X F4 F5 F7
F7
F5
F4
X F4 F5 F8
F8
F5
F4

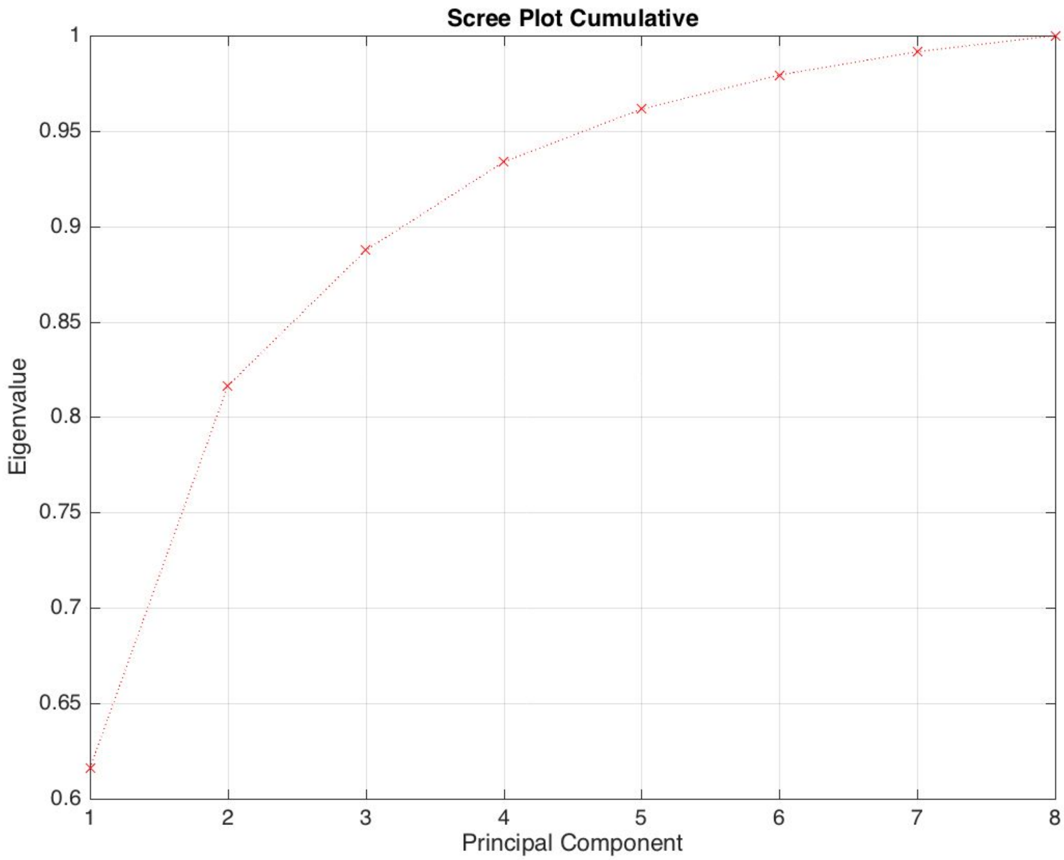

X F4 F6 F7
F7
25
20
15
10
5
0
-5
F6
15
10
5
0
-5
F4
-5
0
5
X F4 F6 F8
F8
15
10
5
0
-5
F6
15
10
5
0
-5
F4
-5
0
5
X F4 F7 F8
F8
15
10
5
0
-5
F7
30
20
10
0
-10
F4
-5
0
5
X F5 F6 F7
F7
25
20
15
10
5
0
-5
F6
15
10
5
0
-5
F5
-10
-5
0
5
10
X F5 F6 F8
F8
15
10
5
0
-5
F6
15
10
5
0
-5
F5
-10
-5
0
5
10
X F5 F7 F8
F8
15
10
5
0
-5
F7
30
20
10
0
-10
F5
-10
-5
0
5
10


X F6 F7 F8
15
10
5
0
-5
F8
30
20
10
0
-10
F7
-5
0
5
10
15
F6

Scree Plots

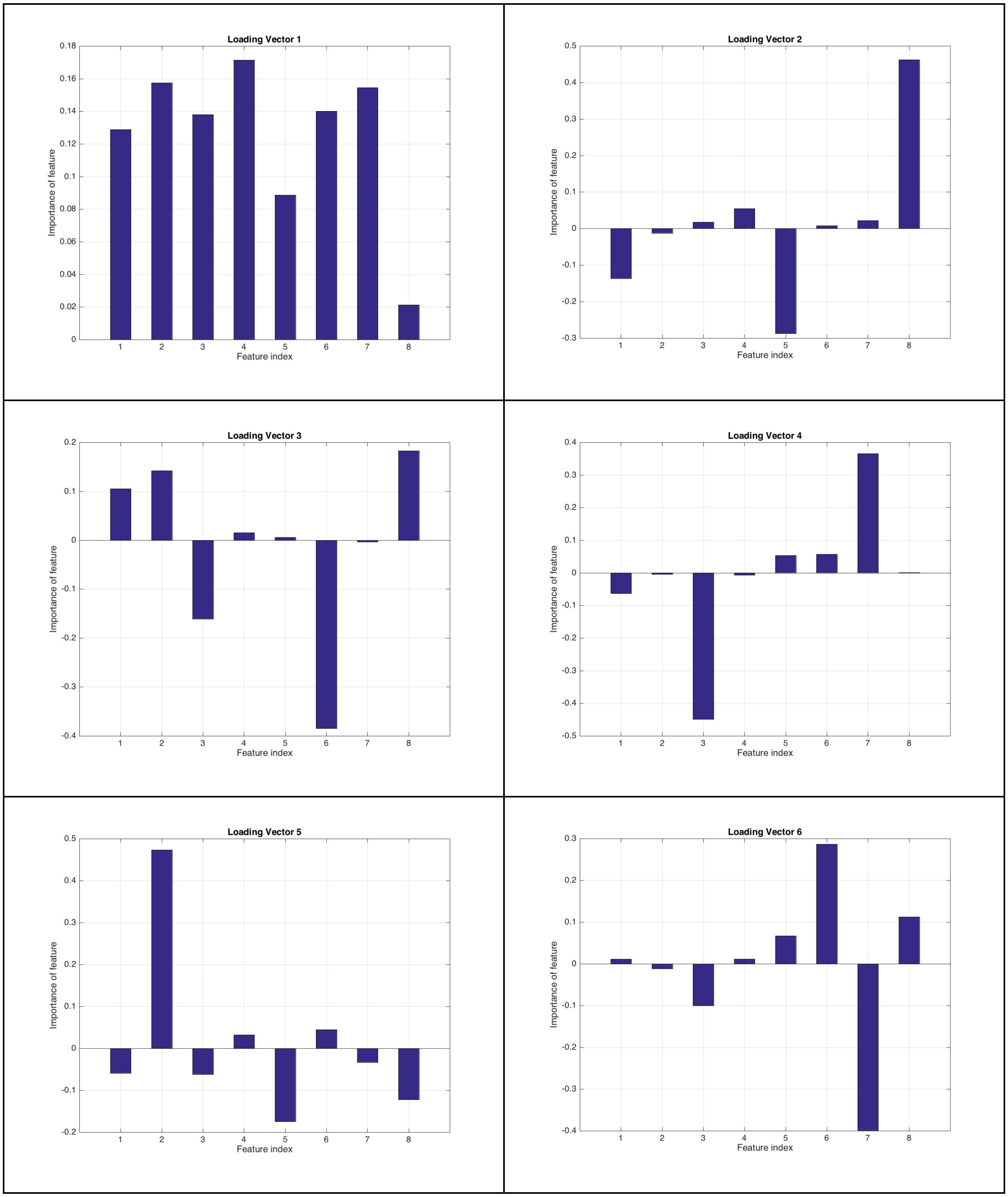


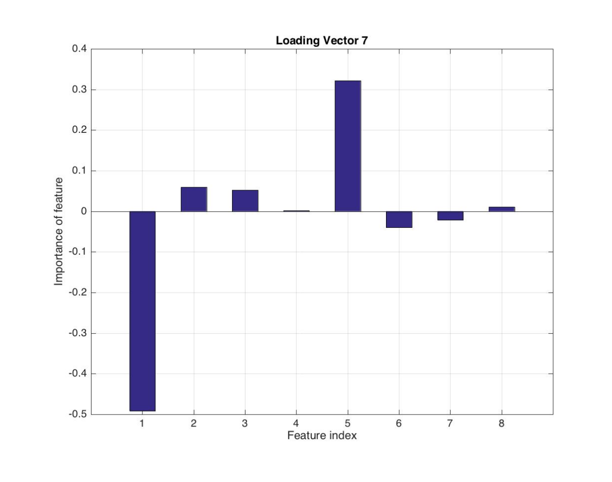
Scree Plot
Eigenvalue
Principal Component
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8

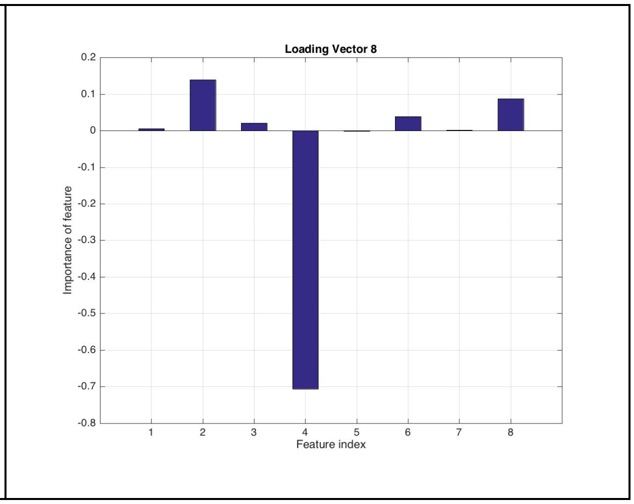


Scree Plot Cumulative
Eigenvalue
Principal Component
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
1
2
3
4
5
6
7
8

Loading Vectors

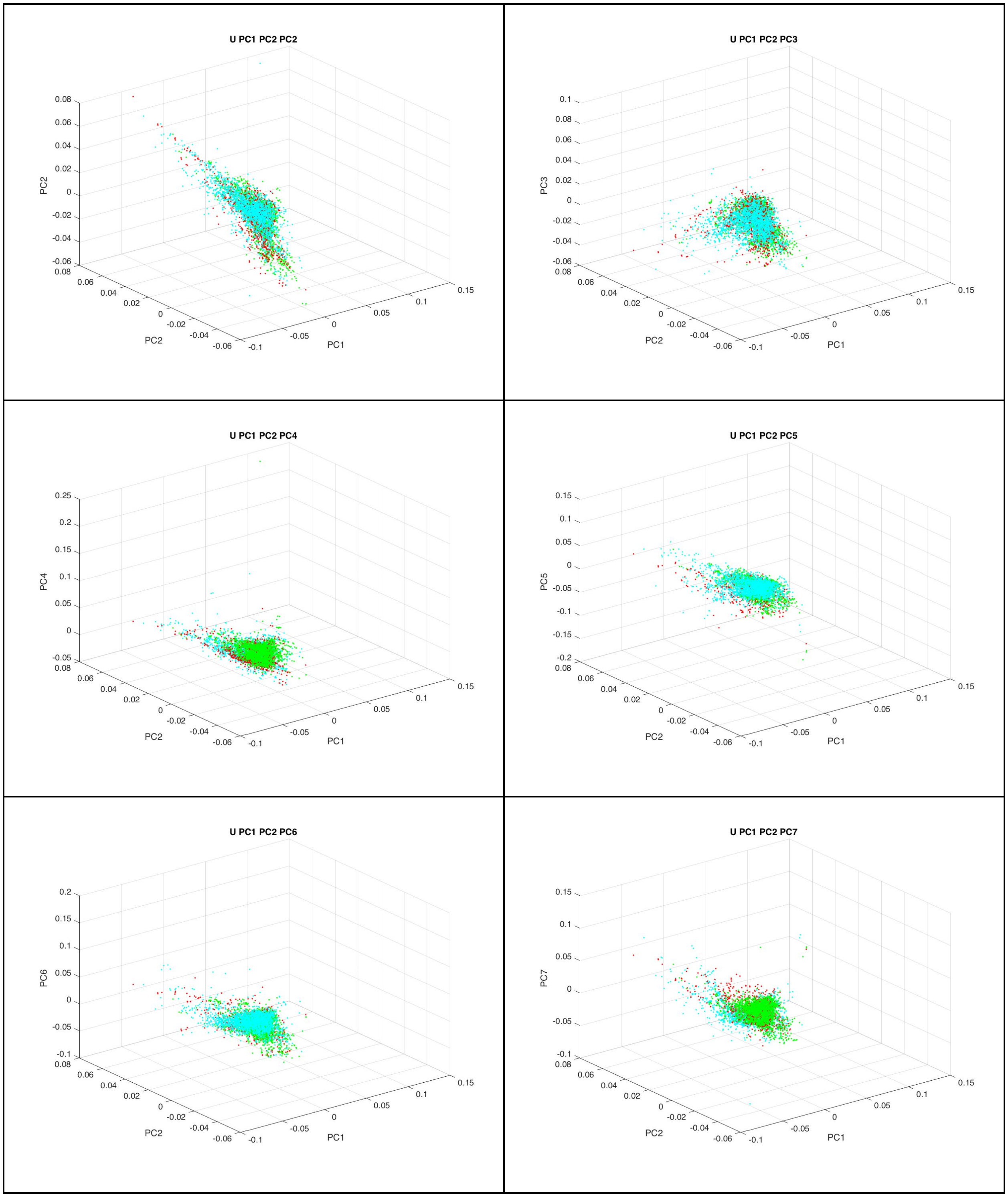


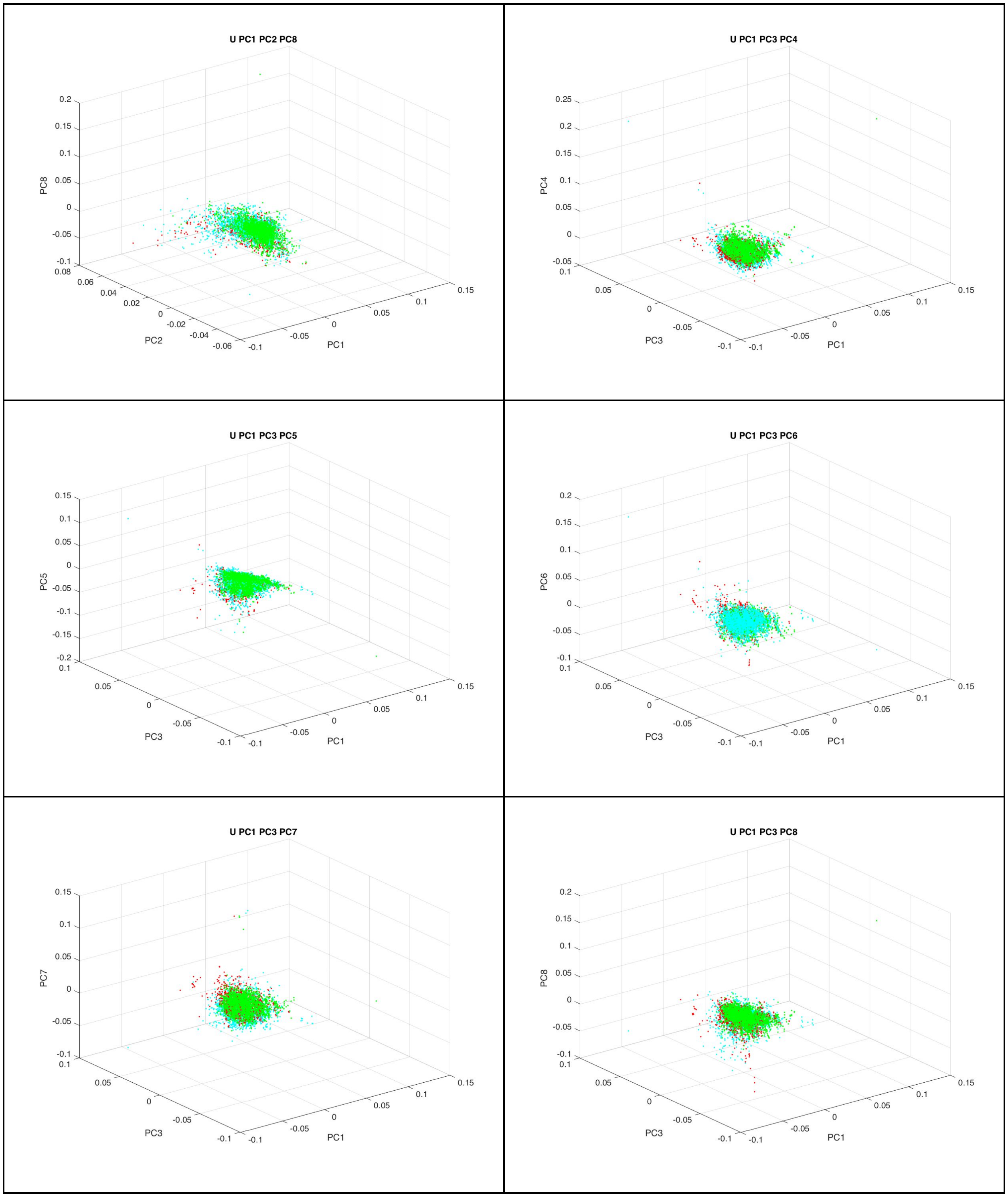

Loading Vector 7
Importance of feature
Feature index
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
1
2
3
4
5
6
7
8

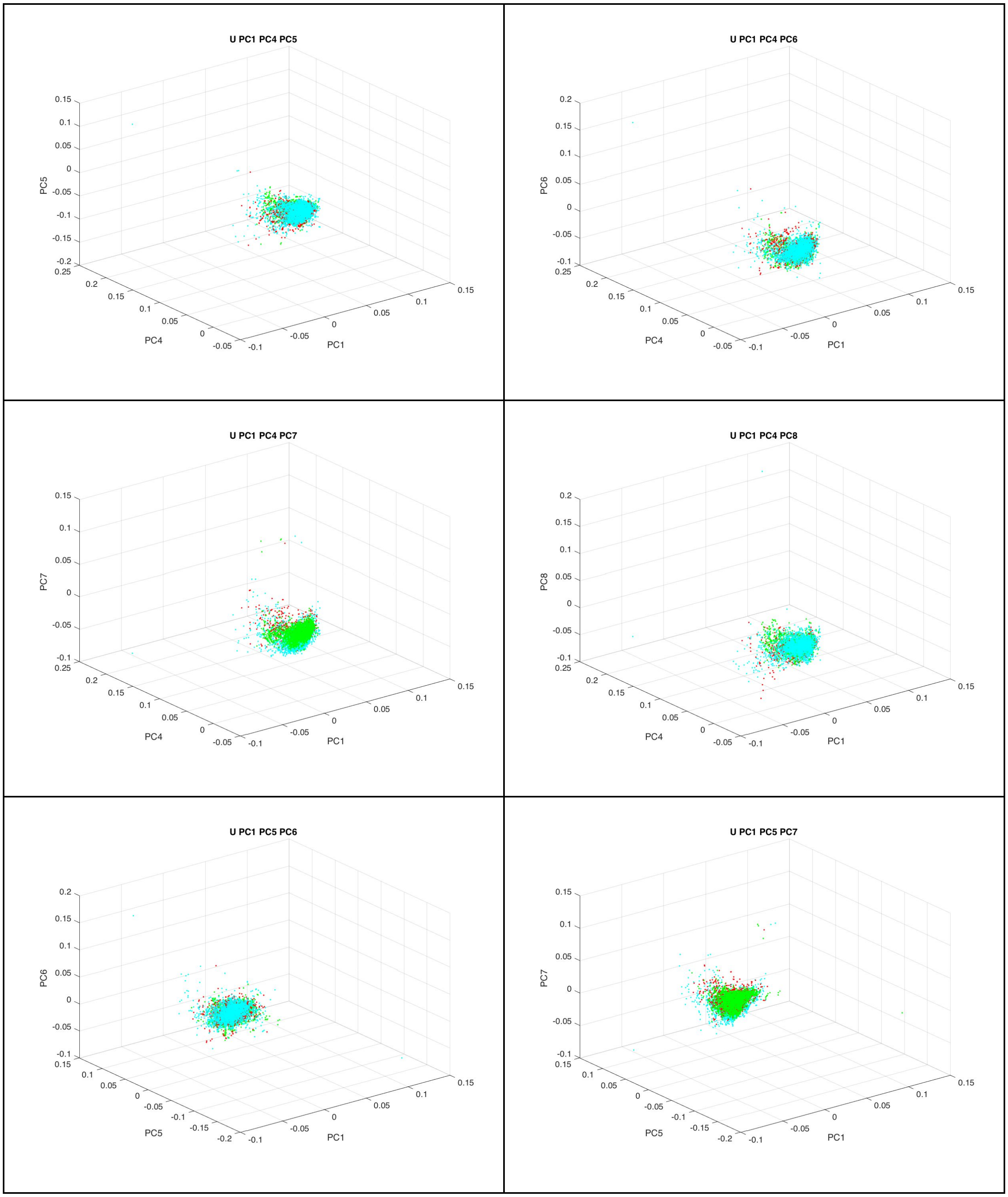

Loading Vector 8
Importance of feature
Feature index
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
1
2
3
4
5
6
7
8

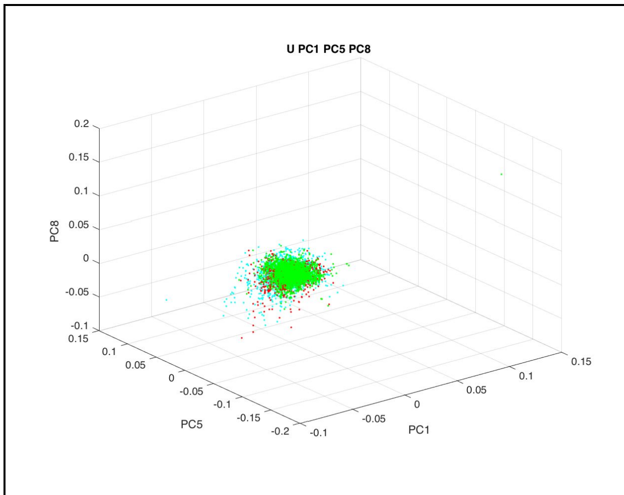
U Scatter Plots

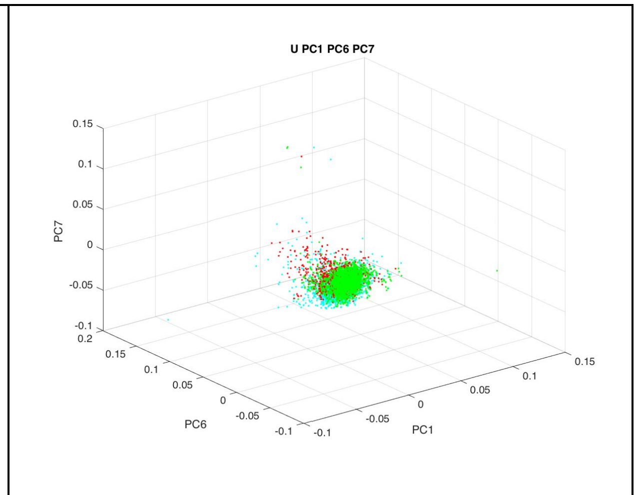
Scatter plots of U across all unique combinations of principle

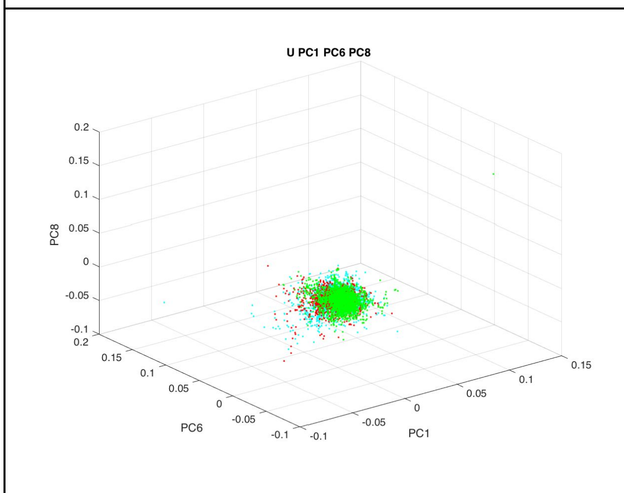


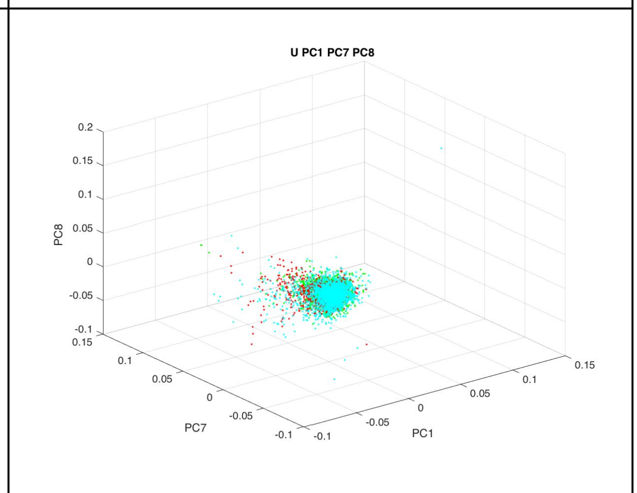

U PC1 PC2 PC8
PC8
PC2
PC1
U PC1 PC3 PC4
PC4
PC3
PC1
U PC1 PC3 PC5
PC5
PC3
PC1
U PC1 PC3 PC6
PC6
PC3
PC1
U PC1 PC3 PC7
PC7
PC3
PC1
U PC1 PC3 PC8
PC8
PC3
PC1

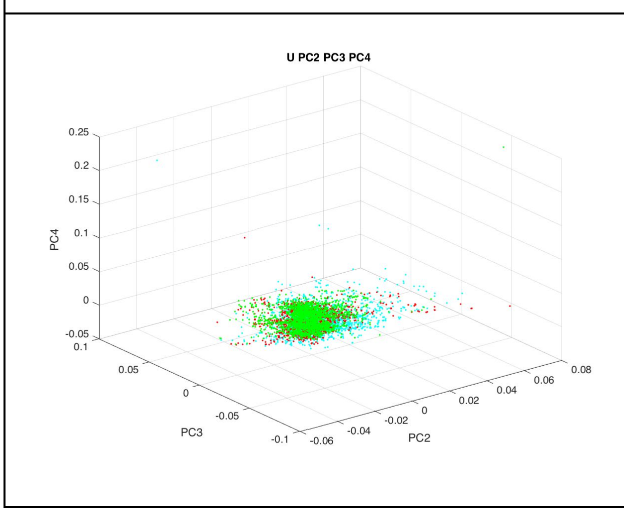

U PC1 PC4 PC5
PC5
PC4
PC1
U PC1 PC4 PC6
PC6
PC4
PC1
U PC1 PC4 PC7
PC7
PC4
PC1
U PC1 PC4 PC8
PC8
PC4
PC1
U PC1 PC5 PC6
PC6
PC5
PC1
U PC1 PC5 PC7
PC7
PC5
PC1

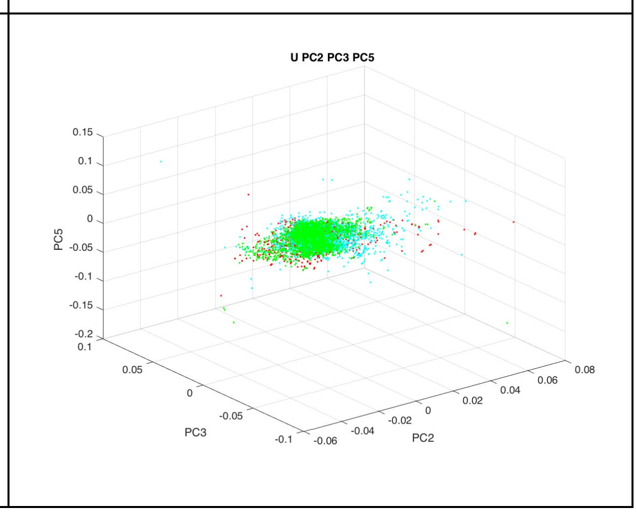

U PC1 PC5 PC8
PC8
PC5
PC1

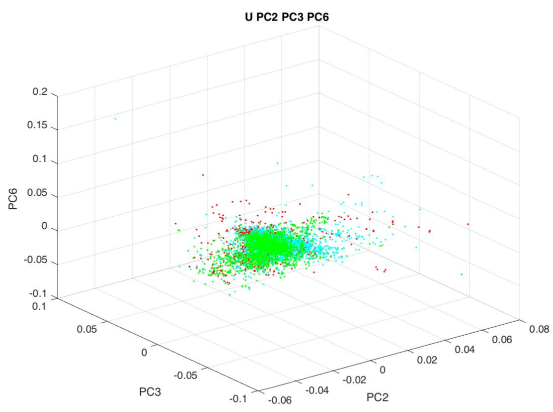

U PC1 PC6 PC7
PC7
PC6
PC1

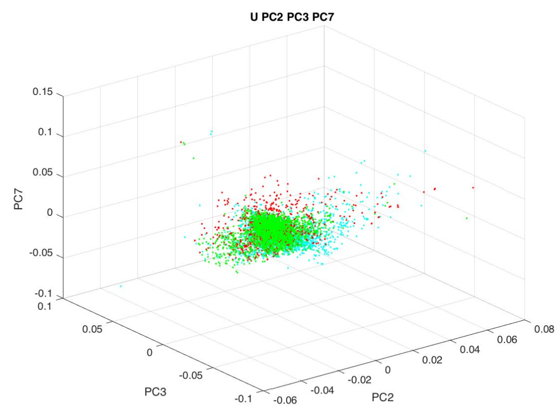

U PC1 PC6 PC8
PC8
PC6
PC1

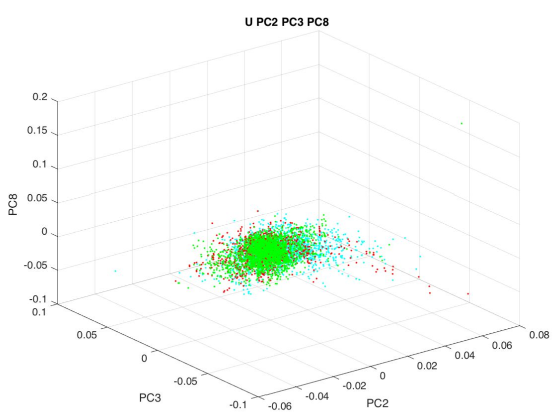

U PC1 PC7 PC8
PC8
PC7
PC1

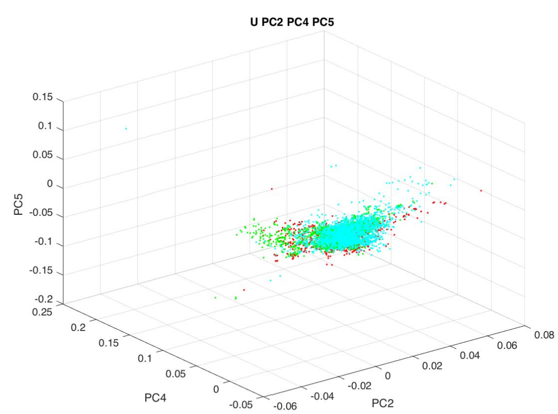

U PC2 PC3 PC4
PC4
PC3
PC2

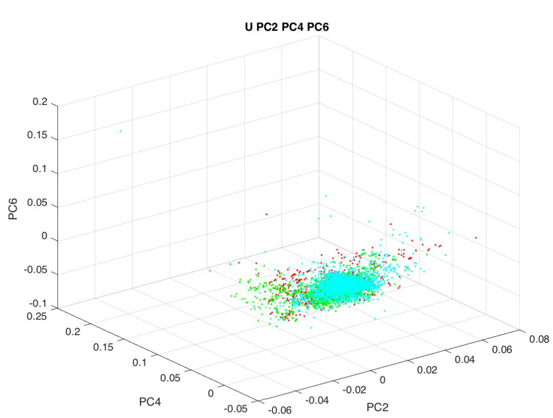

U PC2 PC3 PC5
PC5
PC3
PC2

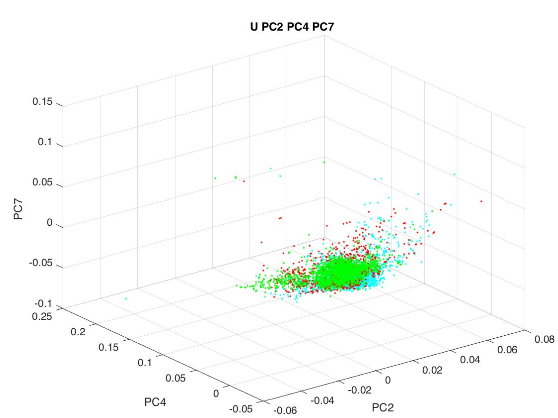

U PC2 PC3 PC6
PC6
PC3
PC2

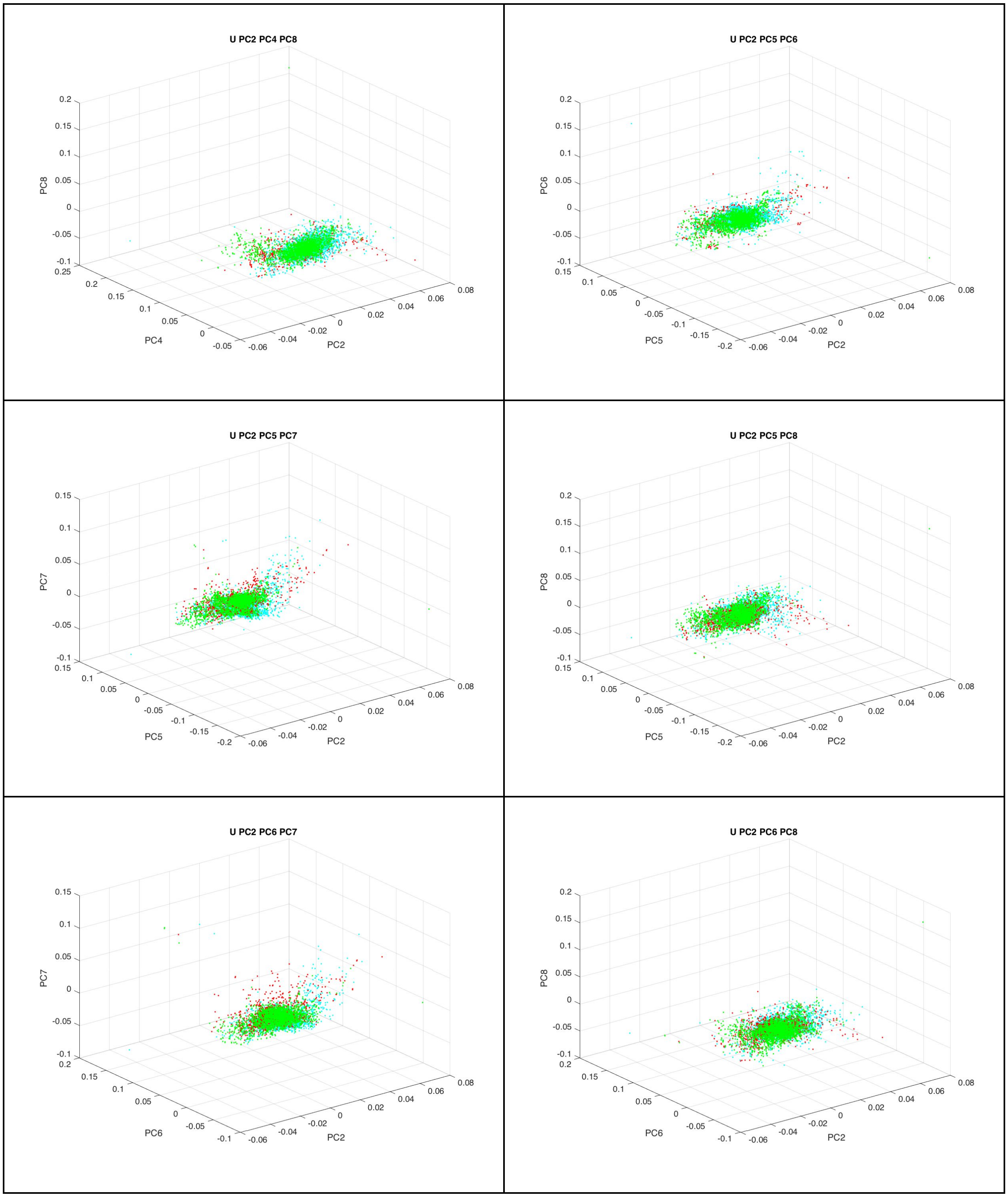

U PC2 PC3 PC7
PC7
PC3
PC2

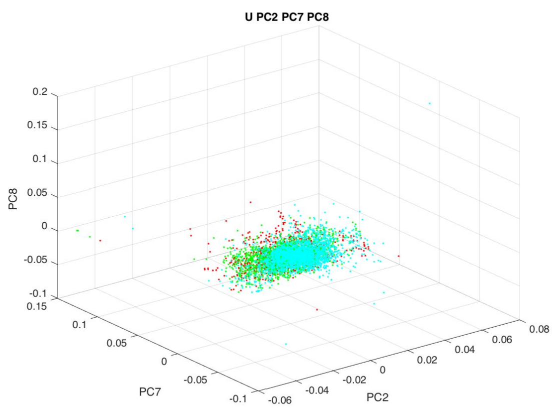

U PC2 PC3 PC8
PC8
PC3
PC2

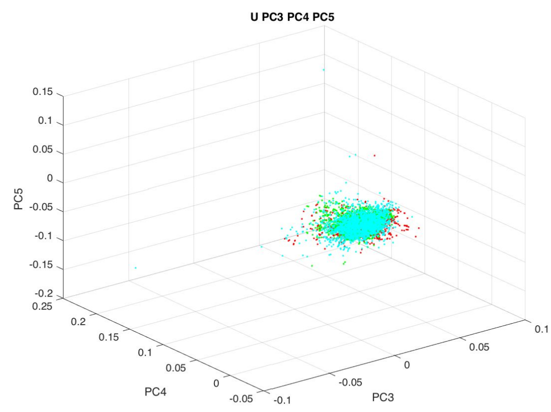

U PC2 PC4 PC5
PC5
PC4
PC2

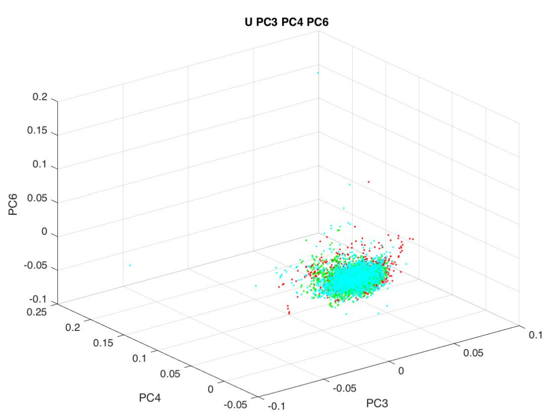

U PC2 PC4 PC6
PC6
PC4
PC2

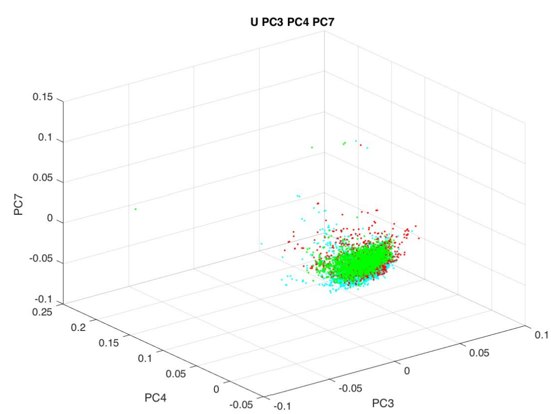

U PC2 PC4 PC7
PC7
PC4
PC2

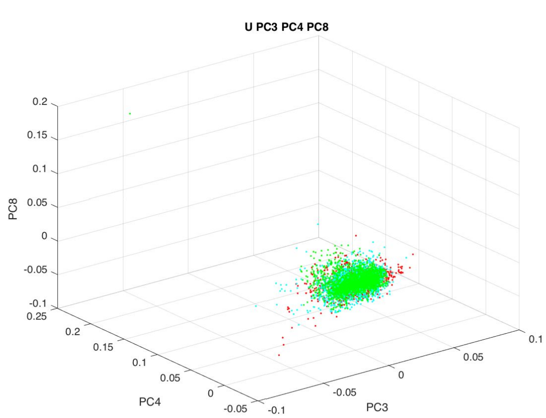

U PC2 PC4 PC8
PC8
PC4
PC2
U PC2 PC5 PC6
PC6
PC5
PC2
U PC2 PC5 PC7
PC7
PC5
PC2
U PC2 PC5 PC8
PC8
PC5
PC2
U PC2 PC6 PC7
PC7
PC6
PC2
U PC2 PC6 PC8
PC8
PC6
PC2

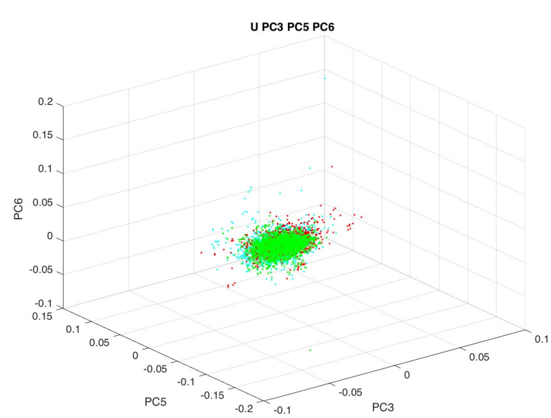

U PC2 PC7 PC8
PC8
PC7
PC2

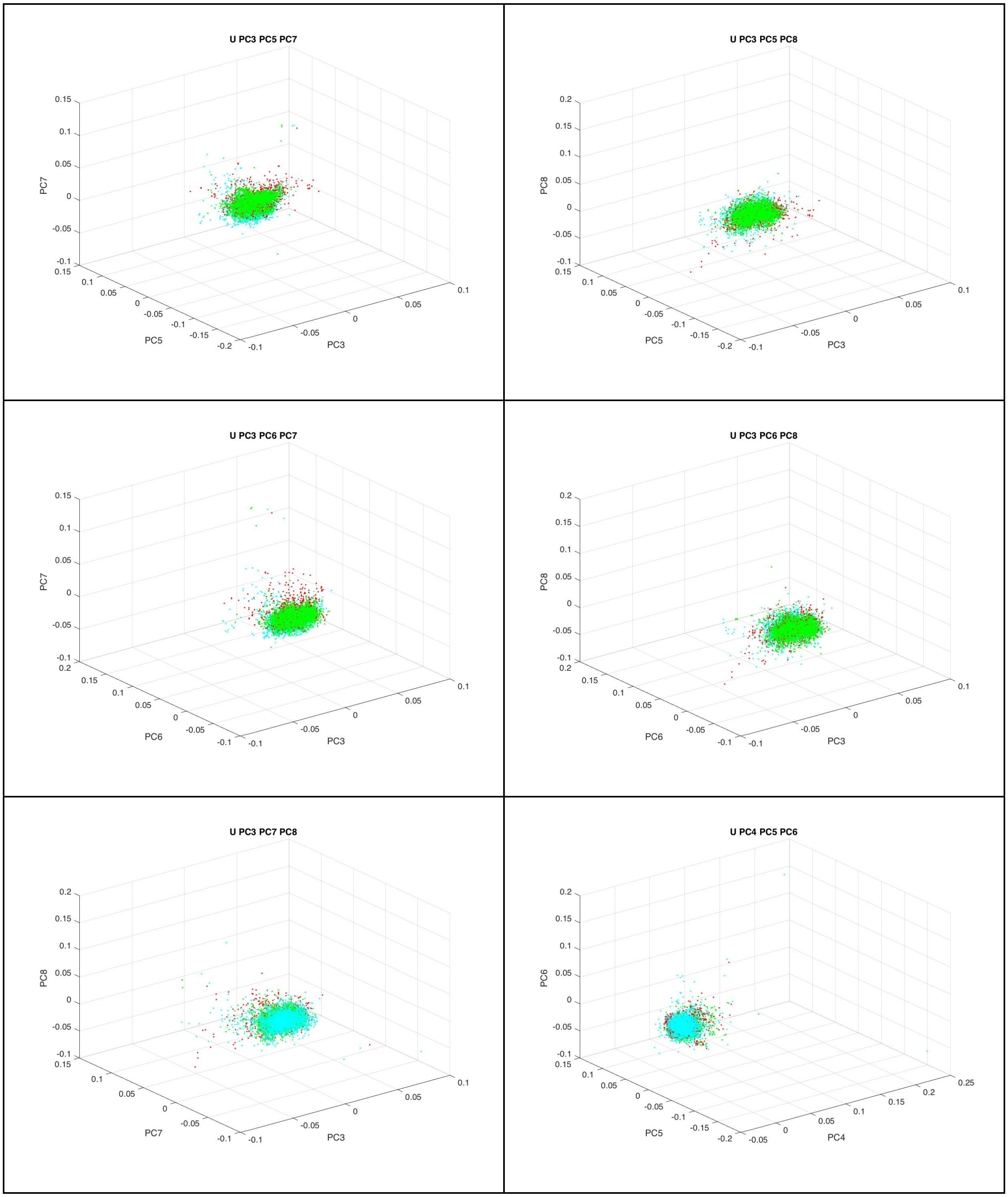

U PC3 PC4 PC5
PC5
PC4
PC3

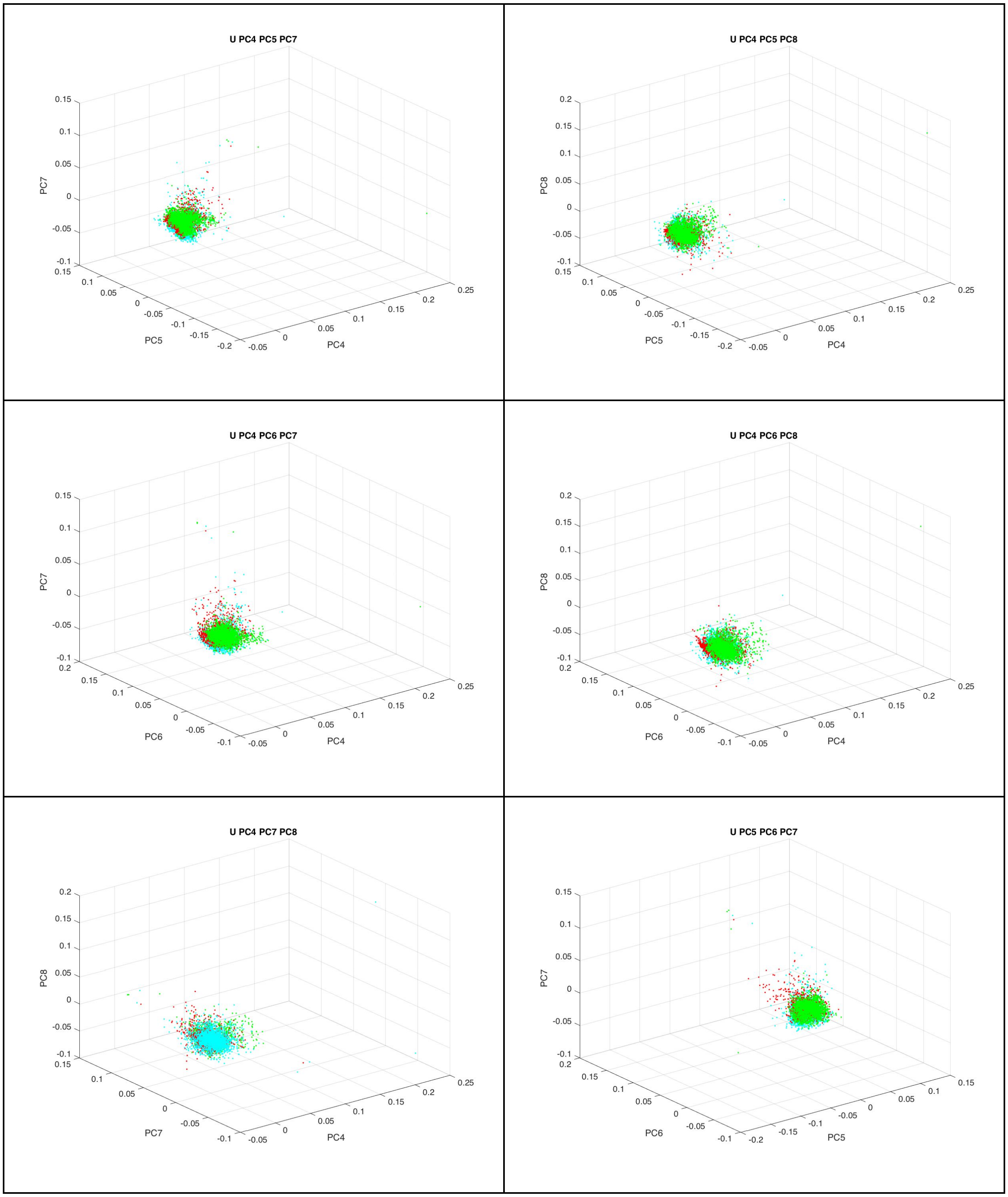

U PC3 PC4 PC6
PC6
PC4
PC3


U PC3 PC4 PC7
PC7
PC4
PC3

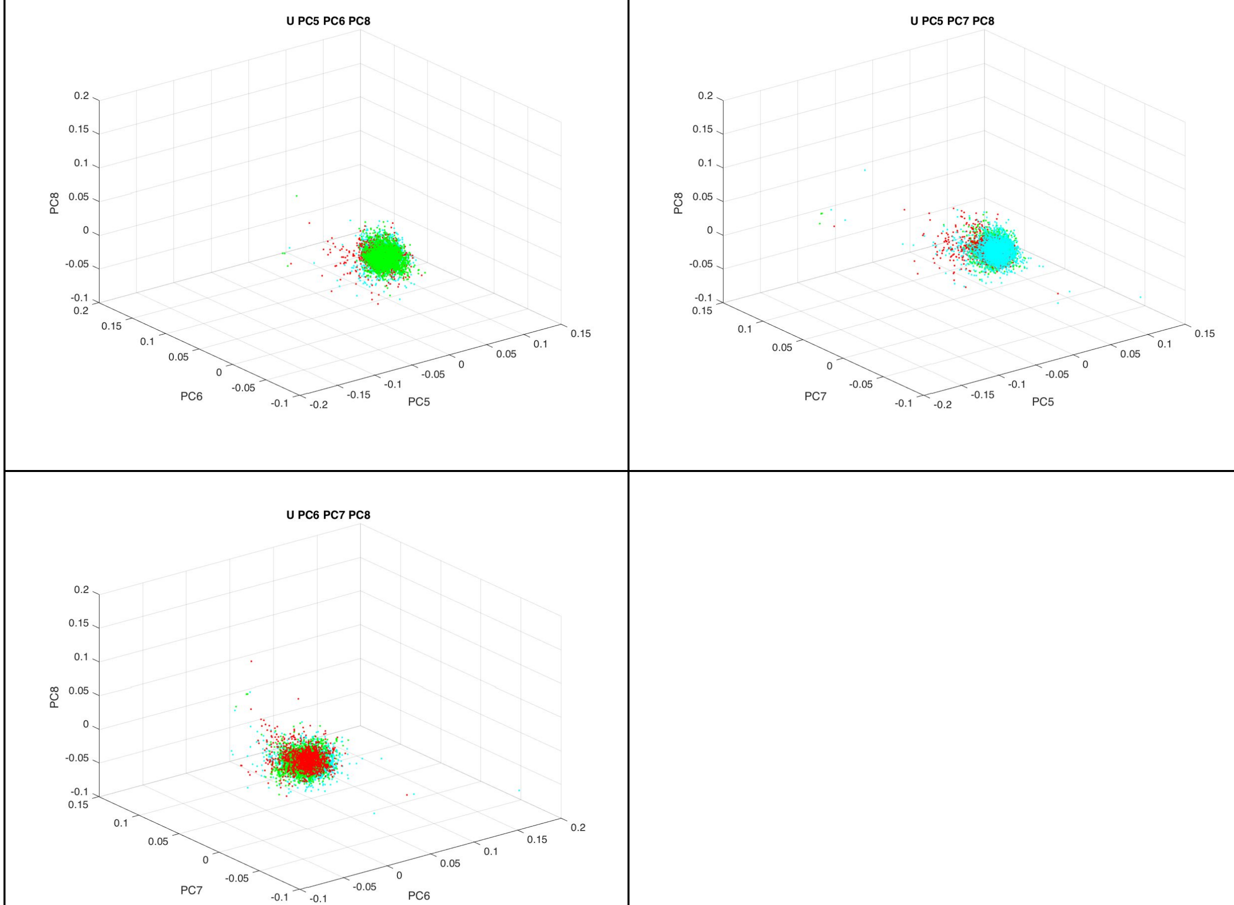

U PC3 PC4 PC8
PC8
PC4
PC3


U PC3 PC5 PC6
PC6
PC5
PC3

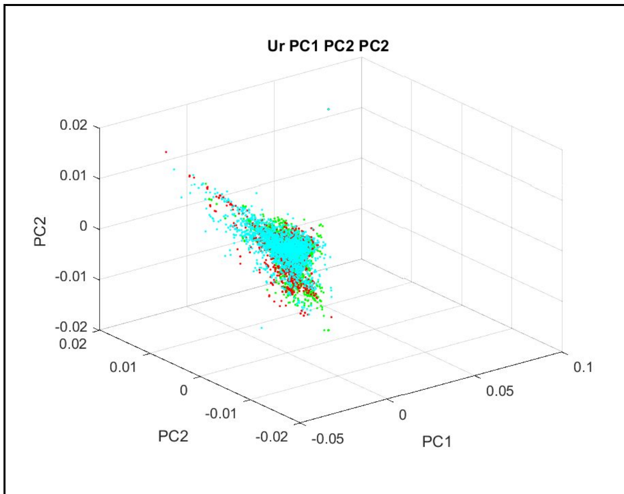

U PC3 PC5 PC7
PC7
0.15
0.1
0.05
0
-0.05
-0.1
PC5
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
PC3
-0.1
-0.05
0
0.05
0.1
U PC3 PC5 PC8
PC8
0.2
0.15
0.1
0.05
0
-0.05
-0.1
PC5
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
PC3
-0.1
-0.05
0
0.05
0.1
U PC3 PC6 PC7
PC7
0.15
0.1
0.05
0
-0.05
-0.1
PC6
0.2
0.15
0.1
0.05
0
-0.05
-0.1
PC3
-0.1
-0.05
0
0.05
0.1
U PC3 PC6 PC8
PC8
0.2
0.15
0.1
0.05
0
-0.05
-0.1
PC6
0.2
0.15
0.1
0.05
0
-0.05
-0.1
PC3
-0.1
-0.05
0
0.05
0.1
U PC3 PC7 PC8
PC8
0.2
0.15
0.1
0.05
0
-0.05
-0.1
PC7
0.15
0.1
0.05
0
-0.05
-0.1
PC3
-0.1
-0.05
0
0.05
0.1
U PC4 PC5 PC6
PC6
0.2
0.15
0.1
0.05
0
-0.05
-0.1
PC5
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
PC4
-0.05
0
0.05
0.1
0.15
0.2
0.25

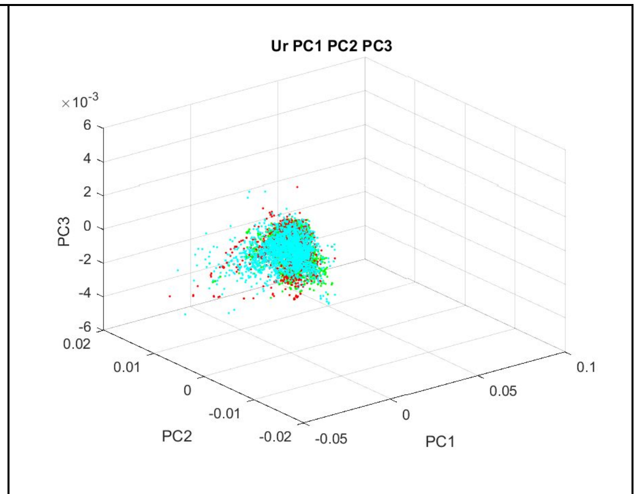


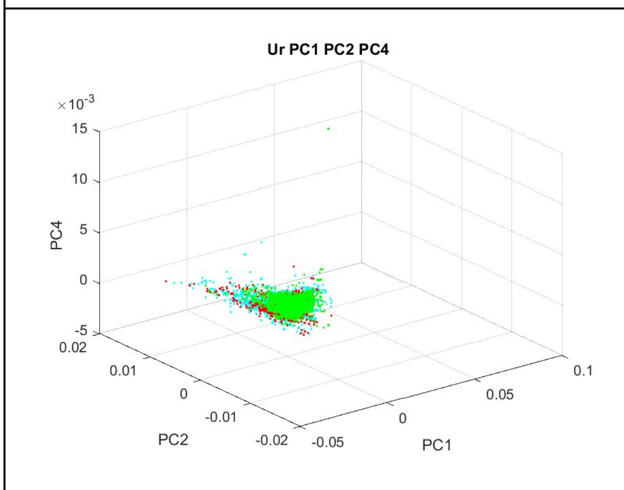
U PC4 PC5 PC7
PC7
PC5
PC4
U PC4 PC5 PC8
PC8
PC5
PC4
U PC4 PC6 PC7
PC7
PC6
PC4
U PC4 PC6 PC8
PC8
PC6
PC4
U PC4 PC7 PC8
PC8
PC7
PC4
U PC5 PC6 PC7
PC7
PC6
PC5

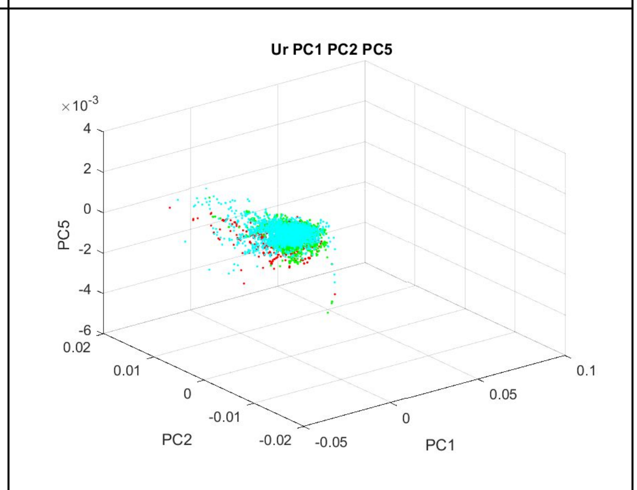

U PC5 PC6 PC8
PC8
PC6
PC5
U PC5 PC7 PC8
PC8
PC7
PC5
U PC6 PC7 PC8
PC8
PC7
PC6

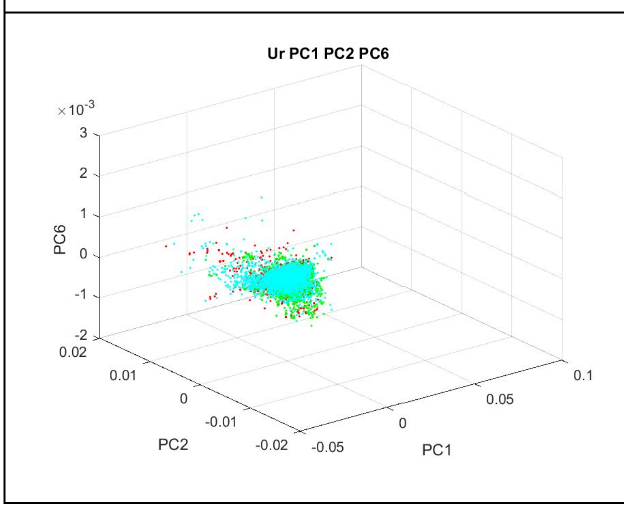
Ur Scatter Plots

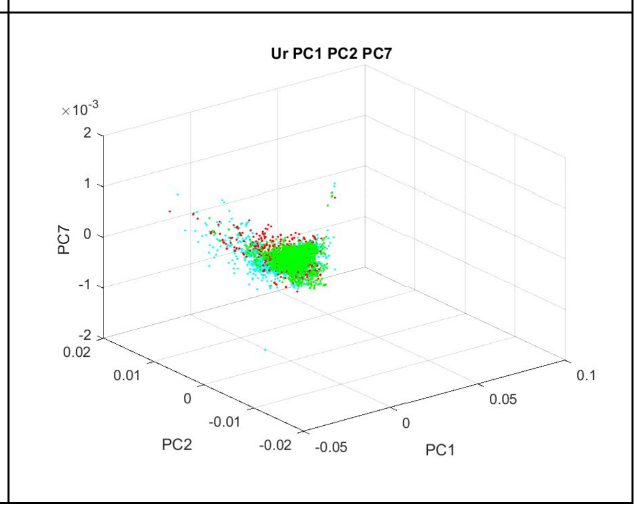


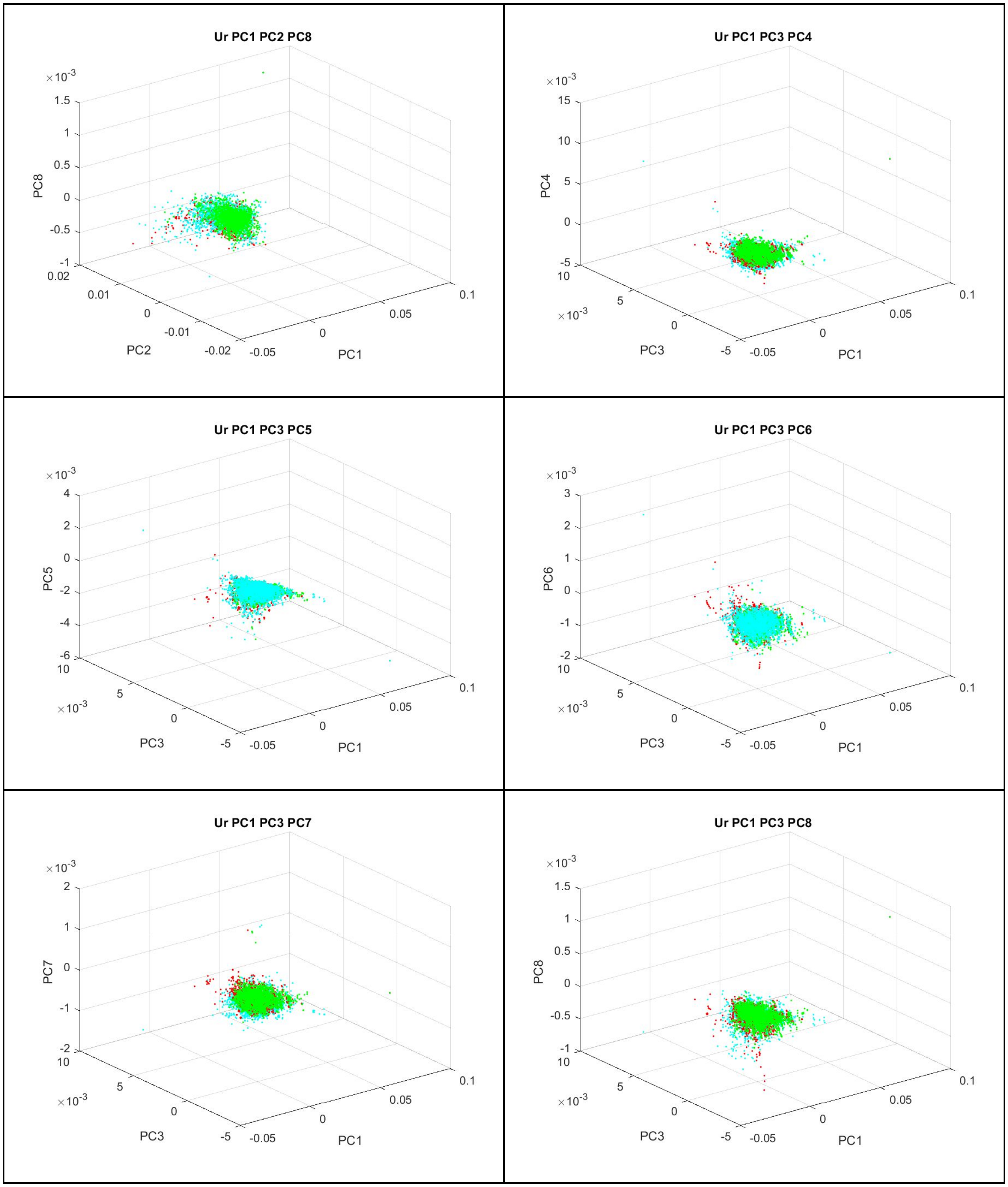


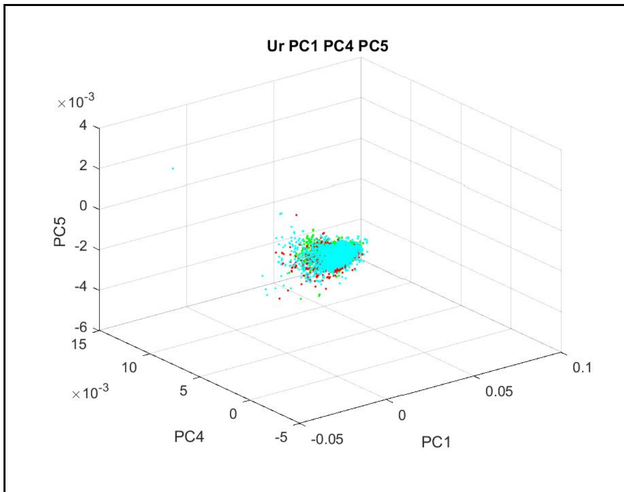


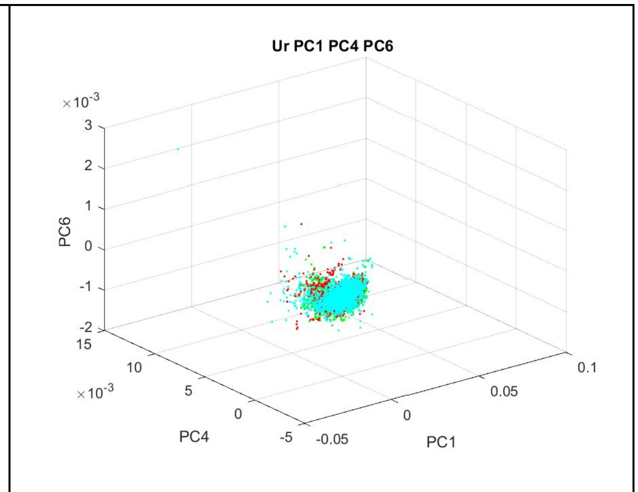


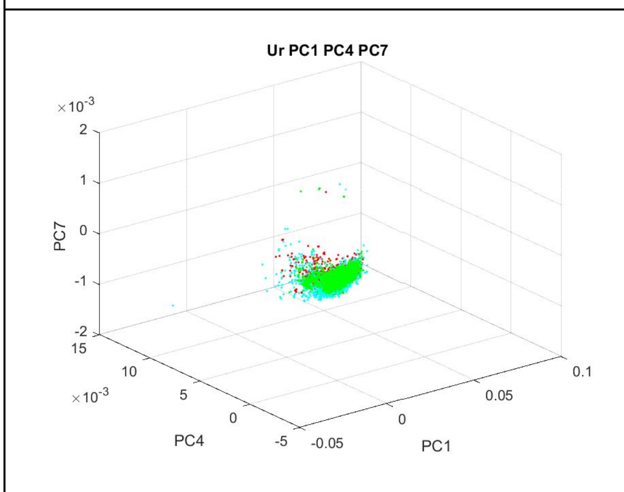


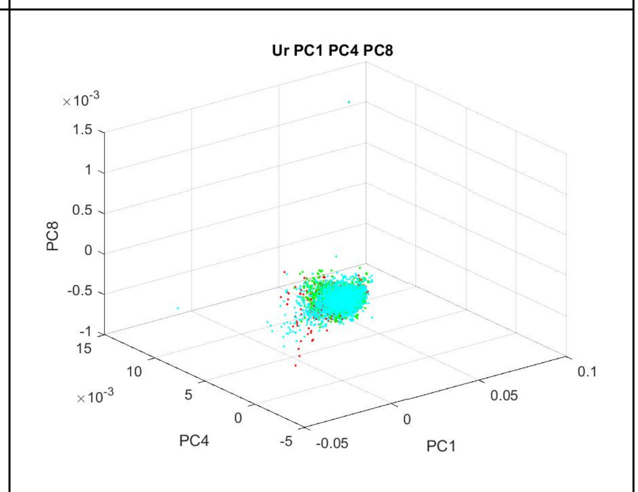


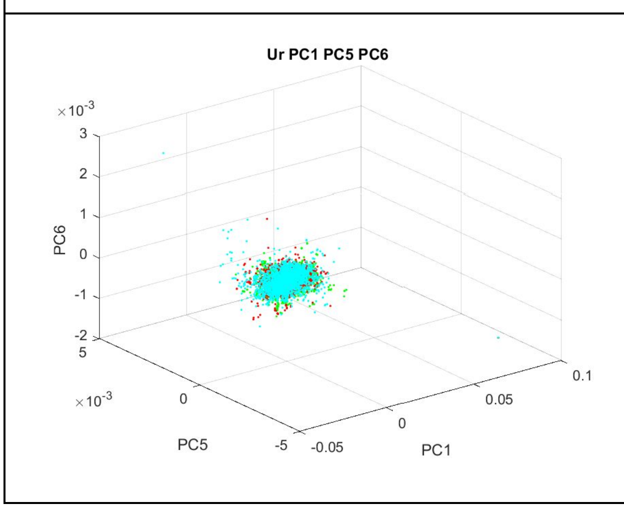

Ur PC1 PC2 PC8
PC8
PC2
PC1
Ur PC1 PC3 PC4
PC4
PC3
PC1
Ur PC1 PC3 PC5
PC5
PC3
PC1
Ur PC1 PC3 PC6
PC6
PC3
PC1
Ur PC1 PC3 PC7
PC7
PC3
PC1
Ur PC1 PC3 PC8
PC8
PC3
PC1

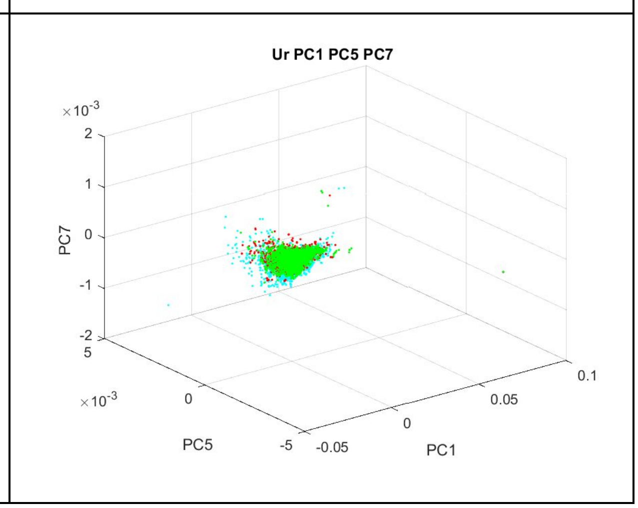

Ur PC1 PC4 PC5
×10^-3
4
2
0
-2
-4
-6
PC5
15
10
×10^-3
5
0
-5
PC4
-0.05
0
0.05
0.1
PC1

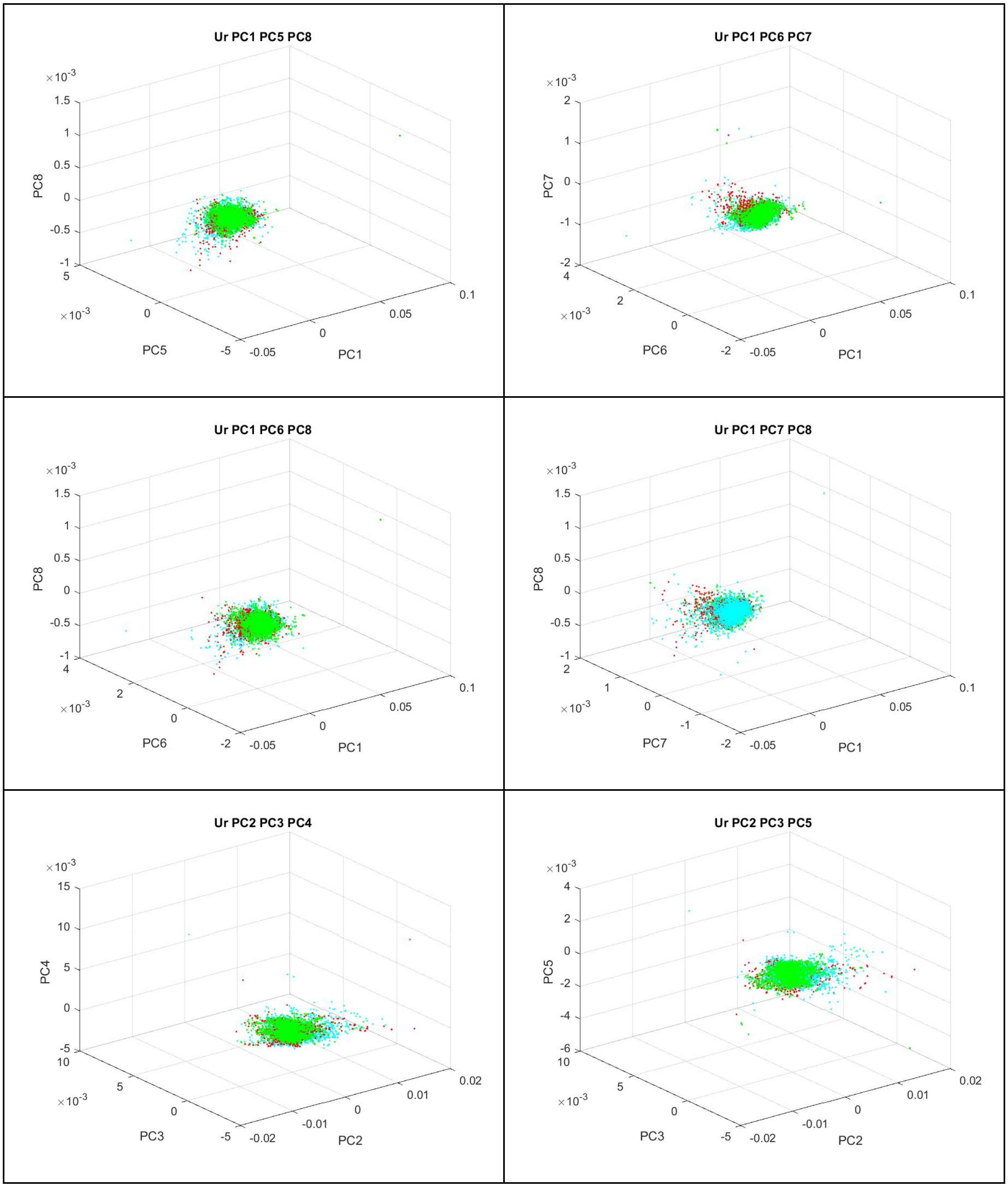

Ur PC1 PC4 PC6
×10^-3
3
2
1
0
-1
-2
PC6
15
10
×10^-3
5
0
-5
PC4
-0.05
0
0.05
0.1
PC1

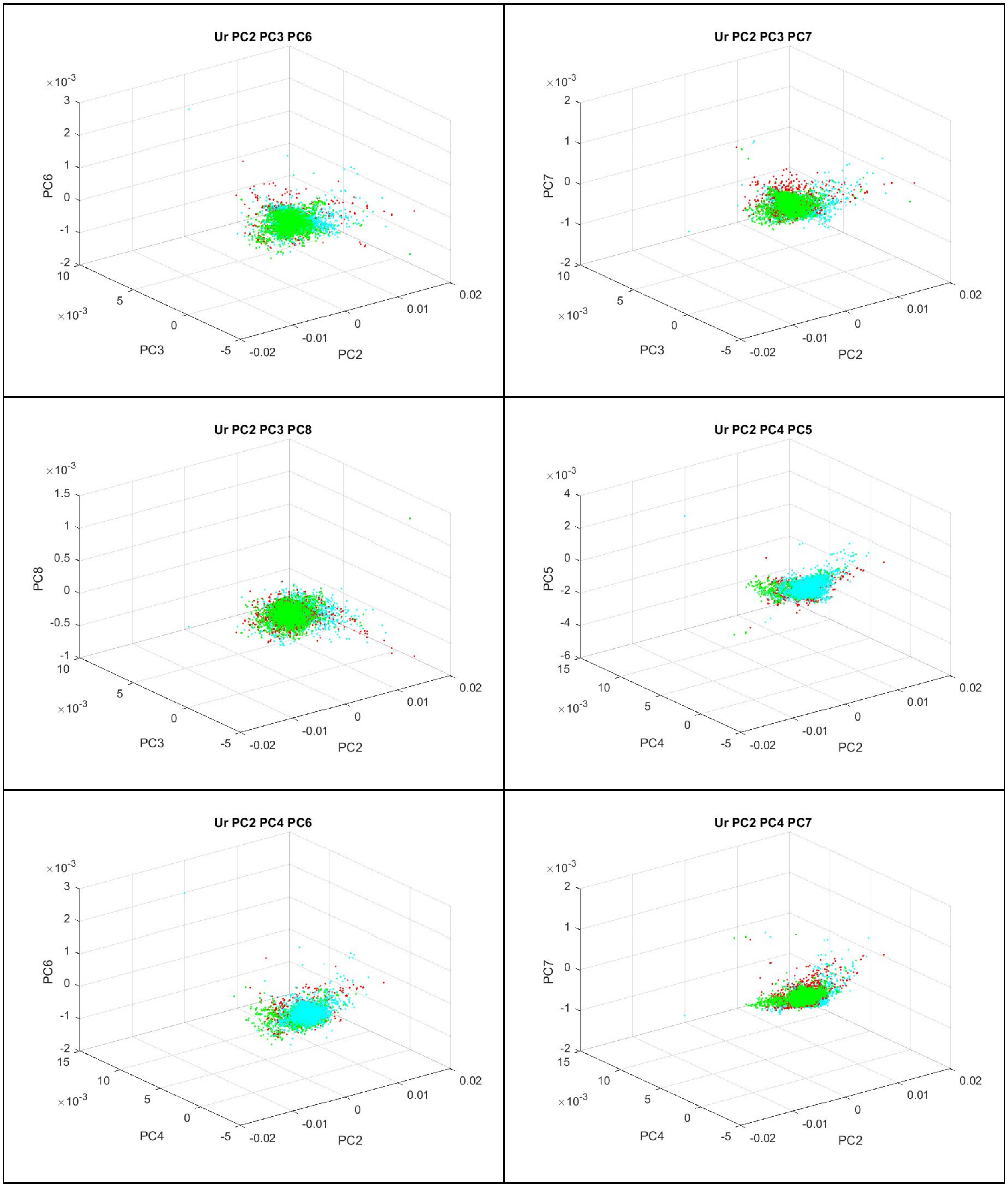

Ur PC1 PC4 PC7
×10^-3
2
1
0
-1
-2
PC7
15
10
×10^-3
5
0
-5
PC4
-0.05
0
0.05
0.1
PC1

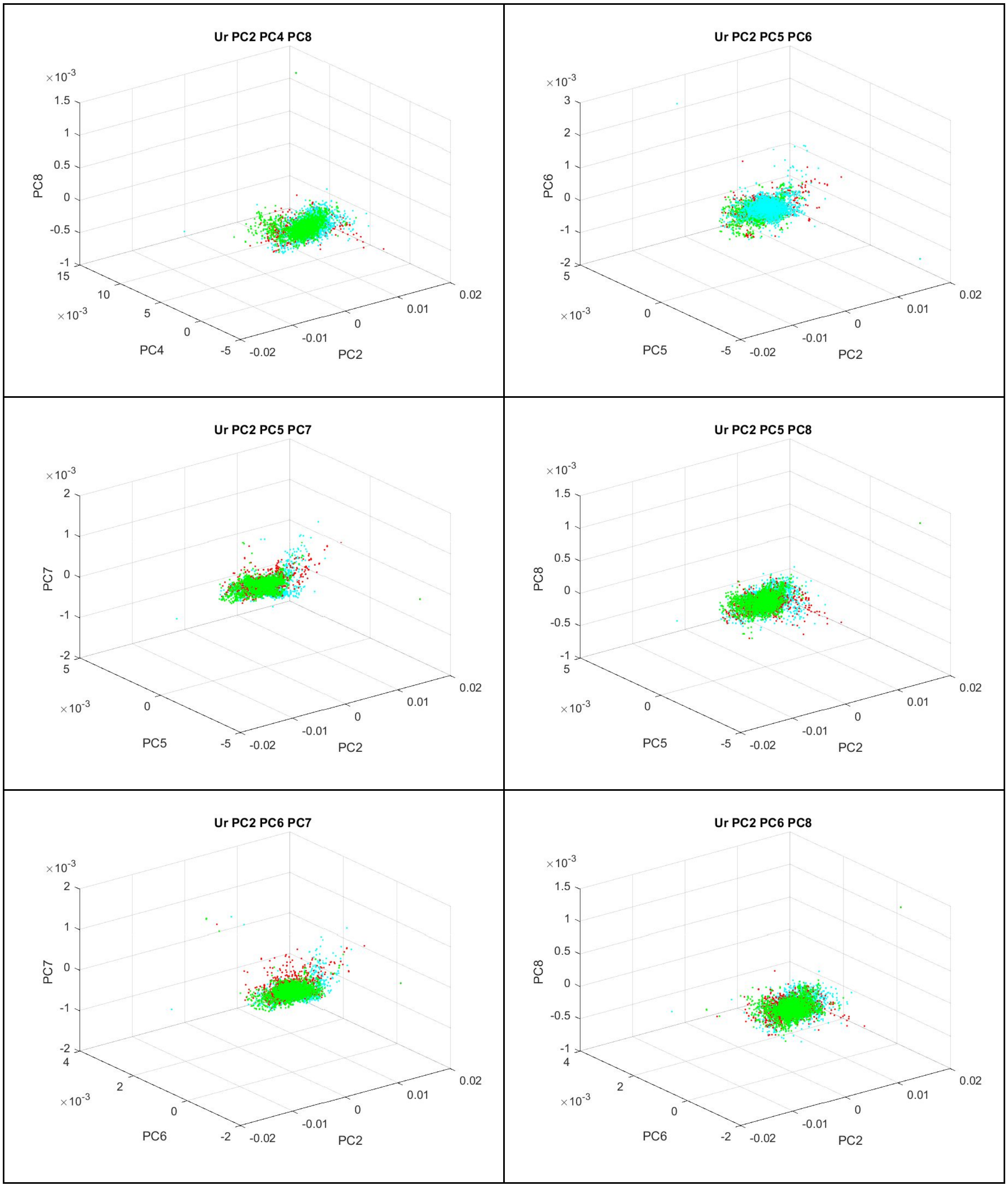

Ur PC1 PC4 PC8
×10^-3
1.5
1
0.5
0
-0.5
-1
PC8
15
10
×10^-3
5
0
-5
PC4
-0.05
0
0.05
0.1
PC1

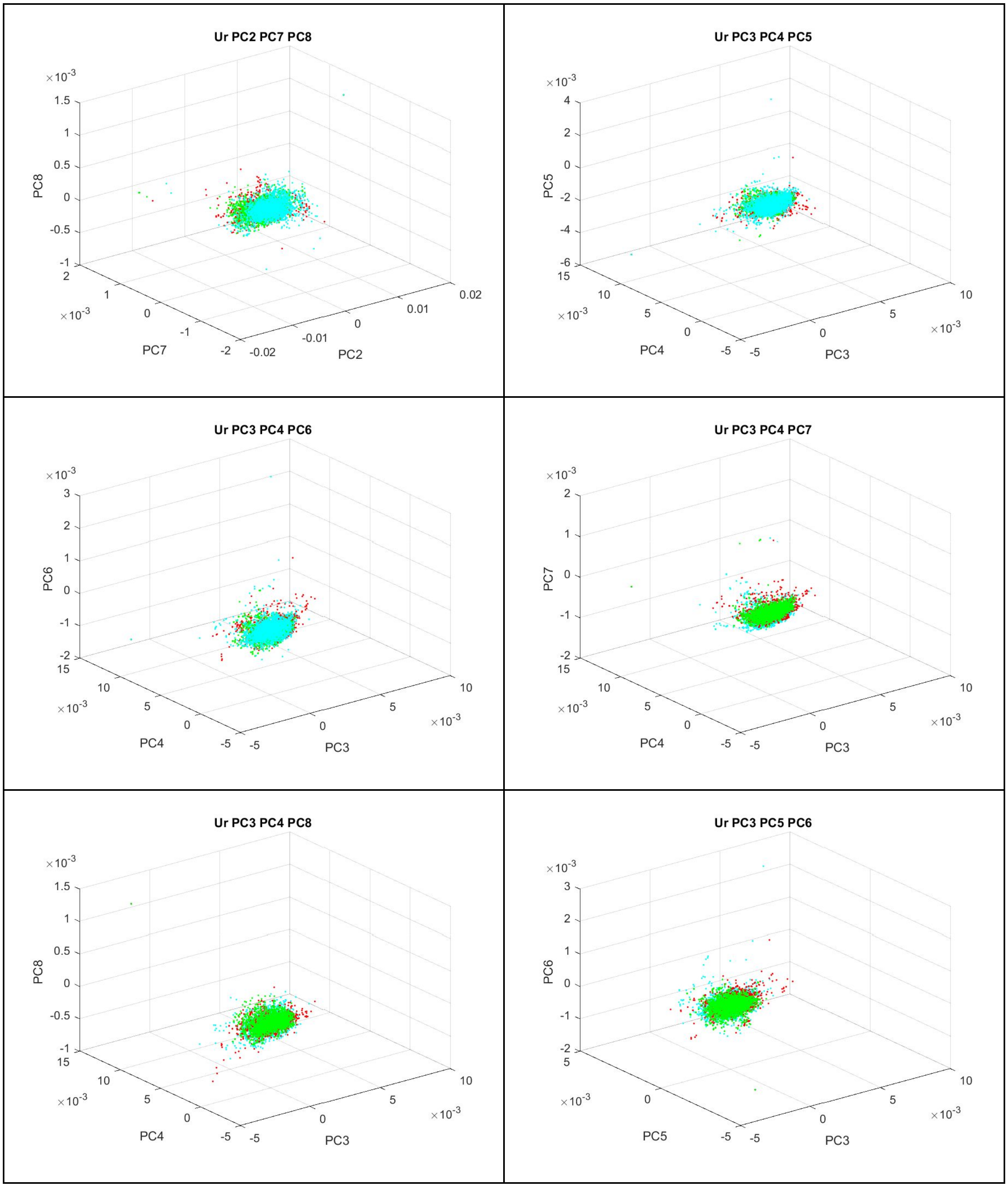

Ur PC1 PC5 PC6
×10^-3
3
2
1
0
-1
-2
PC6
5
×10^-3
0
-5
PC5
-0.05
0
0.05
0.1
PC1

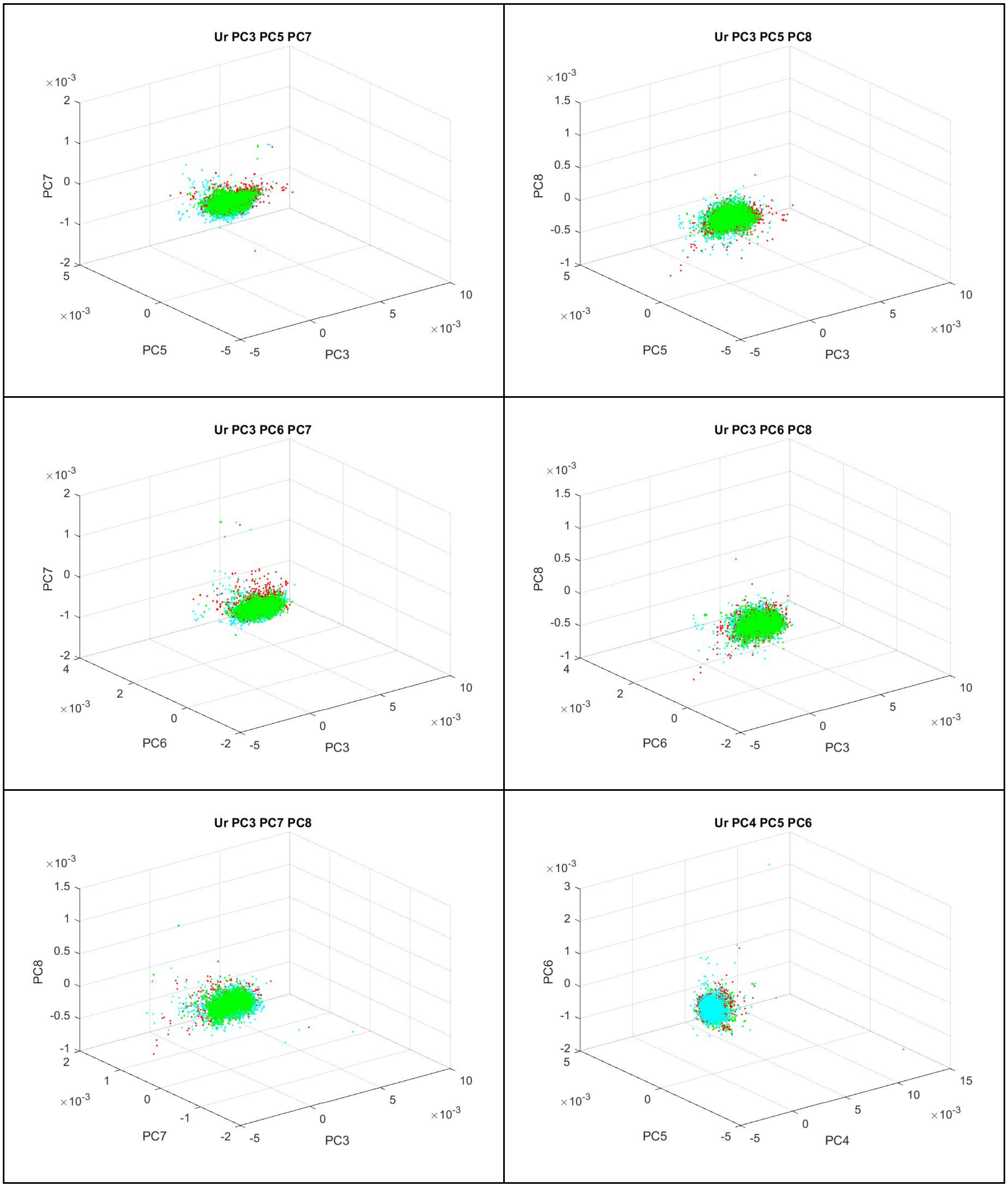

Ur PC1 PC5 PC7
×10^-3
2
1
0
-1
-2
PC7
5
×10^-3
0
-5
PC5
-0.05
0
0.05
0.1
PC1

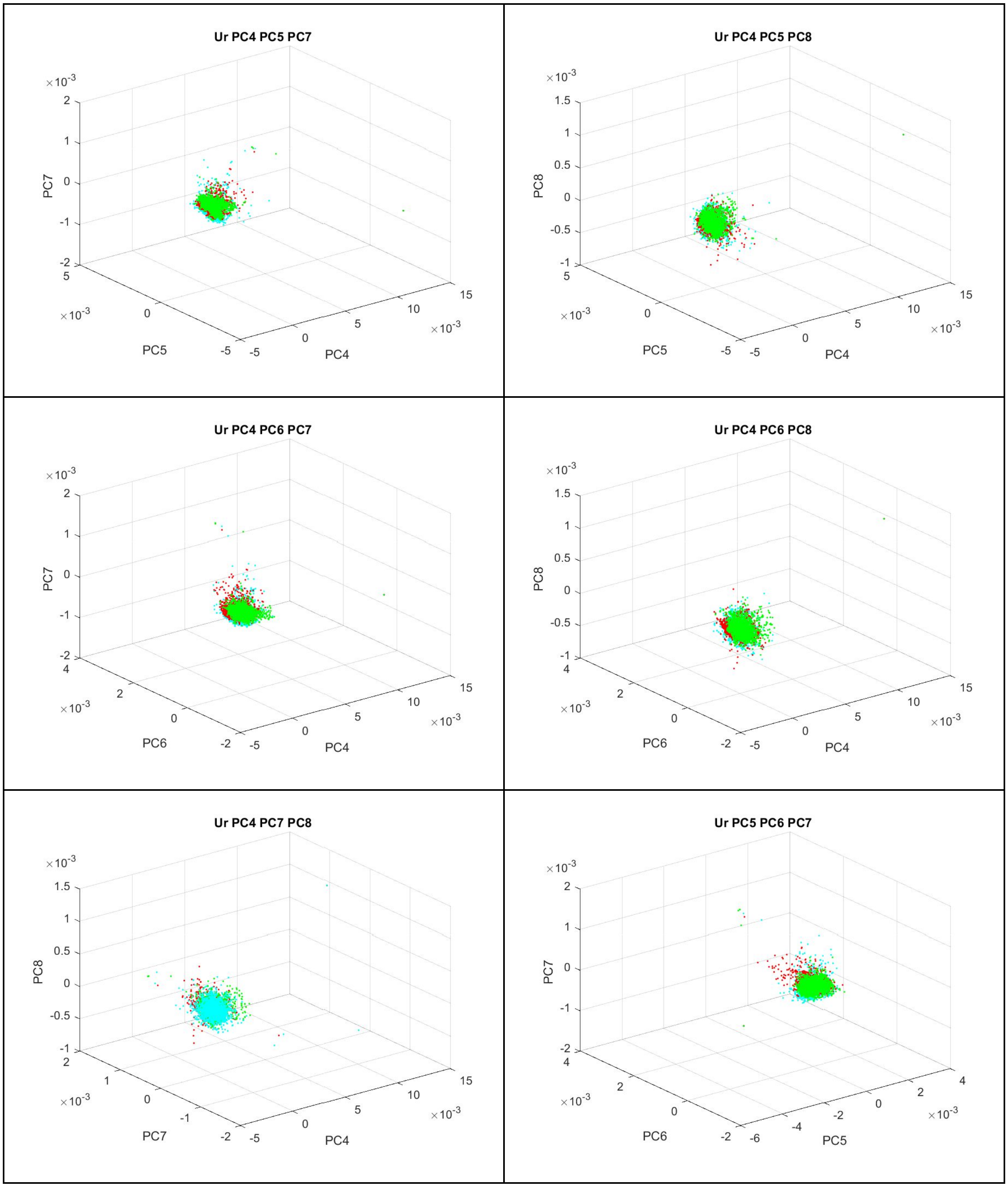

Ur PC1 PC5 PC8
PC8
PC5
PC1
Ur PC1 PC6 PC7
PC7
PC6
PC1
Ur PC1 PC6 PC8
PC8
PC6
PC1
Ur PC1 PC7 PC8
PC8
PC7
PC1
Ur PC2 PC3 PC4
PC4
PC3
PC2
Ur PC2 PC3 PC5
PC5
PC3
PC2

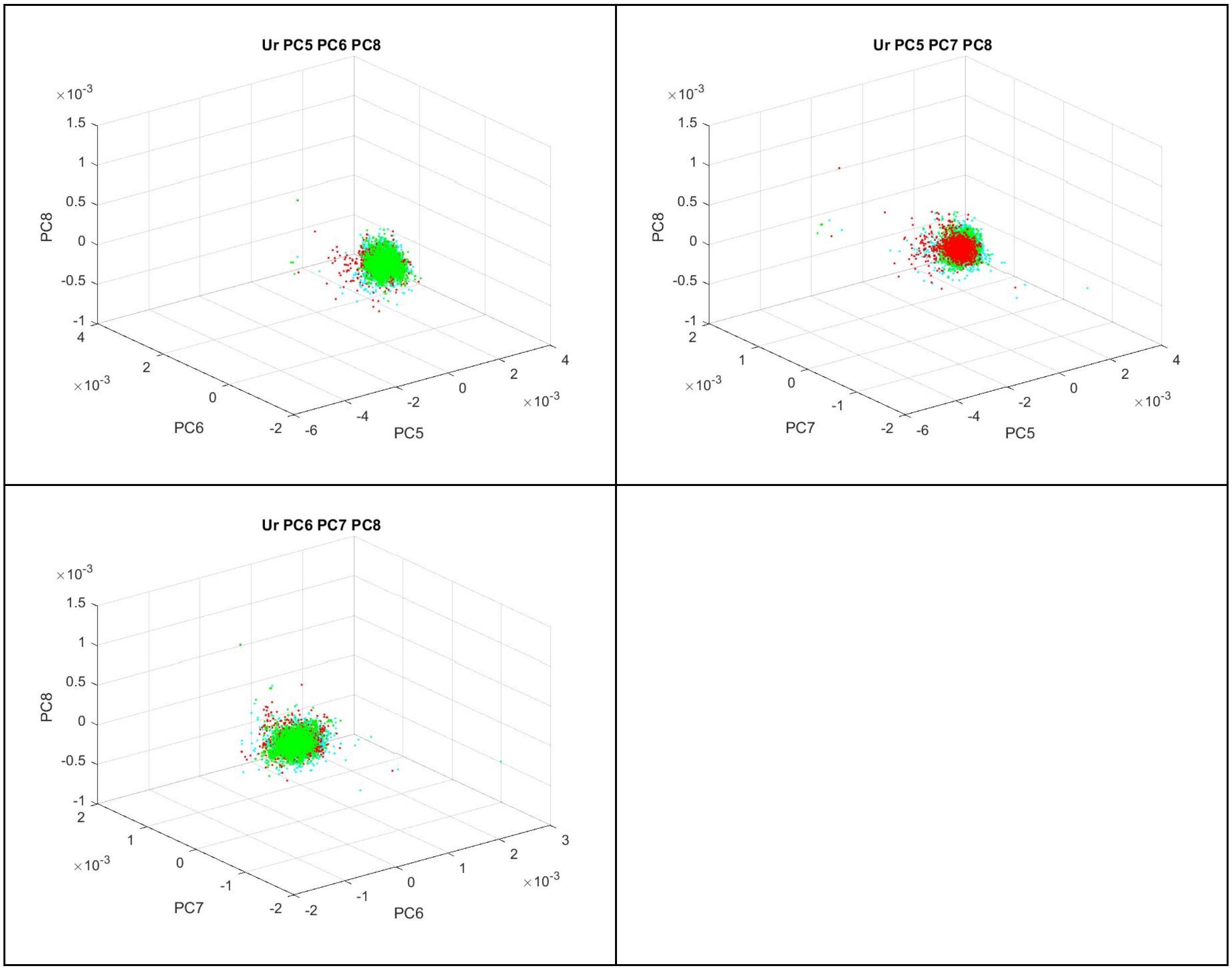

Ur PC2 PC3 PC6
PC6
PC3
PC2
Ur PC2 PC3 PC7
PC7
PC3
PC2
Ur PC2 PC3 PC8
PC8
PC3
PC2
Ur PC2 PC4 PC5
PC5
PC4
PC2
Ur PC2 PC4 PC6
PC6
PC4
PC2
Ur PC2 PC4 PC7
PC7
PC4
PC2


Ur PC2 PC4 PC8
PC8
PC4
PC2
Ur PC2 PC5 PC6
PC6
PC5
PC2
Ur PC2 PC5 PC7
PC7
PC5
PC2
Ur PC2 PC5 PC8
PC8
PC5
PC2
Ur PC2 PC6 PC7
PC7
PC6
PC2
Ur PC2 PC6 PC8
PC8
PC6
PC2


Ur PC2 PC7 PC8
PC8
PC7
PC2
Ur PC3 PC4 PC5
PC5
PC4
PC3
Ur PC3 PC4 PC6
PC6
PC4
PC3
Ur PC3 PC4 PC7
PC7
PC4
PC3
Ur PC3 PC4 PC8
PC8
PC4
PC3
Ur PC3 PC5 PC6
PC6
PC5
PC3

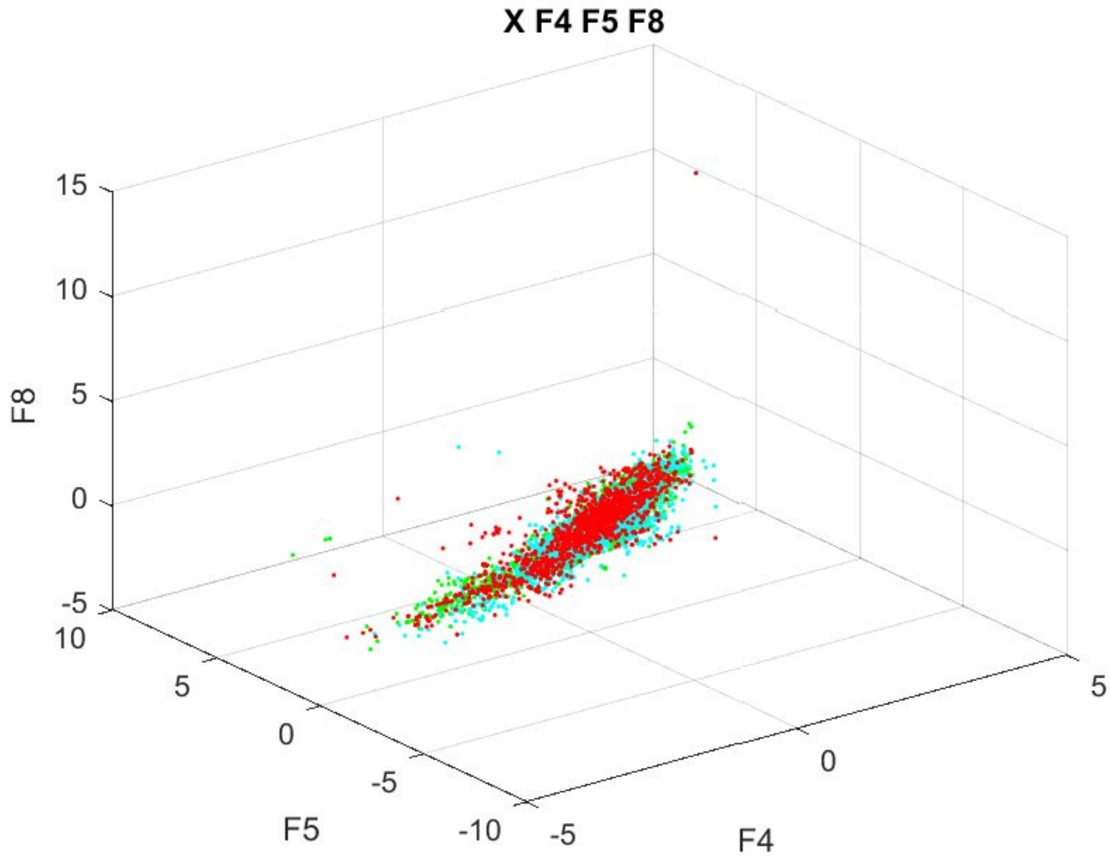

Ur PC3 PC5 PC7
PC7
PC5
PC3
Ur PC3 PC5 PC8
PC8
PC5
PC3
Ur PC3 PC6 PC7
PC7
PC6
PC3
Ur PC3 PC6 PC8
PC8
PC6
PC3
Ur PC3 PC7 PC8
PC8
PC7
PC3
Ur PC4 PC5 PC6
PC6
PC5
PC4


Ur PC4 PC5 PC7
PC7
PC5
PC4
Ur PC4 PC5 PC8
PC8
PC5
PC4
Ur PC4 PC6 PC7
PC7
PC6
PC4
Ur PC4 PC6 PC8
PC8
PC6
PC4
Ur PC4 PC7 PC8
PC8
PC7
PC4
Ur PC5 PC6 PC7
PC7
PC6
PC5


Ur PC5 PC6 PC8
PC8
PC6
PC5
×10-3
Ur PC5 PC7 PC8
PC8
PC7
PC5
×10-3
Ur PC6 PC7 PC8
PC8
PC7
PC6
×10-3

Analysis

We chose to keep three Features: curb weight (Original Feature 4), top speed (Original Feature 5), and height (Original Feature 8). These three Features are highly variable, highly correlated with the Features we eliminated, and capture the most clustering. The scatter plot below shows the clustering using the three kept Features. Although there is little clustering using these three Features, it is the best combination of Features for clustering out of the Feature data we have.



The decision to eliminate the other five Features is explained below.

Looking at our Loading vectors we can see a trend of the PC index 3, and 6 being highly correlated, (fuel tank/ length) with low variability. Of these three Features one Feature can be used to derive the same amount of information that is offered, for example fuel tank capacity is correlated to length of the vehicle, you can also derive how much of the fuel tank is being stretched out underneath the overall length of the vehicle.

I use stretched out because, given a vehicle typically by regulations, fuel tanks are designed to be longer in length than shallow in height, which explains the logic, if a vehicle is short, given the assumption that regulations make manufacturer design fuel tanks not shallow but long, we can say a percentage of the overall length of that vehicle is its fuel tank, given that percent and the length of the vehicle we can derive a fuel tank capacity.

Feature 1 has high correlation with all other Features. This is because with curb weight, and overall size (length width), we can derive that the heavier a vehicle is the less of a top speed it will have. Also with top speed we can derive Feature 1 (maximum horsepower) because with a higher top speed, a vehicle would require higher maximum horsepower. High maximum speed can also be derived from Features 4 and 5 which respectively are curb weight and top speed, because with more weight we can say the vehicle contains a larger engine which is also derived from curb weight.

Feature 4 has very high correlation with all other Features. You can derive curb weight from all of the other Features. Curb weight is highly influenced by the overall size of the vehicle, for instance. This constitutes width, length, and height, Features 6, 7 and 8. We can say the same for many of the other Features generally, for instance a max torque indicates a larger engine, which will necessarily be heavier, and influence curb weight to a great extent. It also adds next to no variability. We see this based on the scree and cumulative scree plots, where PC 8 is explaining almost none of the variability.

Feature 6 and 7 are both highly correlated with Feature 4. In Loading Vector 8, Feature 4 is the most important. This would generally indicate that Feature 4 is therefore a removable feature. However, since it has correlation with both Features 6 and 7, which we know to be length and width, respectively, we will keep Feature 4 to represent these features. This means as a complementary action we will remove Features 6 and 7.

Feature 2 is represented by Feature 4. In most of the Loading Vectors, the two are highly correlated. This makes sense, since the max torque of a car becomes higher by necessity as the total weight of the vehicle increases. Very large vehicles such as buses or semis require high torque in order to initially propel their high mass. Therefore it makes sense that Feature 2 and Feature 4 would be very correlated. Therefore, based on this and the elimination of Features 6 and 7 and their representation by Feature 4, we will eliminate Feature 2.

Team Assessment

Danny

Contributed to analysis of final document, help with brainstorming analysis, and assisted in partial code editing.

Robert

PCA - Worked on creating PCA code. Worked on converting octave PCA code to Matlab code. Created tables and figures. Contributed to analysis.

Kevin

Converted code from Homework 3 from Octave to Matlab. Concatenated previous iterations into one document. Created table of contents and document layout. Organized repository and managed inserting of plots. Created tabular plot schemes.

References

Dr. Erika Parsons

Central limit theorem
http://www.investopedia.com/terms/c/central_limit_theorem.asp

Terminal output in Octave
https://www.gnu.org/software/octave/doc/interpreter/Terminal-Output.html

How to read a CSV file in Matlab/Octave
https://www.mathworks.com/help/matlab/ref/csvread.html

Ultimatespecs website
https://www.ultimatespecs.com

Deleting rows in matrices in Octave
https://lists.gnu.org/archive/html/help-octave/2011-09/msg00179.html

Remove a row from a matrix in Octave
http://octave.1599824.n4.nabble.com/remove-row-from-matrix-td2969149.html

Manipulating plot windows (multiple plots per window - hold on)
https://www.gnu.org/software/octave/doc/interpreter/Manipulation-of-Plot-Windows.html

2D plots in Octave
https://www.gnu.org/software/octave/doc/interpreter/Two_002dDimensional-Plots.html

BHP disambiguation (It means brake horsepower)
https://en.wikipedia.org/wiki/Bhp

Printing and saving plots in Octave
https://www.gnu.org/software/octave/doc/interpreter/Printing-and-Saving-Plots.html

genvarname in Octave for generating unique file names
https://www.mathworks.com/help/matlab/ref/genvarname.html

Plotting multiple lines in one figure
https://www.mathworks.com/examples/matlab/mw/graphics-ex99259345-plot-multiple-lines

Plotting multiple plots on one graph

https://www.mathworks.com/matlabcentral/answers/23591-examples-of-multiple-plots-on-one-graph

Hold for Octave
https://octave.sourceforge.io/octave/function/hold.html

Two functions on one plot in Octave
http://octave.1599824.n4.nabble.com/Same-plot-for-two-functions-td1631271.html

Variables in Octave
https://www.gnu.org/software/octave/doc/v4.0.0/Variables.html

Plotting multiple curves in Octave
http://web.cecs.pdx.edu/~gerry/MATLAB/plotting/multipleCurves.html

Distribution calculation in Octave
https://www.gnu.org/software/octave/doc/interpreter/Distributions.html