Machine Learning and Data Mining II

LABWORK 1: PCA REPORT

Groupwork:

Nguyễn Thị Quỳnh Anh **BI11-029**
Vũ Yến Linh **BI11-152**

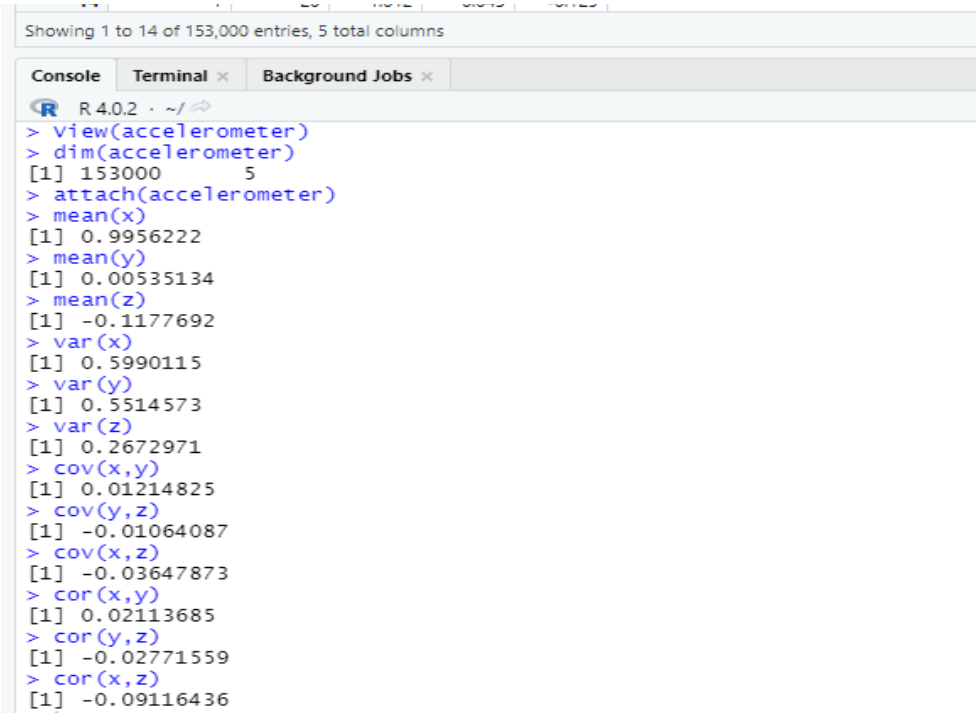
I. Study the dataset

- Choose 2 datasets from UCI Machine Learning Repository with more than 3 dimensions.(http://archive.ics.uci.edu/ml/)
  1.Accelerometer: http://archive.ics.uci.edu/ml/datasets/Accelerometer
  2.Synchronous Machine Data Set:
  http://archive.ics.uci.edu/ml/datasets/Synchronous+Machine+Data+Set

- For each dataset, determine which feature is discrete or continuous? Is it quantitative or qualitative? Explain.
  1.continuous and qualitative because this dataset was generated for use on 'Prediction of Motor Failure Time Using An Artificial Neural Network' project (DOI: 10.3390/s19194342), a cooler fan with weights on its blades was used to generate vibrations.
  2.discrete and quantitative because Synchronous machine data were obtained in real time from the experimental operating environment.

- Calculate mean, variance, covariance, correlation of the selected datasets.
  1.Accelerometer

  - mean:



- variance, covariance, correlation:

```
> var(accelerometer)
              wconfid        pctid             x             y             z
wconfid  0.6666710240   0.00000000 -0.003832724  0.003256623  0.0009406728
pctid    0.0000000000 600.00392159  0.040351898  0.052418643  0.1135272453
x       -0.0038327244   0.04035190  0.599011478  0.012148250 -0.0364787347
y        0.0032566226   0.05241864  0.012148250  0.551457282 -0.0106408723
z        0.0009406728   0.11352725 -0.036478735 -0.010640872  0.2672970885
> cov(accelerometer)
              wconfid        pctid             x             y             z
wconfid  0.6666710240   0.00000000 -0.003832724  0.003256623  0.0009406728
pctid    0.0000000000 600.00392159  0.040351898  0.052418643  0.1135272453
x       -0.0038327244   0.04035190  0.599011478  0.012148250 -0.0364787347
y        0.0032566226   0.05241864  0.012148250  0.551457282 -0.0106408723
z        0.0009406728   0.11352725 -0.036478735 -0.010640872  0.2672970885
> cor(accelerometer)
              wconfid       pctid             x             y             z
wconfid  1.000000000 0.000000000 -0.006065048  0.005371007  0.002228362
pctid    0.000000000 1.000000000  0.002128479  0.002881727  0.008964497
x       -0.006065048 0.002128479  1.000000000  0.021136847 -0.091164358
y        0.005371007 0.002881727  0.021136847  1.000000000 -0.027715593
z        0.002228362 0.008964497 -0.091164358 -0.027715593  1.000000000
```

2.Synchronous Machine Data Set Data Set

- mean

```
> View(synchronous.machine)
> dim(synchronous.machine)
[1] 557   5
> attach(synchronous.machine)
> mean(Iy)
[1] 4.49982
> mean(PF)
[1] 0.8252962
> mean(e)
[1] 0.1747038
> mean(dIf)
[1] 0.3506589
> mean(If)
[1] 1.530659
```

- variance, covariance, correlation:

```
> var(synchronous.machine)
             Iy          PF           e          dIf          If
Iy   0.80285968 -0.00387135  0.00387135  0.06875246  0.06875246
PF  -0.00387135  0.01080050 -0.01080050 -0.01615725 -0.01615725
e    0.00387135 -0.01080050  0.01080050  0.01615725  0.01615725
dIf  0.06875246 -0.01615725  0.01615725  0.03260405  0.03260405
If   0.06875246 -0.01615725  0.01615725  0.03260405  0.03260405
> cov(synchronous.machine)
             Iy          PF           e          dIf          If
Iy   0.80285968 -0.00387135  0.00387135  0.06875246  0.06875246
PF  -0.00387135  0.01080050 -0.01080050 -0.01615725 -0.01615725
e    0.00387135 -0.01080050  0.01080050  0.01615725  0.01615725
dIf  0.06875246 -0.01615725  0.01615725  0.03260405  0.03260405
If   0.06875246 -0.01615725  0.01615725  0.03260405  0.03260405
> cor(synchronous.machine)
             Iy          PF           e         dIf          If
Iy   1.00000000 -0.04157389  0.04157389  0.4249449  0.4249449
PF  -0.04157389  1.00000000 -1.00000000 -0.8610135 -0.8610135
e    0.04157389 -1.00000000  1.00000000  0.8610135  0.8610135
dIf  0.42494491 -0.86101347  0.86101347  1.0000000  1.0000000
If   0.42494491 -0.86101347  0.86101347  1.0000000  1.0000000
```

- Find the most correlated couple of features of each dataset. Comment on the results.

  1.Accelerometer

  cor(x,z): -0.09116436

  -> Strong negative correlation

  2.Synchronous Machine Data Set Data Set

  cor(dIF,IF): 1

  -> Strong positive correlation