Chuong Vu

COMP 5800

Homework 3

Prof. Byung Kim

PCA Analysis AAs

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Vol.** | **Bulk.** | **Polarity** | **pI** | **Hyd. 1** | **Hyd. 2** | **Surf. area** | **Fract. Area** |
| Alanine | Ala | A | 67 | 11.50 | 0.00 | 6.00 | 1.8 | 1.6 | 113 | 0.74 |
| Arginine | Arg | R | 148 | 14.28 | 52.00 | 10.76 | -4.5 | -12.3 | 241 | 0.64 |
| Asparagine | Asn | N | 96 | 12.28 | 3.38 | 5.41 | -3.5 | -4.8 | 158 | 0.63 |
| Aspartic acid | Asp | D | 91 | 11.68 | 49.70 | 2.77 | -3.5 | -9.2 | 151 | 0.62 |
| Cysteine | Cys | C | 86 | 13.46 | 1.48 | 5.05 | 2.5 | 2.0 | 140 | 0.91 |
| Glutamine | Gln | Q | 114 | 14.45 | 3.53 | 5.65 | -3.5 | -4.1 | 189 | 0.62 |
| Glutamic acid | Glu | E | 109 | 13.57 | 49.90 | 3.22 | -3.5 | -8.2 | 183 | 0.62 |
| Glycine | Gly | G | 48 | 3.40 | 0.00 | 5.97 | -0.4 | 1.0 | 85 | 0.72 |
| Histidine | His | H | 118 | 13.69 | 51.60 | 7.59 | -3.2 | -3.0 | 194 | 0.78 |
| Isoleucine | Ile | I | 124 | 21.40 | 0.13 | 6.02 | 4.5 | 3.1 | 182 | 0.88 |
| Leucine | Leu | L | 124 | 21.40 | 0.13 | 5.98 | 3.8 | 2.8 | 180 | 0.85 |
| Lysine | Lys | K | 135 | 15.71 | 49.50 | 9.74 | -3.9 | -8.8 | 211 | 0.52 |
| Methionine | Met | M | 124 | 16.25 | 1.43 | 5.74 | 1.9 | 3.4 | 204 | 0.85 |
| Phenylalanine | Phe | F | 135 | 19.80 | 0.35 | 5.48 | 2.8 | 3.7 | 218 | 0.88 |
| Proline | Pro | P | 90 | 17.43 | 1.58 | 6.30 | -1.6 | -0.2 | 143 | 0.64 |
| Serine | Ser | S | 73 | 9.47 | 1.67 | 5.68 | -0.8 | 0.6 | 122 | 0.66 |
| Threonine | Thr | T | 93 | 15.77 | 1.66 | 5.66 | -0.7 | 1.2 | 146 | 0.70 |
| Tryptophan | Trp | W | 163 | 21.67 | 2.10 | 5.89 | -0.9 | 1.9 | 259 | 0.85 |
| Tyrosine | Tyr | Y | 141 | 18.03 | 1.61 | 5.66 | -1.3 | -0.7 | 229 | 0.76 |
| Valine | Val | V | 105 | 21.57 | 0.13 | 5.96 | 4.2 | 2.6 | 160 | 0.86 |
| **Mean** |  |  | **109** | **15.35** | **13.59** | **6.03** | **-0.5** | **-1.4** | **175** | **0.74** |
| **Std. dev.** |  |  | **28** | **4.53** | **21.36** | **1.72** | **2.9** | **4.8** | **44** | **0.11** |

The data I use to do PCA Analysis is from the Physico-chemical properties of the amino acids table.

Then I make the visual clustering based on volume and pI

|  |  |  |
| --- | --- | --- |
| Label | **Vol.(Y)** | **Pl (X)** |
| A | 67 | 6.00 |
| R | 148 | 10.76 |
| N | 96 | 5.41 |
| D | 91 | 2.77 |
| C | 86 | 5.05 |
| Q | 114 | 5.65 |
| E | 109 | 3.22 |
| G | 48 | 5.97 |
| H | 118 | 7.59 |
| I | 124 | 6.02 |
| L | 124 | 5.98 |
| K | 135 | 9.74 |
| M | 124 | 5.74 |
| F | 135 | 5.48 |
| P | 90 | 6.30 |
| S | 73 | 5.68 |
| T | 93 | 5.66 |
| W | 163 | 5.89 |
| Y | 141 | 5.66 |
| V | 105 | 5.96 |

That is how we make a graph based on two properties. What if we want to make a 2D graph from all eight properties? This can be cone with principal component analysis (PCA)

The first step in PCA is begin with the data in the form of an N x P matrix where N = 20 amino acids and P = 8 properties.

Since I was allowed to use any library that perform PCA, I choice R programming language to do this PCA because it already have the build-in PCA library.

In R, first I need to load the data file into object.

> data <- read.csv(file="data.csv", header=TRUE)

> data

X Vol. Bulk. Pol. pI Hyd.1 Hyd.2 SA FA

1 A 67 11.50 0.00 6.00 1.8 1.6 113 0.74

2 R 148 14.28 52.00 10.76 -4.5 -12.3 241 0.64

3 N 96 12.28 3.38 5.41 -3.5 -4.8 158 0.63

4 D 91 11.68 49.70 2.77 -3.5 -9.2 151 0.62

5 C 86 13.46 1.48 5.05 2.5 2.0 140 0.91

6 Q 114 14.45 3.53 5.65 -3.5 -4.1 189 0.62

7 E 109 13.57 49.90 3.22 -3.5 -8.2 183 0.62

8 G 48 3.40 0.00 5.97 -0.4 1.0 85 0.72

9 H 118 13.69 51.60 7.59 -3.2 -3.0 194 0.78

10 I 124 21.40 0.13 6.02 4.5 3.1 182 0.88

11 L 124 21.40 0.13 5.98 3.8 2.8 180 0.85

12 K 135 15.71 49.50 9.74 -3.9 -8.8 211 0.52

13 M 124 16.25 1.43 5.74 1.9 3.4 204 0.85

14 F 135 19.80 0.35 5.48 2.8 3.7 218 0.88

15 P 90 17.43 1.58 6.30 -1.6 -0.2 143 0.64

16 S 73 9.47 1.67 5.68 -0.8 0.6 122 0.66

17 T 93 15.77 1.66 5.66 -0.7 1.2 146 0.70

18 W 163 21.67 2.10 5.89 -0.9 1.9 259 0.85

19 Y 141 18.03 1.61 5.66 -1.3 -0.7 229 0.76

20 V 105 21.57 0.13 5.96 4.2 2.6 160 0.86

So I make another data object that replace the row name from X column

> d <- data[,-1] # copy all columns expect the first one

> rownames(d) <- data[,1] # replace the row name from the first column of original data

> d

Vol. Bulk. Pol. pI Hyd.1 Hyd.2 SA FA

A 67 11.50 0.00 6.00 1.8 1.6 113 0.74

R 148 14.28 52.00 10.76 -4.5 -12.3 241 0.64

N 96 12.28 3.38 5.41 -3.5 -4.8 158 0.63

D 91 11.68 49.70 2.77 -3.5 -9.2 151 0.62

C 86 13.46 1.48 5.05 2.5 2.0 140 0.91

Q 114 14.45 3.53 5.65 -3.5 -4.1 189 0.62

E 109 13.57 49.90 3.22 -3.5 -8.2 183 0.62

G 48 3.40 0.00 5.97 -0.4 1.0 85 0.72

H 118 13.69 51.60 7.59 -3.2 -3.0 194 0.78

I 124 21.40 0.13 6.02 4.5 3.1 182 0.88

L 124 21.40 0.13 5.98 3.8 2.8 180 0.85

K 135 15.71 49.50 9.74 -3.9 -8.8 211 0.52

M 124 16.25 1.43 5.74 1.9 3.4 204 0.85

F 135 19.80 0.35 5.48 2.8 3.7 218 0.88

P 90 17.43 1.58 6.30 -1.6 -0.2 143 0.64

S 73 9.47 1.67 5.68 -0.8 0.6 122 0.66

T 93 15.77 1.66 5.66 -0.7 1.2 146 0.70

W 163 21.67 2.10 5.89 -0.9 1.9 259 0.85

Y 141 18.03 1.61 5.66 -1.3 -0.7 229 0.76

V 105 21.57 0.13 5.96 4.2 2.6 160 0.86

Now we compute the mean and std. dev from the table

#Get mean value

> meandata <- ddply(data, .(),numcolwise(mean))

> meandata <- meandata[2:9]

> rownames(meandata)[1] <- "Mean"

> meandata[,-1] <- round(meandata[,-1], 2)

> meandata

Vol. Bulk. Pol. pI Hyd.1 Hyd.2 SA FA

Mean 109.2 15.34 13.59 6.03 -0.49 -1.37 175.4 0.74

#Get Std.Dev

> sddata <- ddply(data, .(), numcolwise(sd))

> sddata[,-1] <- round(sddata[,-1], 2)

> sddata <- sddata[2:9]

> rownames(sddata)[1] <- "Std.Dev"

> sddata

Vol. Bulk. Pol. pI Hyd.1 Hyd.2 SA FA

Std.Dev 29.05 4.65 21.91 1.77 2.99 4.89 44.69 0.12

#Get Variance value

> vardata <- ddply(data, .(), numcolwise(var))

> vardata[,-1] <- round(vardata[,-1], 2)

> vardata <- vardata[2:9]

> rownames(vardata)[1] <- "Variance"

> vardata

Vol. Bulk. Pol. pI Hyd.1 Hyd.2 SA FA

Variance 843.64 21.62 480.23 3.13 8.92 23.94 1996.78 0.01

Now, by using this commands, we can get the correlation coefficient matrix.

> r = cor(d) # this generate the correlation coefficient matrix

> r <- round(r, 2) # round to 2 decimal value

> r

Vol. Bulk. Pol. pI Hyd.1 Hyd.2 SA FA

Vol. 1.00 0.73 0.24 0.37 -0.08 -0.16 0.99 0.18

Bulk. 0.73 1.00 -0.20 0.08 0.44 0.32 0.64 0.49

Pol. 0.24 -0.20 1.00 0.27 -0.67 -0.85 0.29 -0.53

pI 0.37 0.08 0.27 1.00 -0.20 -0.27 0.36 -0.18

Hyd.1 -0.08 0.44 -0.67 -0.20 1.00 0.85 -0.18 0.84

Hyd.2 -0.16 0.32 -0.85 -0.27 0.85 1.00 -0.23 0.79

SA 0.99 0.64 0.29 0.36 -0.18 -0.23 1.00 0.12

FA 0.18 0.49 -0.53 -0.18 0.84 0.79 0.12 1.00

Now we pass the value in d to do Principal component analysis

# Principal component analysis, copy data to pca library

> pca1 <- princomp(d, scores=TRUE, cor=TRUE)

> summary(pca1)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

Standard deviation 1.8902097 1.6772929 0.88574888 0.65337000 0.52747405 0.272352088 0.215854849

Proportion of Variance 0.4466116 0.3516640 0.09806888 0.05336154 0.03477861 0.009271957 0.005824164

Cumulative Proportion 0.4466116 0.7982756 0.89634444 0.94970598 0.98448459 0.993756548 0.999580712

Comp.8

Standard deviation 0.057916350

Proportion of Variance 0.000419288

Cumulative Proportion 1.000000000

>

> # Loadings of Principal components

> loadings(pca1)

Loadings:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8

Vol. -0.585 -0.106 -0.169 0.116 -0.215 0.739

Bulk. -0.221 -0.481 -0.169 0.145 0.690 -0.249 0.347 -0.109

Pol. 0.444 -0.104 -0.143 -0.733 0.161 -0.394 -0.230

pI 0.185 -0.253 0.942

Hyd.1 -0.493 0.128 -0.354 0.343 0.497 -0.495

Hyd.2 -0.508 0.129 0.132 -0.186 -0.713 -0.405

SA 0.100 -0.566 -0.141 0.112 -0.367 0.109 -0.235 -0.660

FA -0.453 -0.172 -0.525 -0.425 0.555

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8

SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000

Proportion Var 0.125 0.125 0.125 0.125 0.125 0.125 0.125 0.125

Cumulative Var 0.125 0.250 0.375 0.500 0.625 0.750 0.875 1.000

To get eigenvalue we use:

> library("factoextra")

> eig.val <- get\_eigenvalue(d.pca)

> eig.val

eigenvalue variance.percent cumulative.variance.percent

Dim.1 3.572892791 44.6611599 44.66116

Dim.2 2.813311627 35.1663953 79.82756

Dim.3 0.784551075 9.8068884 89.63444

Dim.4 0.426892356 5.3361545 94.97060

Dim.5 0.278228871 3.4778609 98.44846

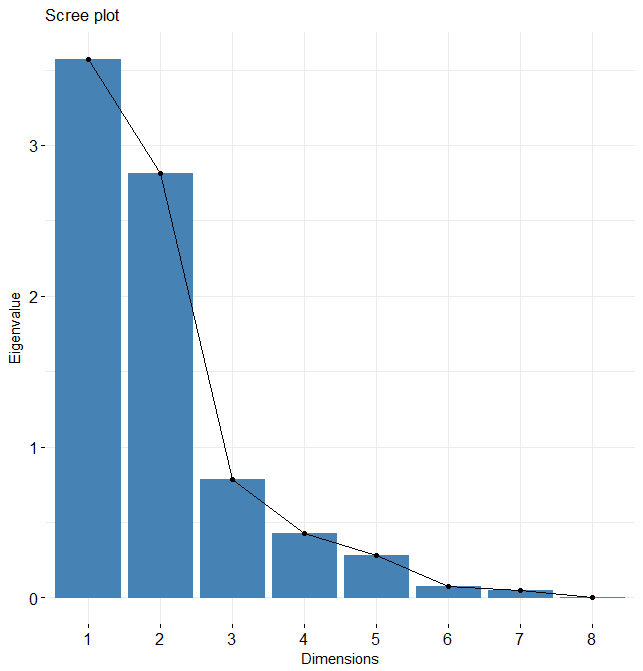
Dim.6 0.074175660 0.9271957 99.37565

Dim.7 0.046593316 0.5824164 99.95807

Dim.8 0.003354304 0.0419288 100.00000

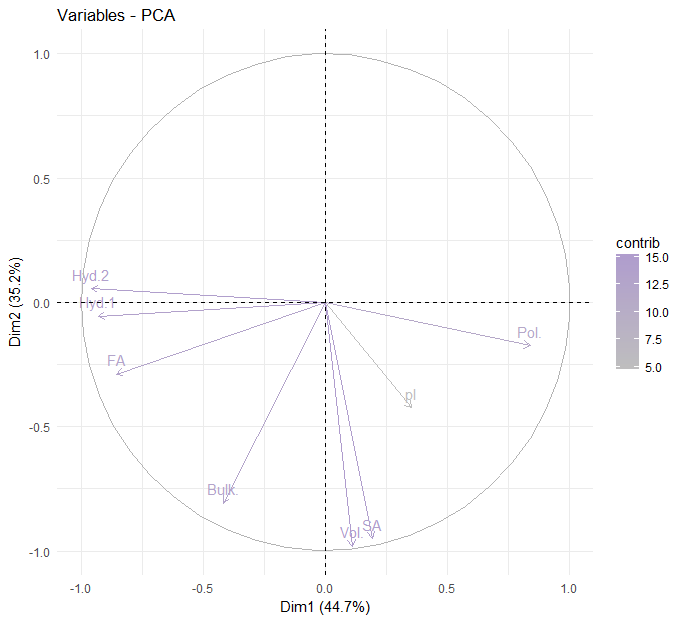
#Use to graph a scree plot to visually assess which components/factors explain most of the variability in the data

> fviz\_screeplot(d.pca, ncp=10, choice="eigenvalue")



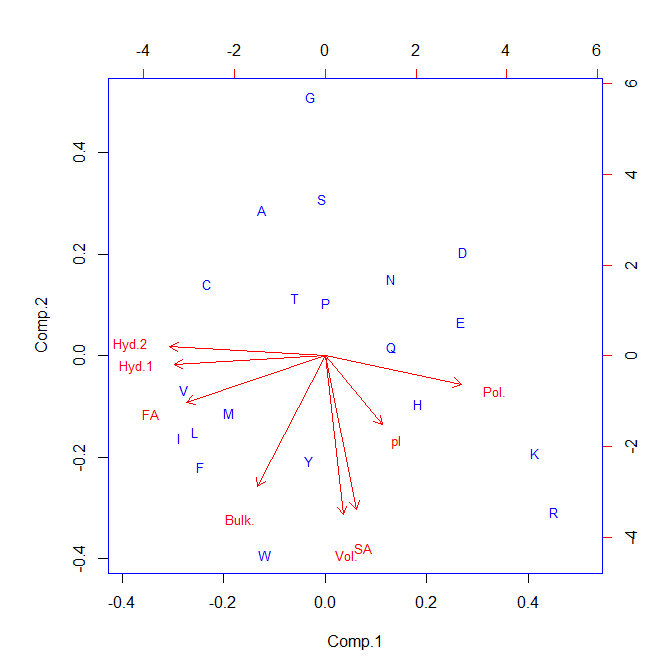
#create Variables factors graph to graph all variables

>fviz\_pca\_var(d.pca, col.var="contrib") + scale\_color\_gradient2(low="gray", mid="blue", high="red", midpoint=55) + theme\_minimal()



And lasting is create the graph based on first two components of the principal component analysis.

> biplot(pca, cex = 0.8, col = c("blue", "red"))



This is the result of PCA for amino acid in the given data on first page. For component 1, the largest contributions in the vector are the negative contributions from the hydrophobicity scales. Thus hydrophobic amino acids appear on the left side and hydrophilic ones on the right.

For the component 2, the largest contributions are positive ones from volume, bulkiness, and surface area. Thus large amino acids appear near the top of the figure and small ones near the bottom.

All the properties contribute to some extent to each of the components so the result from the graph is not the same as we would have got by simply plotting hydrophobicity against volume.