Multivariate Analysis: Finding Structure in Data

Or Goldreich

June 17, 2023

1. Motivation
2. Principal Component Analysis
3. Visualisation
4. Clustering

Dimension reduction

- In many cases, we may be trying to explain a large quantity of numerical values that may be highly correlated with each other.
- Condensing said measurements into a smaller set that encompasses most of the same information allows us to get a clearer view of the big picture.

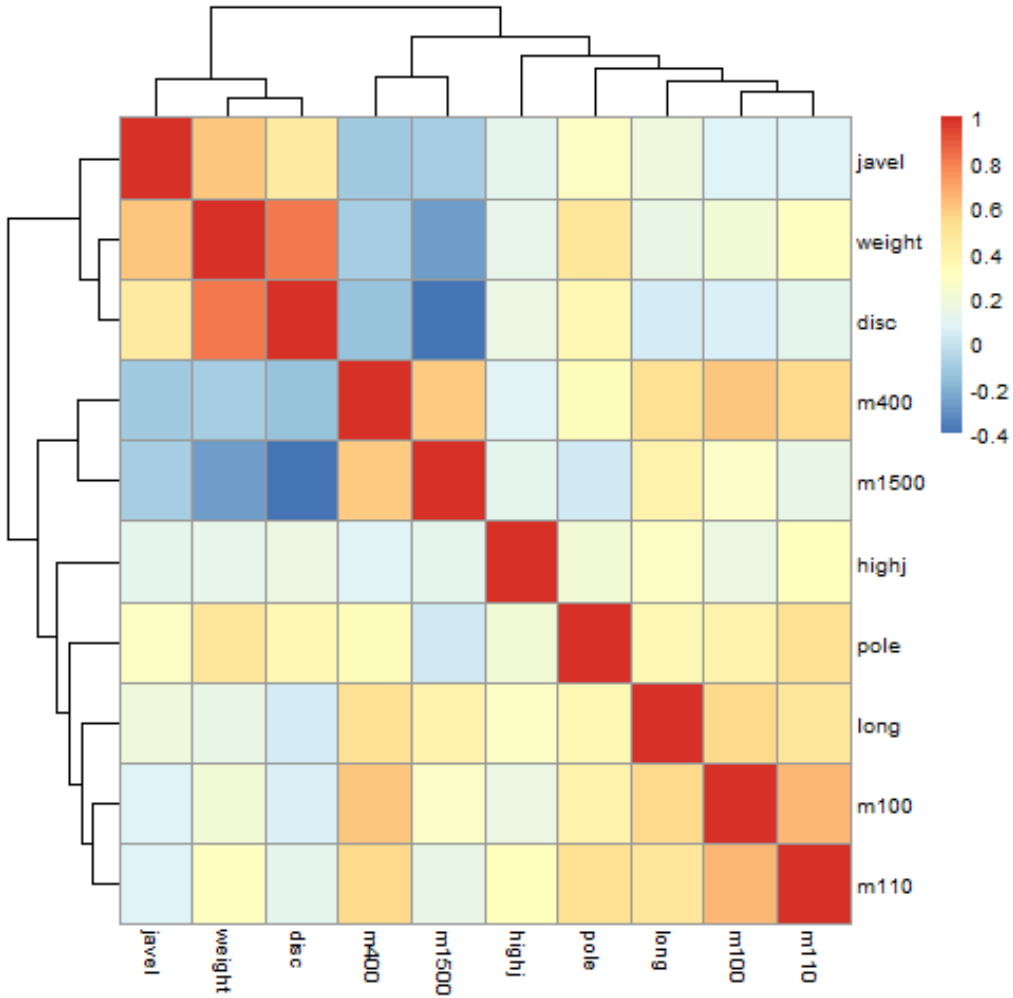
Example - Decathlon results

Let's view the first few lines of this dataset:

```
##    m100 long weight highj  m400  m110  disc pole javel  m1500
## 1 11.25 7.43  15.48  2.27 48.90 15.13 49.28  4.7 61.32 268.95
## 2 10.87 7.45  14.97  1.97 47.71 14.46 44.36  5.1 61.76 273.02
## 3 11.18 7.44  14.20  1.97 48.29 14.81 43.66  5.2 64.16 263.20
## 4 10.62 7.38  15.02  2.03 49.06 14.72 44.80  4.9 64.04 285.11
## 5 11.02 7.43  12.92  1.97 47.44 14.40 41.20  5.2 57.46 256.64
```

Example - Decathlon results

And the correlation heatmap:



Linear combinations

- Given a set of vectors $v_1, \ldots, v_n$, a vector u is said to be a linear combination of these vectors if there exist parameters $\alpha_1, \ldots, \alpha_n$ such that $u = \sum_{i=1}^{n} \alpha_i v_i$. The parameters need not be unique, though in many practical applications this is a desirable feature.
- Examples:
  - $u = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ is a linear combination of $v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ with parameters $\alpha_1 = 2$ and $\alpha_2 = 3$.
  - $u = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ is a linear combination of $v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ with parameters $\alpha_1 = 1$ and $\alpha_2 = -1$.
  - $u = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ is not a linear combination of $v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Linear combinations

- There is generally no need to solve these equations manually, as R has these capabilities.
- We are interested in finding linear combinations of our quantities of interest that represent most of the variance within them.

Preprocessing

- The processes we will go over can be sensitive to location and scaling. It is generally best practice to standardise each of the columns (that is, centre them all with mean zero, and scale them to have a standard deviation of 1).
- We apply this to the variables weight and disc in our data for the next demonstration.

Principal Component Analysis

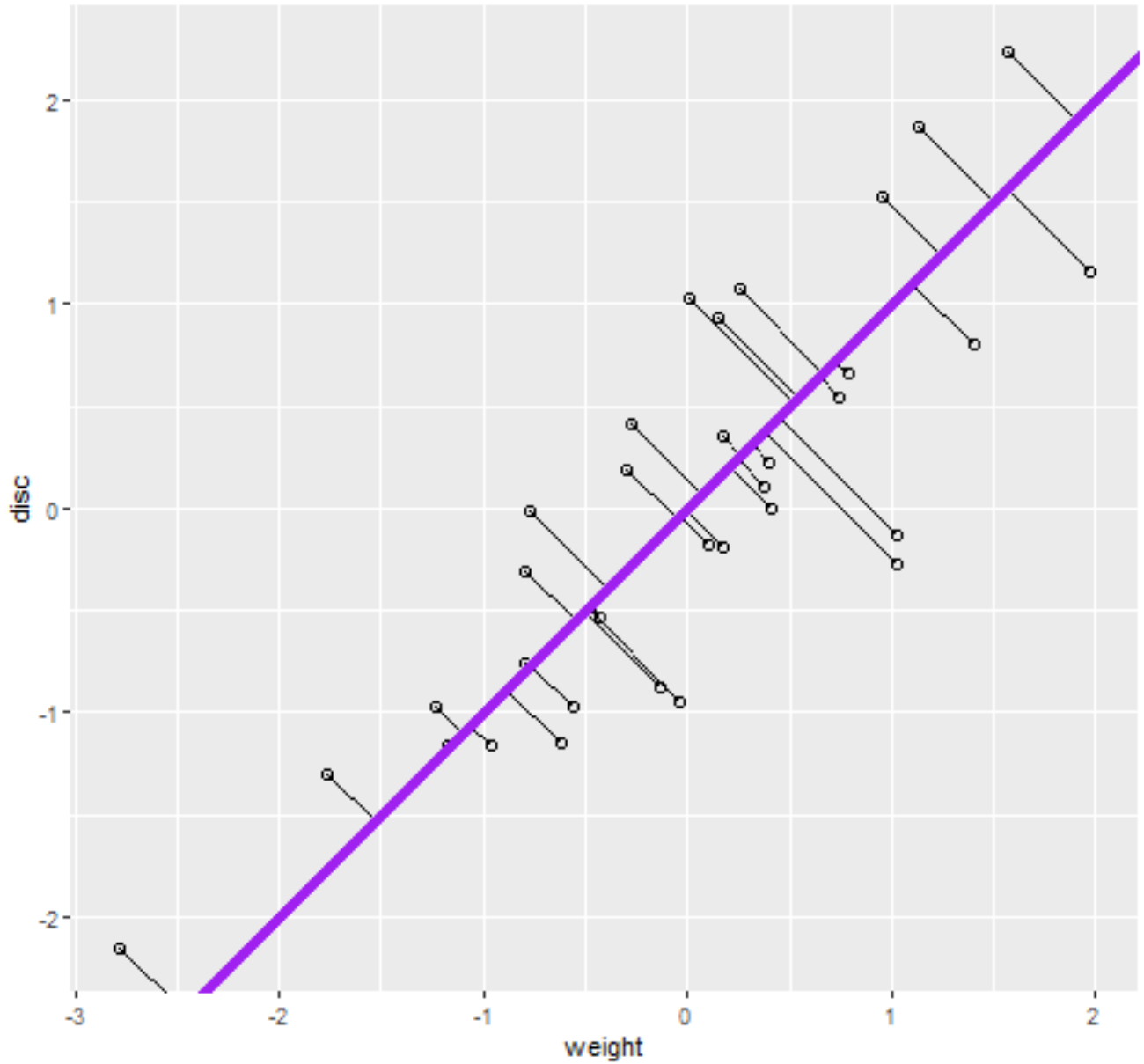
- Said linear combinations are known as the principal components of our data.
- In the case of two variables, we can visualise this in a manner similar to the least squares regression line.
- As opposed to OLS, where we minimise the sum of squared vertical distances between the fitted values and the observations, here we minimise this with respect to the two-dimensional distances.

PCA Line - code

```
X = cbind(athletes$disc, athletes$weight)
svda = svd(X)
pc = X %*% svda$v[, 1] %*% t(svda$v[, 1])
bp = svda$v[2, 1] / svda$v[1, 1]
ap = mean(pc[, 2]) - bp * mean(pc[, 1])

p + geom_segment(xend = pc[,1], yend = pc[,2]) +
geom_abline(intercept = ap, slope = bp, col = "purple", lwd = 1.5) +
    coord_fixed()
```

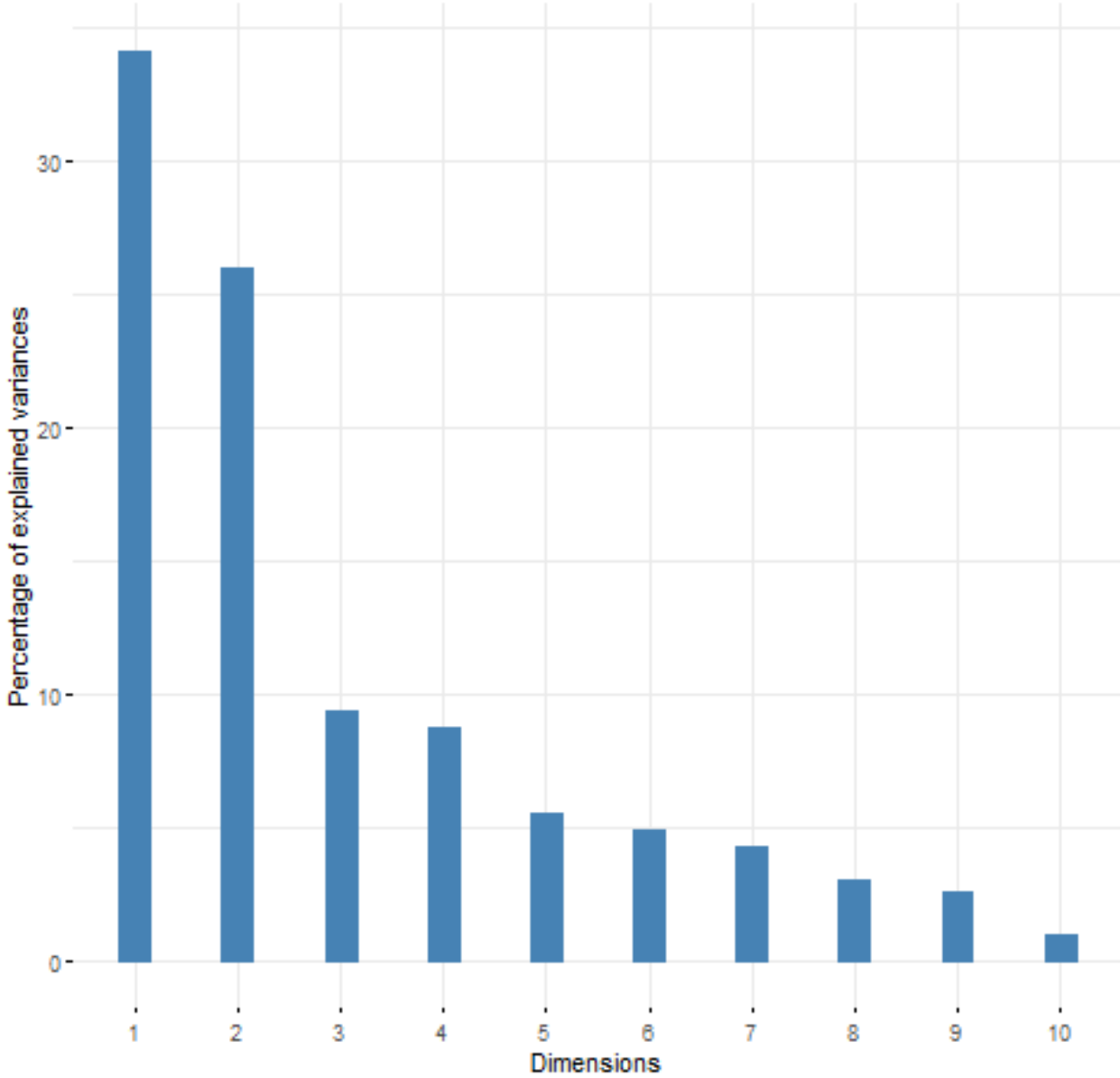
PCA Line - output



Scree Plots

- How do we decide how many dimensions to include after the reduction?
- The most useful tool is the scree plot, showing us what part of the variance in observed value is explained by each axis.

Scree Plots


Percentage of explained variances
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Dimensions

Hierarchical Clustering

- Given a set of datapoints and a corresponding distance matrix (that can be computer in R with the dist function), we can use the base R function hclust.
- Any vector can be treated as a datapoint for this purpose. In fact, we've already seen one instance of hierarchical clustering earlier today, with each discipline in the decathlon being treated as a datapoint.

k-means

- It might be prudent to assume that out data comes from several underlying distributions, each with its own mean. In that case, by setting k, the amount of said means, we have a tool to assign our data to once of k clusters corresponding to said means in an unsupervised fashion.
- We are required to choose the value of k in advance. In the case that said value is not obvious, we can use cross-validation in order to make that decision.