# Aspects of multivariate analysis

**Woojoo Lee**

# The objectives of multivariate data analysis

1. Data reduction : e.g. project multidimensional data on 2-d plane
   how to make 1-d score by summarizing multidimensional data → principal component analysis, factor analysis

2. Sorting and grouping : e.g. clustering based on a proper similarity measure, classification and discrimination

3. Investigation of the dependence among variables

   : e.g. (canonical) correlation analysis, factor analysis, pca

4. Prediction : by exploiting the relationship between variables

5. Hypothesis testing : e.g. multiple comparison

Data type

$x_{jk}$ : measurement of the $k$th variable on the $j$th item

Note the dimension of the subscript !

→

$$X = \begin{pmatrix} & \text{Variable1} & \text{Variable2} & \cdots & \text{Variable p} \\ \text{Item1} & x_{11} & \cdots & & \\ \text{Item2} & \vdots & & & \\ \vdots & & & & \\ \text{Item n} & & & & \end{pmatrix}$$

Remark) Item/ Individual/ experimental unit are often exchangeable .
Variables/ Characters are often exchangeable.

Descriptive statistics

Sample means : $\frac{1}{p}\sum_j x_{ij}$ vs $\frac{1}{n}\sum_i x_{ij} (= \bar{x}_j)$

Sample covariance matrix :

$$s_{jk} = \frac{1}{n}\sum_i (x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})$$

Sample correlation matrix :

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} (= r_{kj})$$

Q) What kind of information can you get from the sample correlation ?

# Arrays of Basic Descriptive Statistics

Sample means

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Sample variances and covariances

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Sample correlations

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

From Johnson and Wichern, 6ed

# Graphical techniques

Box plots

Scatter plots and their variants

Growth curves

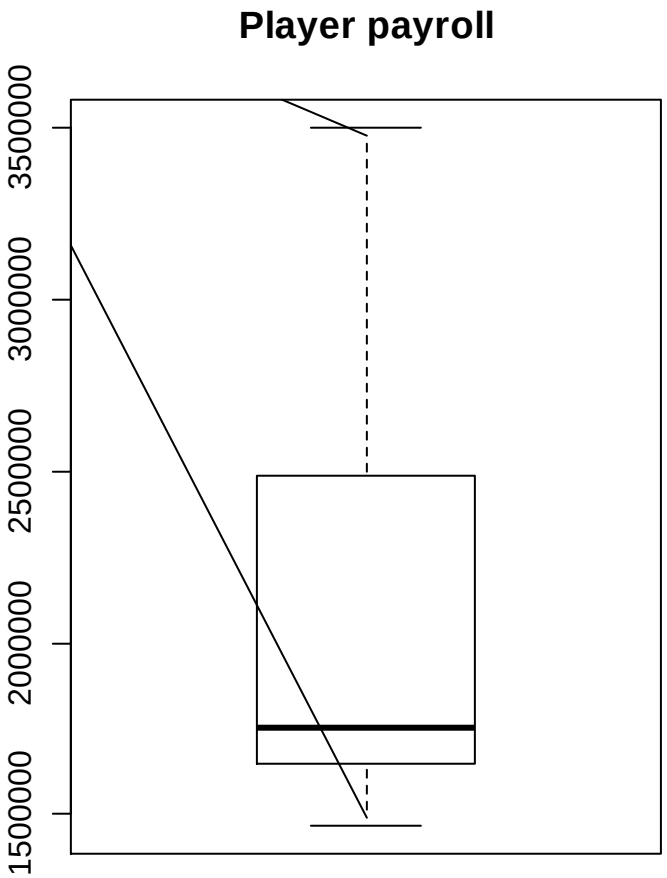
Chernoff faces

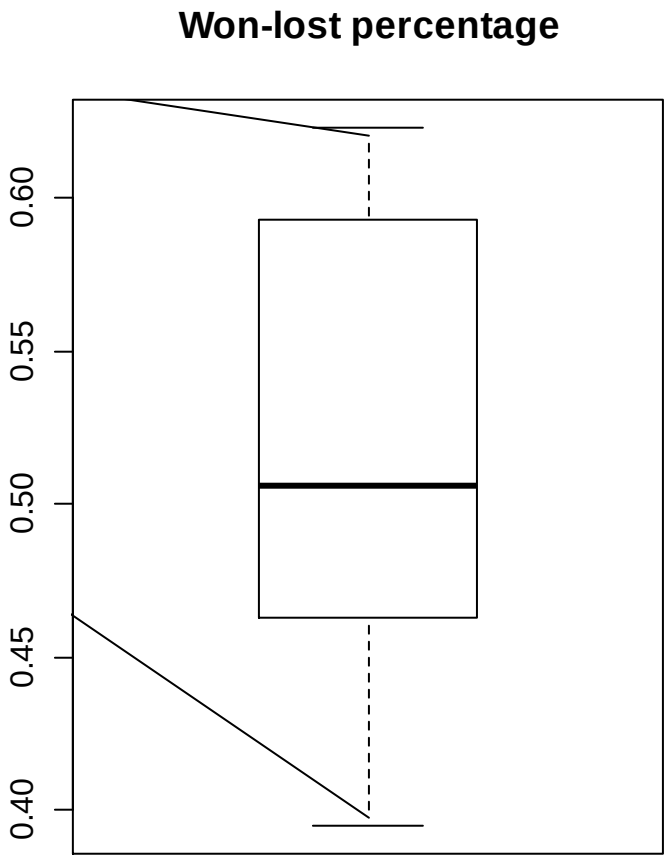
These graphical tools are usually informative in data analysis.
We will see some of them in the next slides.

Box plot ( Table.1.1)

Salary and final record for the national league east baseball team in 1977

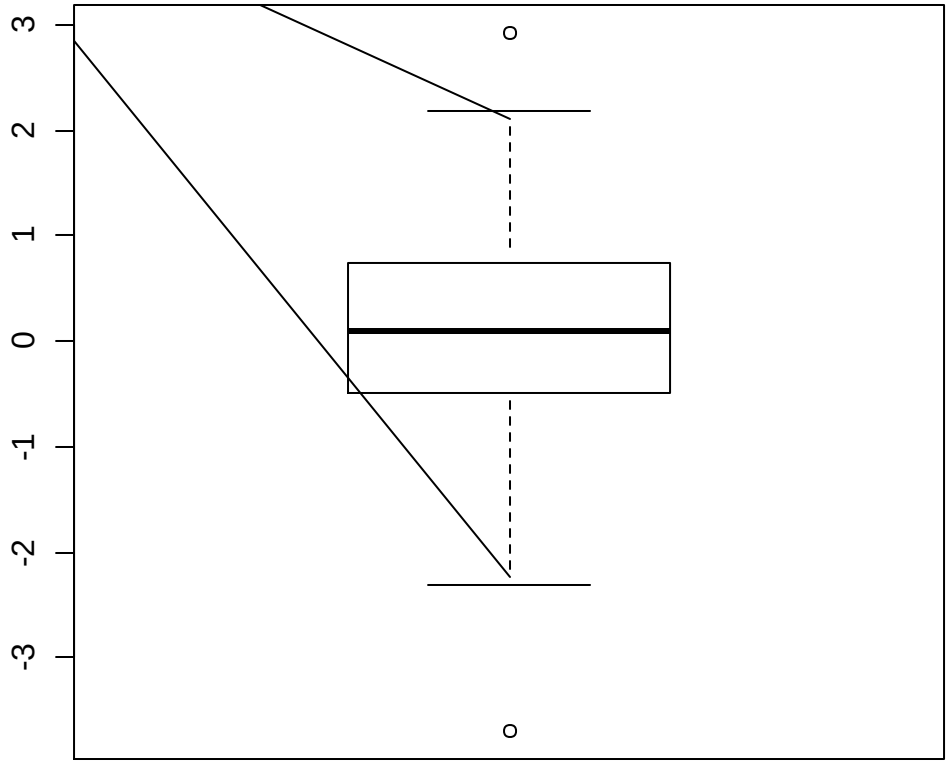
| | Player | payroll | Won-lost percentage |
|---|---|---|---|
| 1 | | 3497900 | 0.623 |
| 2 | | 2485475 | 0.593 |
| 3 | | 1782875 | 0.512 |
| 4 | | 1725450 | 0.500 |
| 5 | | 1645575 | 0.463 |
| 6 | | 1469800 | 0.395 |




Dispersion, skewness

A simulated example

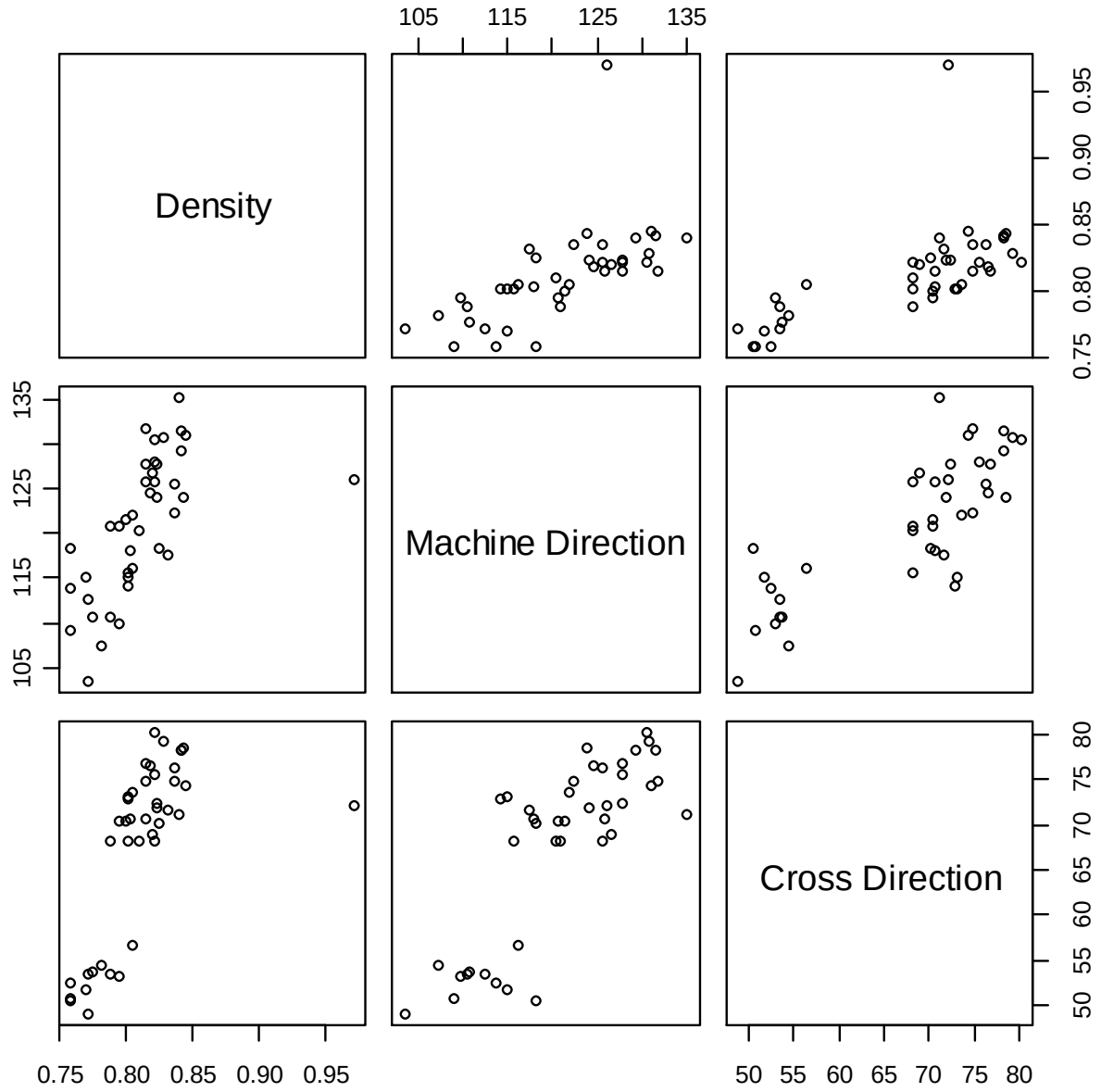



possible to identify outliers

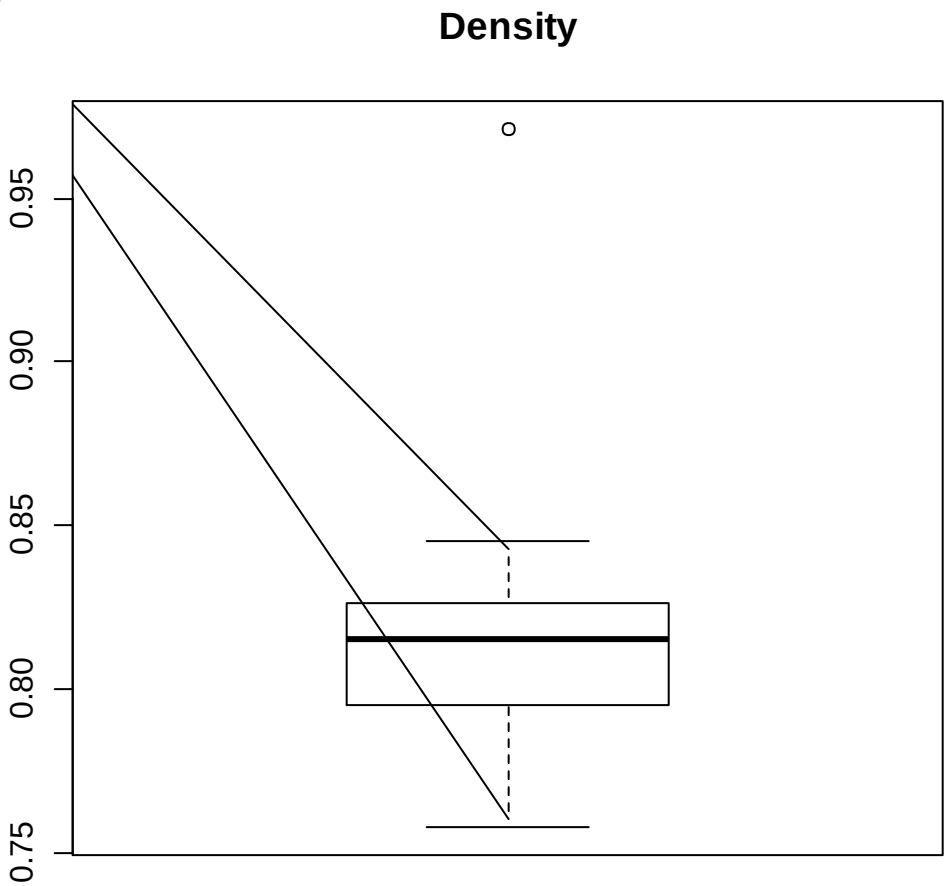
Table 1.2. Paper-quality measurements

$x_1$: density (grams/cubic centimeter)
$x_2$: strength (pounds) in the machine direction
$x_3$: strength (pounds) in the cross direction

| | Density | Machine Direction | Cross Direction |
|---|---|---|---|
| 1 | 0.801 | 121.41 | 70.42 |
| 2 | 0.824 | 127.70 | 72.47 |
| 3 | 0.841 | 129.20 | 78.20 |
| 4 | 0.816 | 131.80 | 74.89 |
| 5 | 0.840 | 135.10 | 71.21 |
| 6 | 0.842 | 131.50 | 78.39 |


Density
Machine Direction
Cross Direction
105
115
125
135
0.75
0.80
0.85
0.90
0.95
50
55
60
65
70
75
80

Boxplot, again



Using
"ggpairs"
With tips data
in R

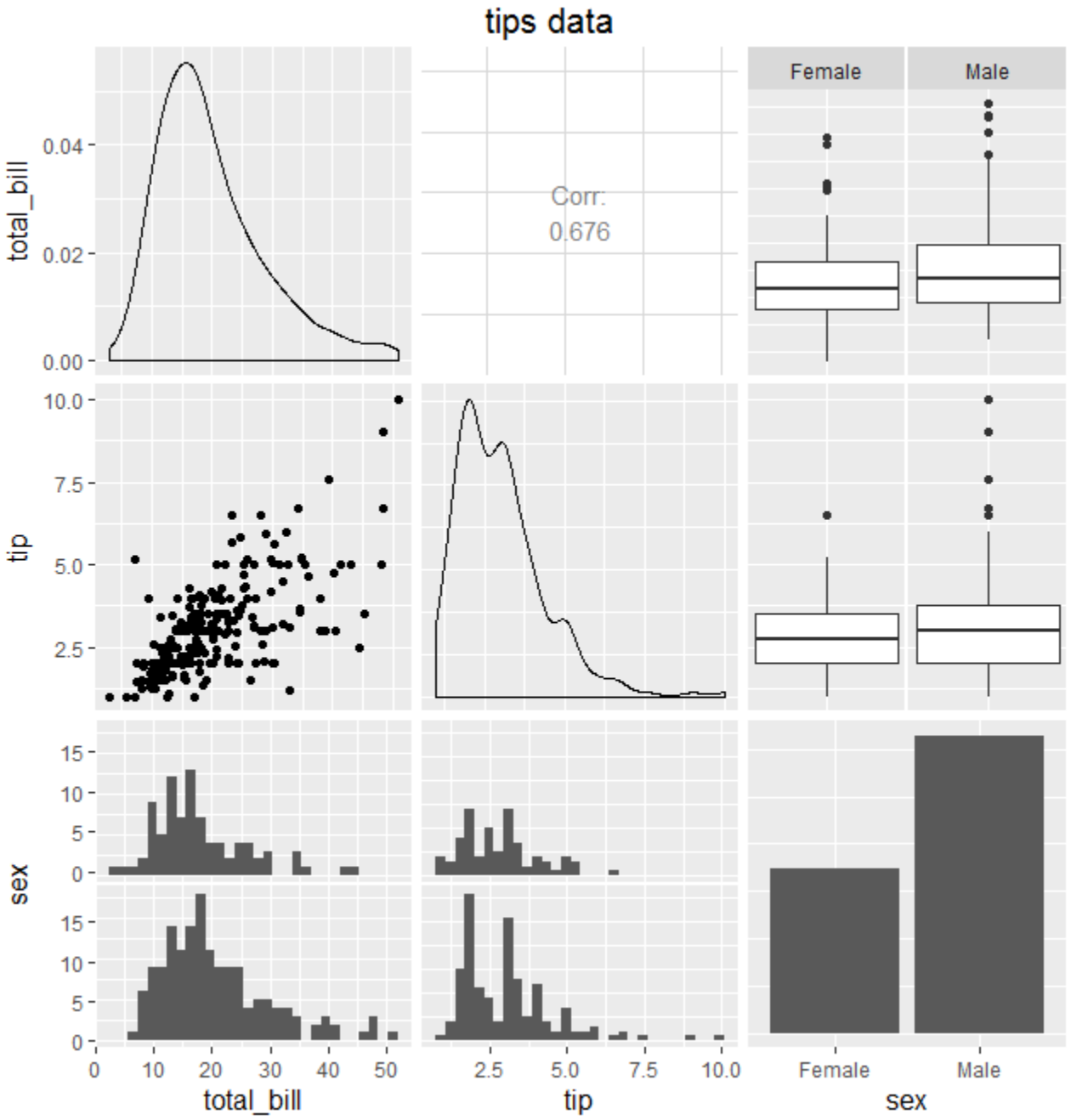
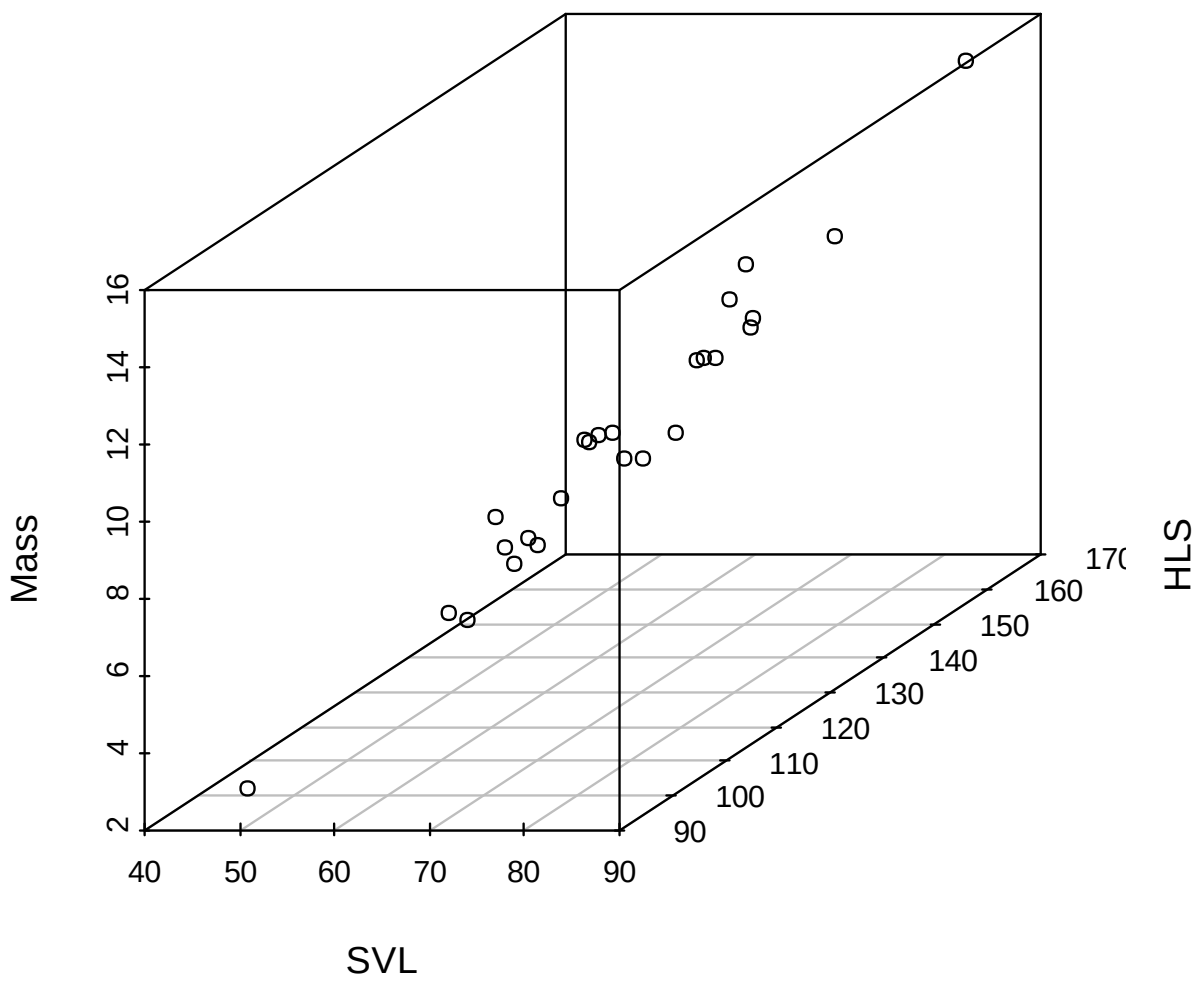


Table.1.3. Lizard size data

A zoologist obtained measurements on 25 lizards. The weight (or mass) is given in grams while the snoutvent length (SVL) and hind limb span (HLS) are given in milimeters.

| | Mass | SVL | HLS |
|---|---|---|---|
| 1 | 5.526 | 59.0 | 113.5 |
| 2 | 10.401 | 75.0 | 142.0 |
| 3 | 9.213 | 69.0 | 124.0 |
| 4 | 8.953 | 67.5 | 125.0 |
| 5 | 7.063 | 62.0 | 129.5 |
| 6 | 6.610 | 62.0 | 123.0 |




Remark) If one variable has much larger variability, the standardization can be considered.

Spinning 3d Scatterplot

A snapshot of spinning 3d scatter plot (Table.1.2)

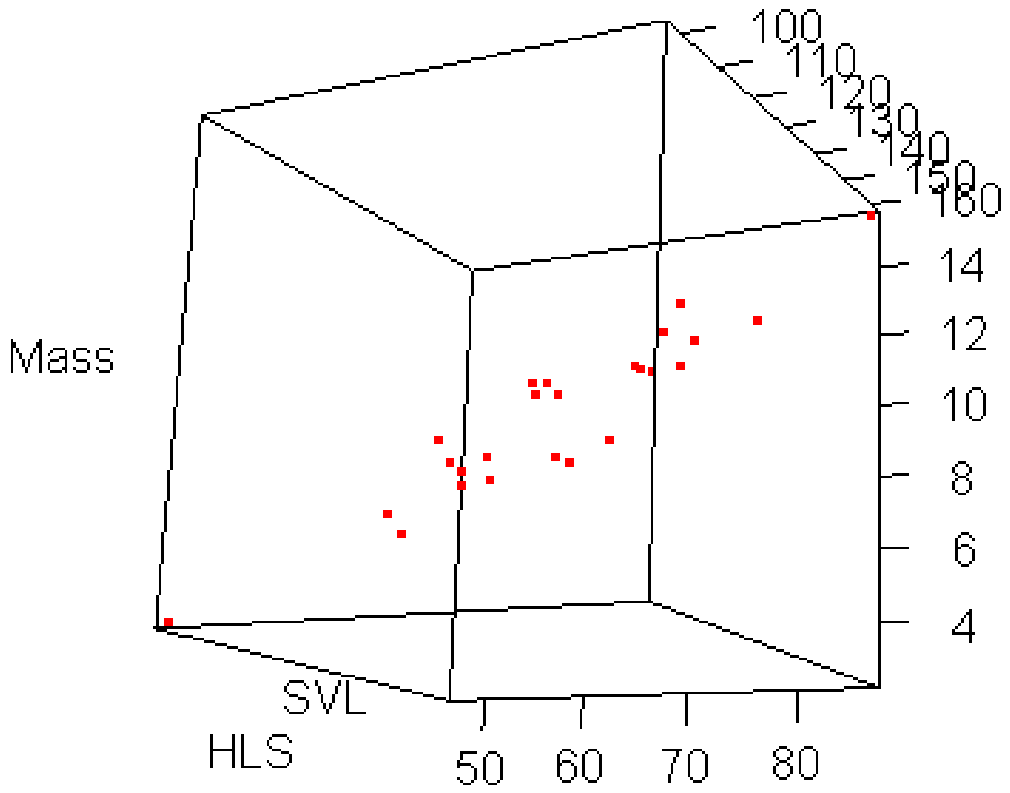
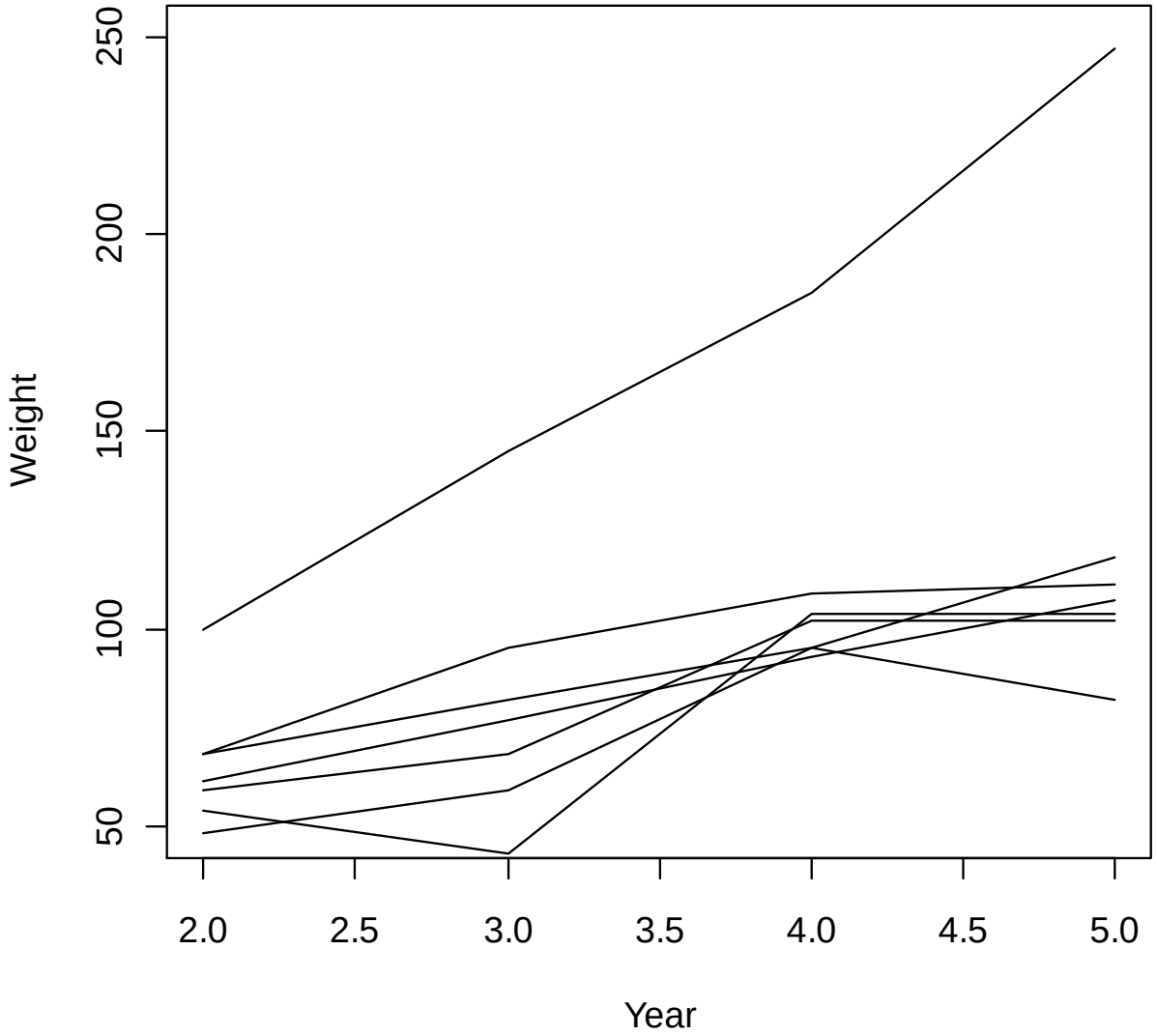


Table. 1.4. Female Bear Data

Repeated measurements of the same characteristic on the same unit or subject can give rise to a growth curve.

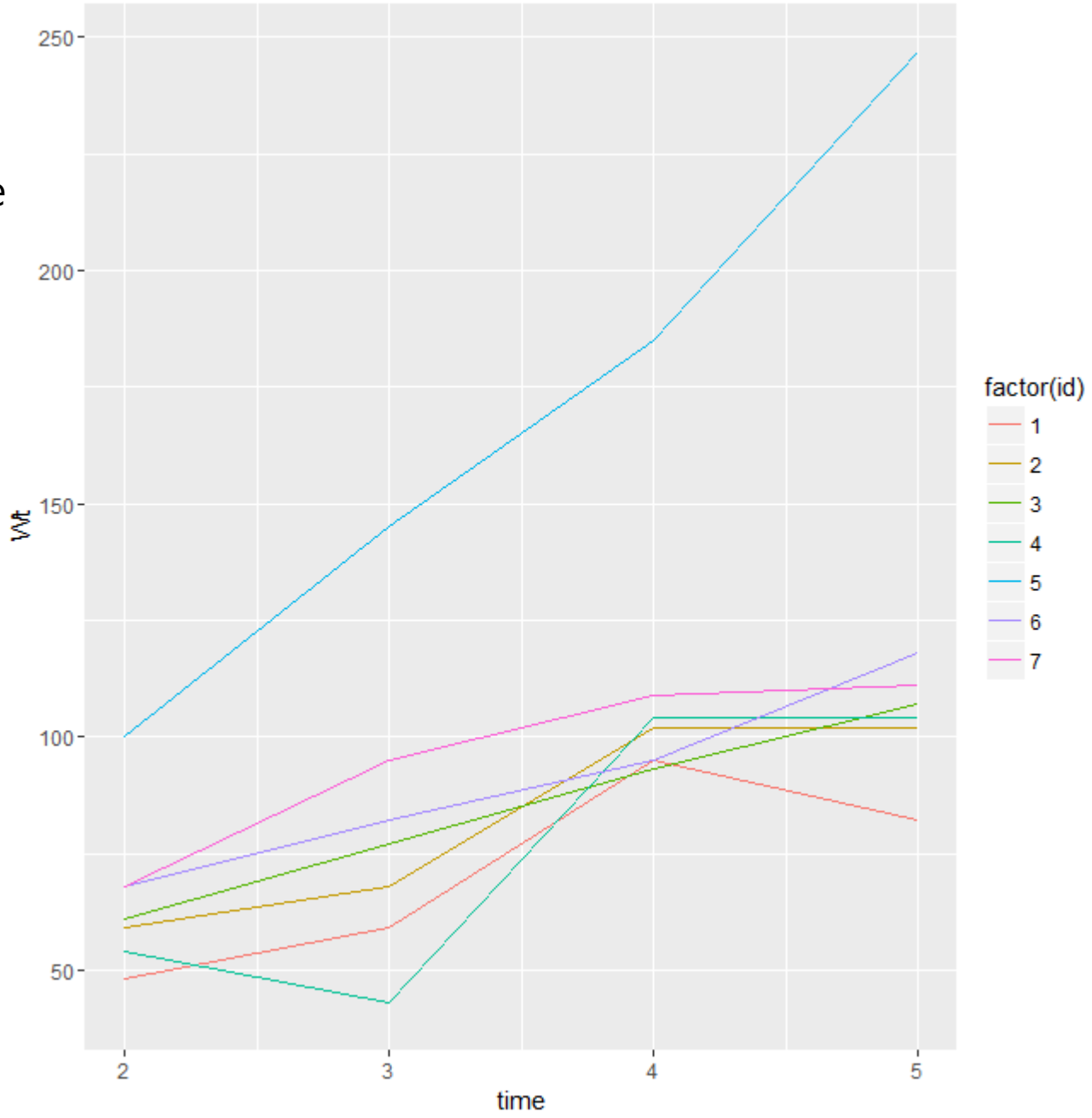
| | Wt2 | Wt3 | Wt4 | Wt5 | Length2 | Length3 | Length4 | Length5 |
|---|---|---|---|---|---|---|---|---|
| 1 | 48 | 59 | 95 | 82 | 141 | 157 | 168 | 183 |
| 2 | 59 | 68 | 102 | 102 | 140 | 168 | 174 | 170 |
| 3 | 61 | 77 | 93 | 107 | 145 | 162 | 172 | 177 |
| 4 | 54 | 43 | 104 | 104 | 146 | 159 | 176 | 171 |
| 5 | 100 | 145 | 185 | 247 | 150 | 158 | 168 | 175 |
| 6 | 68 | 82 | 95 | 118 | 142 | 140 | 178 | 189 |

Combined growth curves of weight for seven female grizzly bears




Outlier vs natural variation in the population ?

Using “qplot”
in ggplot2 package

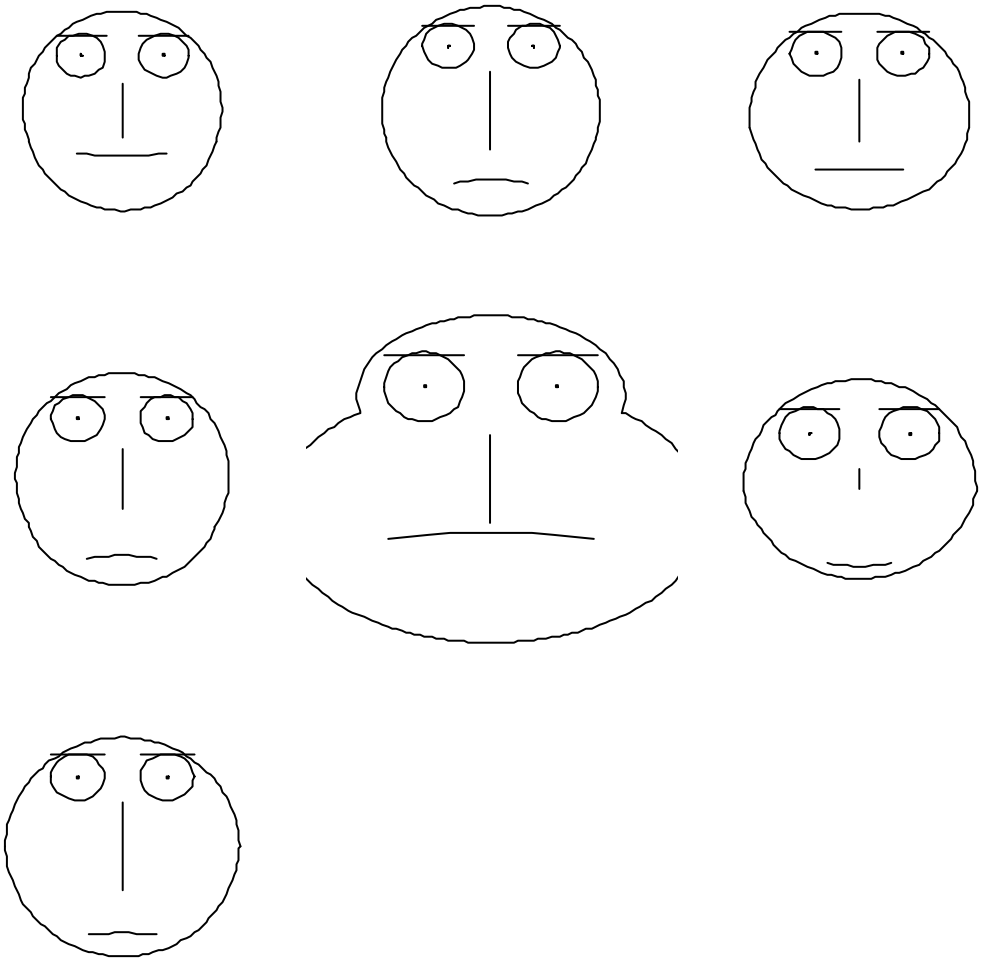


Chernoff faces : people react to faces !

For instance, Face2 in TeachingDemos

1: Width of Center
2:Top vs Bottom width
3:Height of Face
4:Width of top half of face
⋮
17:Angle of Eyebrows
18:Width of Eyebrows

Chernoff faces ( using Table 1.4.)



Distance

Most multivariate techniques are based on the concept of distance.

Distance : a measure to define similarity !

$$P = (x_1, \cdots, x_p)^T \quad Q = (y_1, \cdots, y_p)^T$$

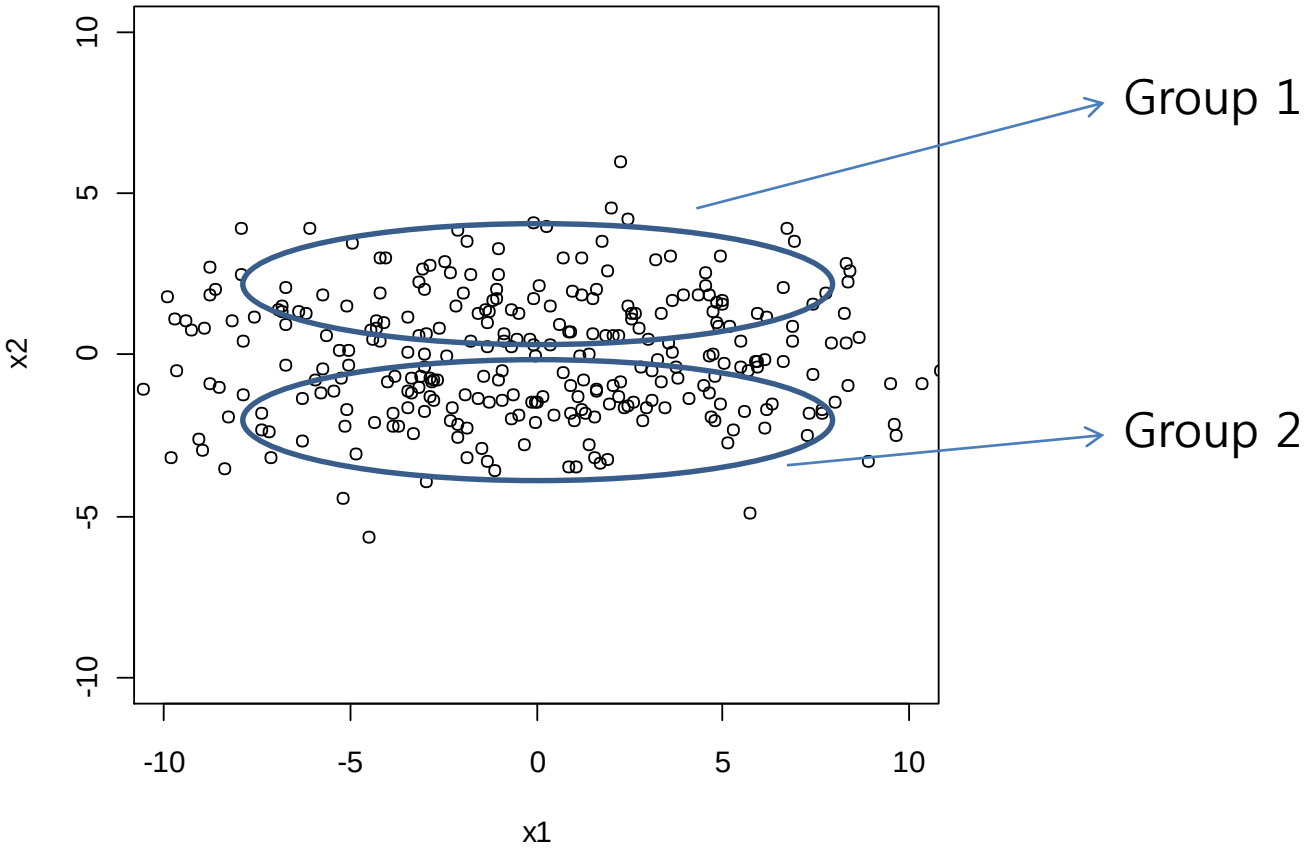
1. $d(P, Q) = d(Q, P)$
2. $d(P, Q) \geq 0$
3. $d(P, Q) \leq d(P, R) + d(R, Q)$

The above function $d(\cdot, \cdot)$ is called "distance" !

e.g. Euclidean distance

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

But,

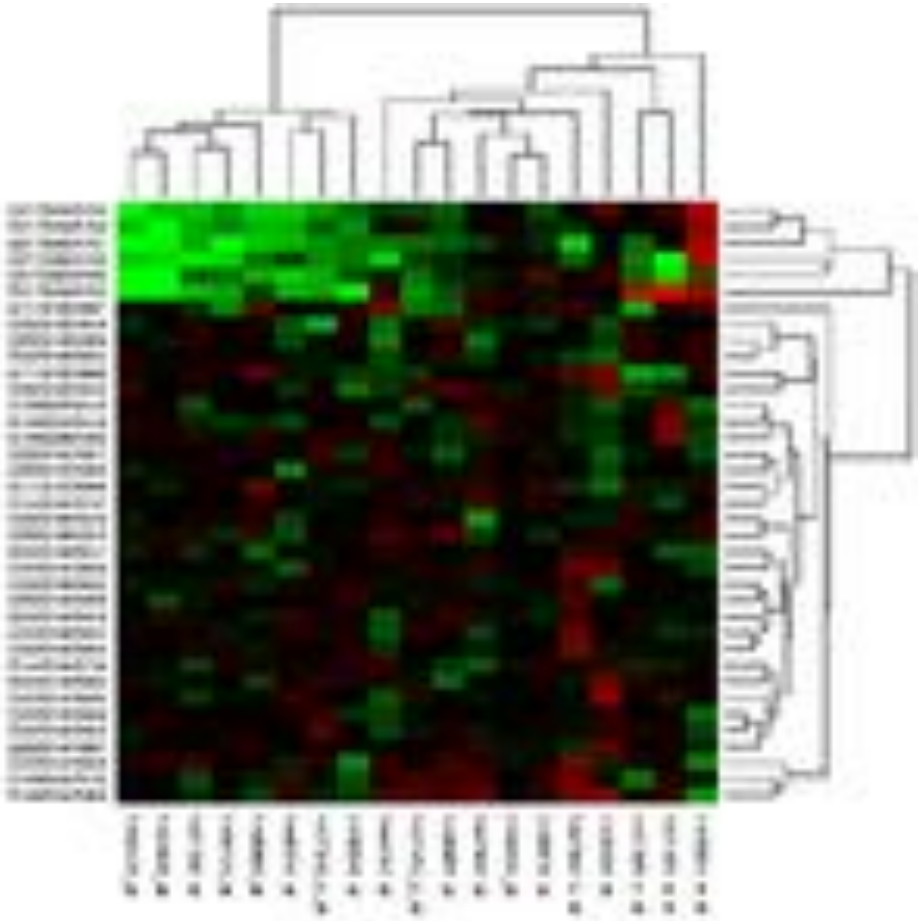


Statistical distance

$$d(P,Q) = \sqrt{(\frac{x_1 - y_1}{\sqrt{s_{11}}})^2 + (\frac{x_2 - y_2}{\sqrt{s_{22}}})^2 + \cdots + (\frac{x_p - y_p}{\sqrt{s_{pp}}})^2}$$

Is this distance ?

Recently, high-dimensional data (very large number of variables !) are everywhere !



Examples: microarrays, sequencing data, SNPs, fMRIs, EEGs, rating data

An example of rating data

| | Anne | Ben | Charlie | Doug | Eve | ... |
|---|---|---|---|---|---|---|
| Star Wars | 2 | 5 | 4 | 4 | 3 | ... |
| Harry Potter | 3 | 4 | 5 | 3 | ? | ... |
| Pretty Woman | 4 | ? | 2 | ? | 5 | ... |
| Titanic | 5 | ? | 2 | 1 | 3 | ... |
| Lord of the Rings | ? | 5 | 5 | 4 | 4 | ... |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ |

Movie ratings : Scale of 1-5

The usual distance is still meaningful in high dimensional problems ?

$$X = (X_1, \cdots, X_p)^T \text{ and } Y = (Y_1, \cdots, Y_p)^T$$
where $X_i$ and $Y_i$ are independent $N(0,1)$ and $p$ is very large

1) $d(X, 0) =$
2) $d(X, Y) =$
3) $\text{angle}(X, Y) =$

What is your conclusion from 1), 2) and 3) ?

K-nearest neighbor algorithm (k-NN)

# *k*-nearest neighbors algorithm

From Wikipedia, the free encyclopedia

In pattern recognition, the ***k*-nearest neighbors algorithm** (***k*-NN**) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

From https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm