Why do we need Multivariate Analyses — what does it do? What do we achieve?

---

① Dimension reduction

② Structural simplification

③ Understanding dependence among variables

④ Sorting & grouping

Cluster analyses is a sorting algorithm

EDA
- PCA, FA ⟶ dimension reduction of the covariate set/features
- CA — grouping of obs. ⟶ non-parametric

Classification — grouping algorithm (parametric)

CA

① Identify groups
② Identify no. of groups } based on corr.
③ Group new obs.

↓

measure of proximity

What we need?

① Measurement on ϕ variables for n objects ($n \gg \phi$)
② Proximity measure
③ Standardisation (?)

Solid Waste generated per year in US

Year / Waste	1990	1995	. . .	2020
X_1 (plastic waste)				
X_2 (paper waste)				
X_3 (metal waste)				
X_4 (glass waste)				

dependence across col^n
not multivariate data
rather multivariate time series data

Proximity measures

① Distance measures

(i) $d(i,j) \geq 0$

(ii) $d(i,j) = 0 \Rightarrow i = j$

(iii) $d(i,j) = d(j,i)$

Triangle inequality ← (iv) $d(i,j) \leq d(i,k) + d(k,j)$

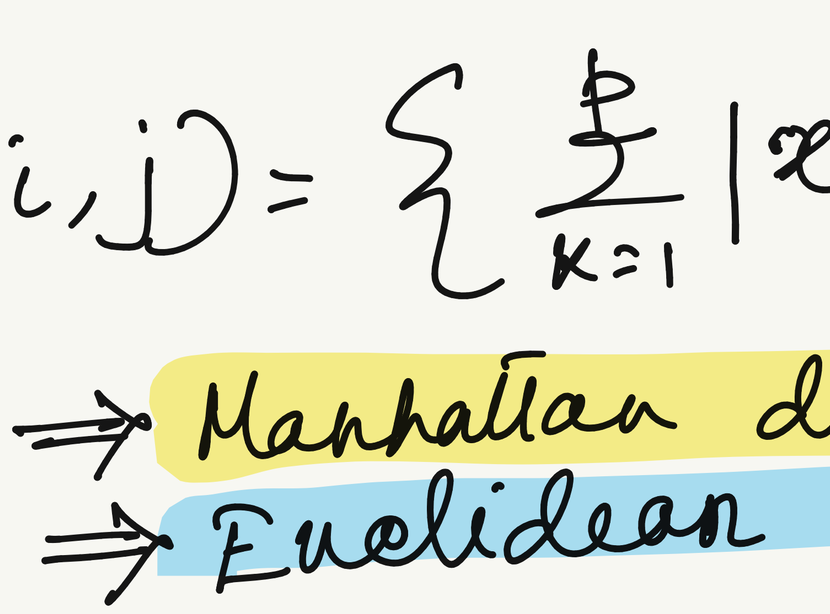
$$d(i,j) = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^{q} \right\}^{1/q}, \quad \begin{matrix} i \neq j \\ i,j = 1(1)n \end{matrix}$$

$q = 1 \Rightarrow$ Manhattan distance (Taxicab distance)

$q = 2 \Rightarrow$ Euclidean distance

$q \Rightarrow$ Minkowski distance

$p = 2$



Weighted Euclidean : $d(i,j) = \sqrt{\sum \omega_k (x_{ik} - x_{jk})^2}$

Dissimilarity / Similarity measures

(2) Correlation based measures

$$d(i,j) = \frac{1 - r_{ij}}{2}$$

$$d(i,j) = 1 - |r_{ij}|$$

$$s(i,j) = |r_{ij}|$$

$$s(i,j) = \frac{1 + r_{ij}}{2}$$

Measure based on both correlation and distance : Mahalanobis distance

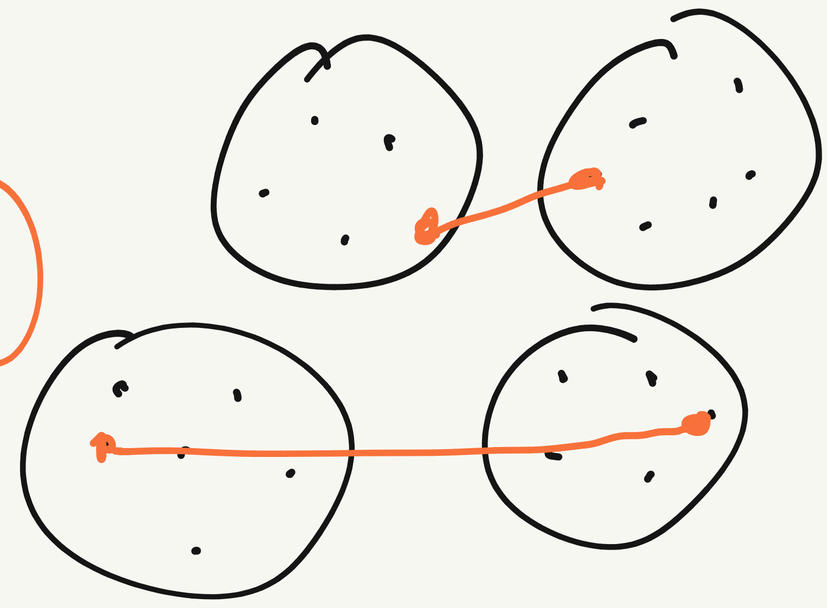
(most popular and useful in practice)

$$d(i,j) = \sqrt{(\underset{\sim}{x}_i - \underset{\sim}{x}_j)' S^{-1} (\underset{\sim}{x}_i - \underset{\sim}{x}_j)}$$

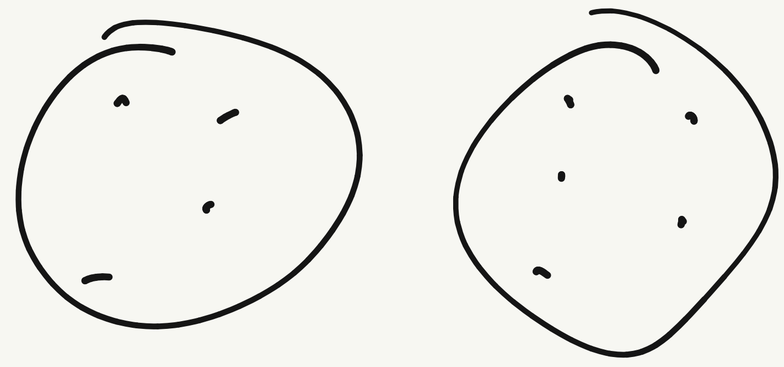
Distance between clusters

(1) Single linkage
(Nearest neighbour)

(2) Complete linkage
(Farthest neighbour)



(3) Average linkage
(average distance)



Clustering

Hierarchical

Divisive
1 cluster
↓
n clusters

Agglomerative
n clusters
↓
1 cluster

Non-hierarchical

k-means clustering

Centroid : $\bar{x}(A) = (\bar{x}_1(A), \ldots, \bar{x}_p(A))$

Cluster A : $\bar{x}_j(A) = \frac{1}{|A|} \sum_{i \in A} x_{ij}$, $|A| = n(A)$

$d(A,B) = \sqrt{\sum_{j=1}^{p} \{\bar{x}_j(A) - \bar{x}_j(B)\}^2}$ distance between 2 clusters

How to find no. of clusters in k-means method?

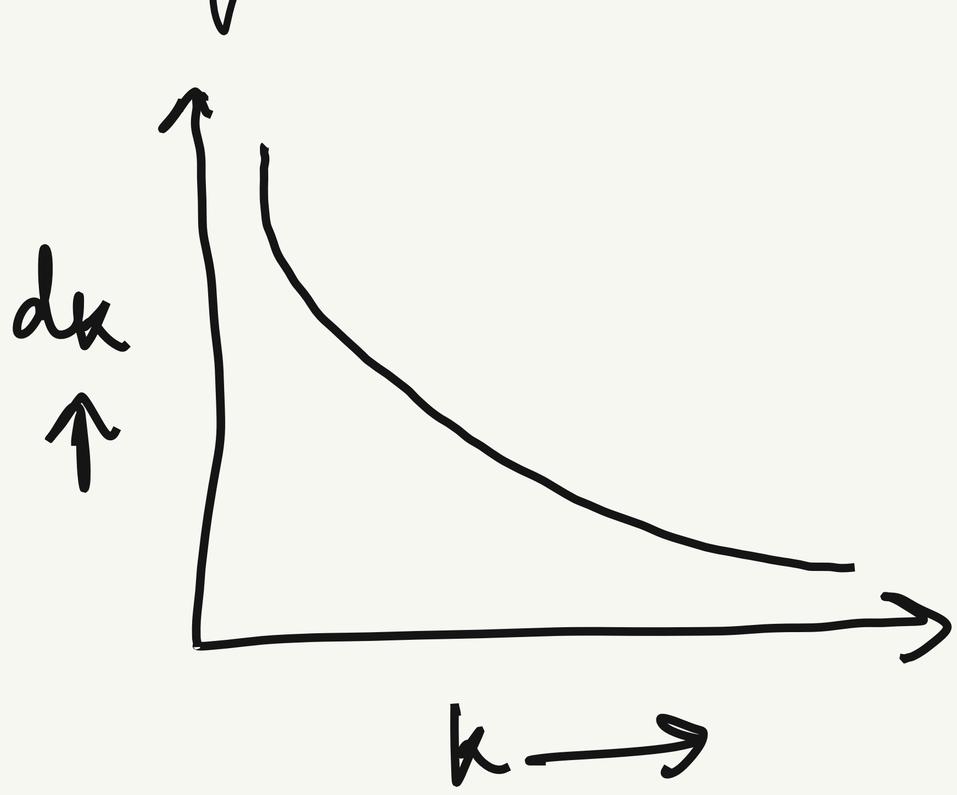
① Run k-means algorithm for all ($k = 1, 2, \ldots$) and calculate the estimates

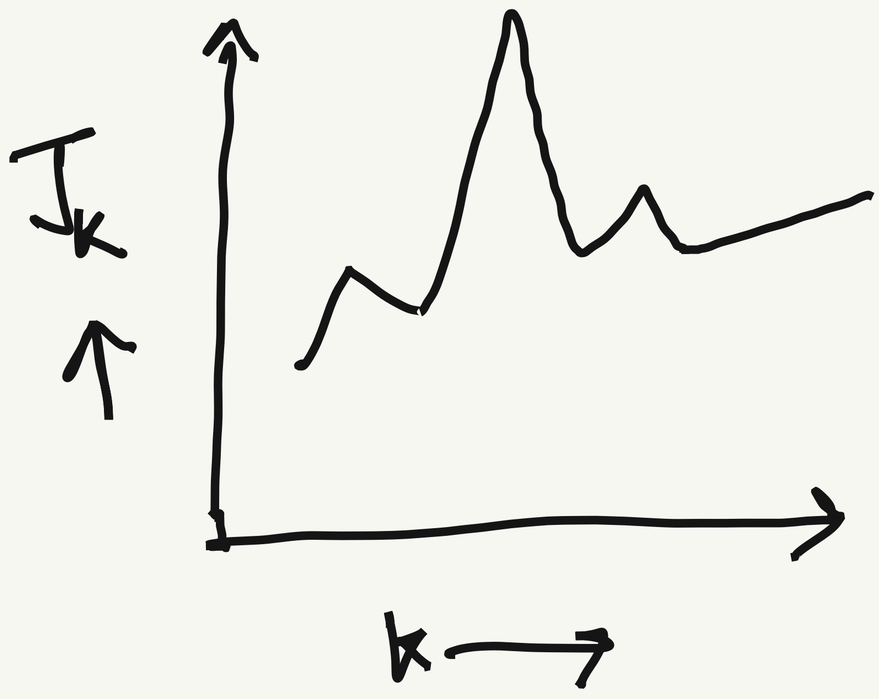
$$d_k = \frac{1}{p} \sum_{i,j} (x_{ij} - \bar{x}_i)^2$$

② Select a power index, say $\gamma = \frac{p}{2}$

③ Calculate jumps $J_k = d_k^{-\gamma} - d_{k-1}^{-\gamma}$

④ Estimate k which gives you largest jump





Hartigan method: Choose smallest k such that

$H(k) \leq 10$, where, $H(k) = (n-k-1)\left[\frac{\omega(k)}{\omega(k+1)} - 1\right] \sim F_{p,(n-k-1)p}$

$\omega(k)$: within cluster SS