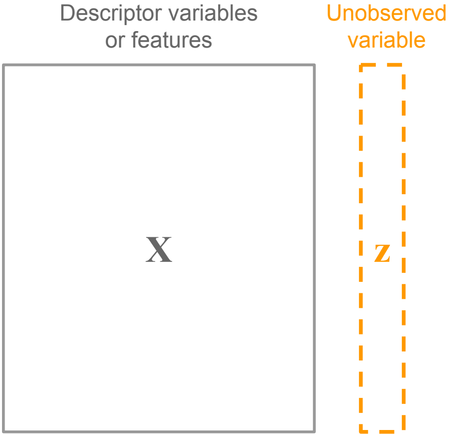
Introduction to Clustering

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Descriptive/Unsupervised Methods



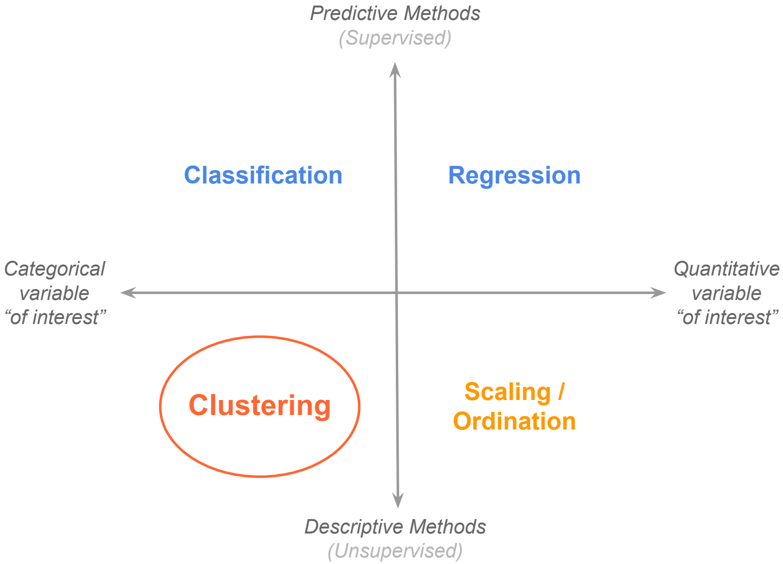
Descriptor variables or features

Unobserved variable

$\mathbf{X}$

$\mathbf{z}$

Predictive and Descriptive Methods



Predictive Methods
(Supervised)

Classification

Regression

Categorical variable "of interest"

Quantitative variable "of interest"

Clustering

Scaling / Ordination

Descriptive Methods
(Unsupervised)

Clustering

Clustering refers to a very broad set of techniques for finding groups, or clusters, in a data set.

Clustering Examples

- **Marketing**: discover groups of customers and used them for targeted marketing.
- **Medical Field**: discover groups of patients suitable for particular treatment protocols.
- **Astronomy**: find groups of similar stars and galaxies.
- **Genomics**: find groups of genes with similar expressions.

Clustering (from Tuffery 2011)

Clustering is the statistical operation of grouping objects (individuals or variables) into a limited number of groups known as clusters.

On the one hand, groups are not defined in advanced by the analyst, but are discovered during the analysis, unlike the classes used in classification.

On the other hand, the clusters are combinations of objects having similar characteristics, which are separated from objects having different characteristics (in other clusters).

Clustering Idea

Group a set of n objects in K groups such that:

- each group is as much homogeneous as possible
  i.e. *within-groups homogeneity*.
- groups are as distinct as possible among them
  i.e. *between-groups heterogeneity*

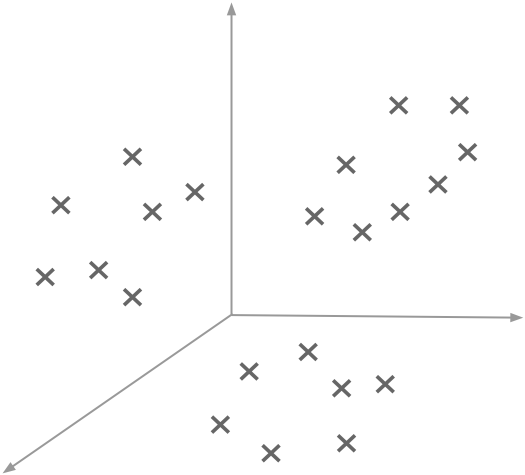
Geometric Perspective

Sum of Squared Distances

Before looking at some clustering approaches, let's define the **between-cluster** and **within-cluster** sum of squared distances.

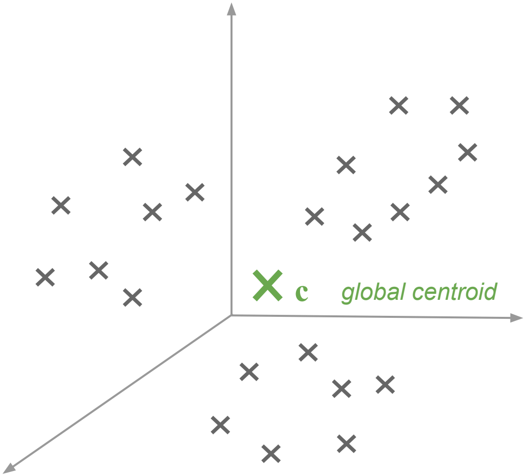
Why? Because we will focus on methods that are based on some criterion of such sums of squares.

Data as a cloud of points in p-dim space



Cloud of n points in p-dimensional space

Global centroid



The global *centroid* $\mathbf{c}$ is the point of averages

Global Centroid

The global centroid $\mathbf{c}$ is the point of averages which consists of the point formed with all the variable means:

$$\mathbf{c} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]$$

where:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

If all variables are mean-centered, the centroid is the origin

$$\mathbf{c} = \underbrace{[0, 0, \dots, 0]}_{p \text{ times}}$$

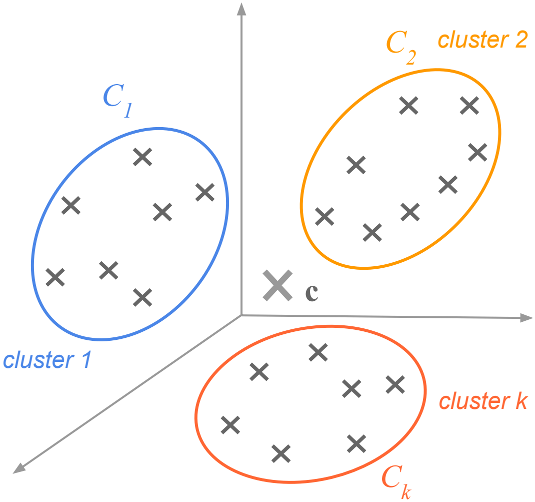
Total Dispersion

We can look at the amount of spread or dispersion in the data with respect to the global centroid.

The sum of all squared distances from the centroid is:

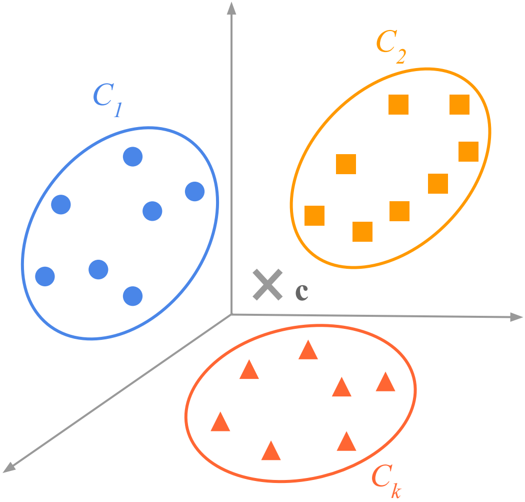
$$\mathsf{T} = \sum_{i=1}^{n} d^2(\mathbf{x_i}, \mathbf{c})$$

Clustering Configuration



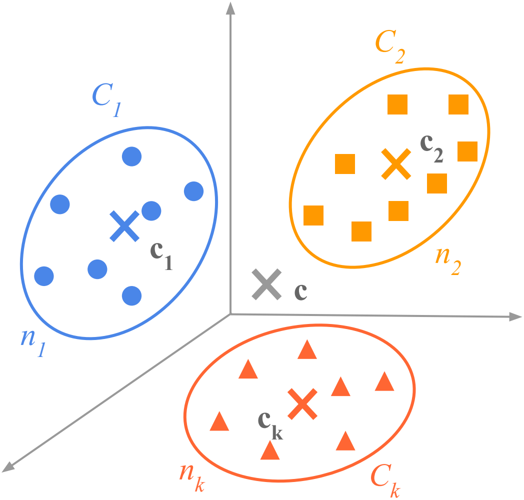
The objects can be divided into clusters or groups

Cloud of points for each cluster



Each cluster C_k forms its own cloud

Cluster centroids



Each cluster C_k has its own centroid $\mathbf{c_k}$

Cluster Centroids

The cluster centroid $\mathbf{c_k}$ is the point of averages for those observations in cluster k:

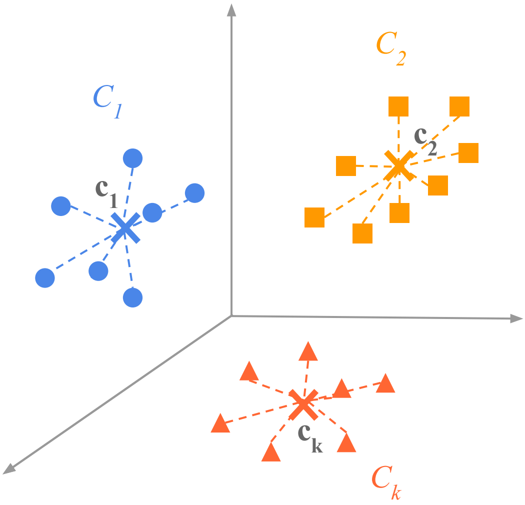
$$\mathbf{c_k} = [\bar{x}_{1k}, \bar{x}_{2k}, \dots, \bar{x}_{pk}]$$

with:

$$\bar{x}_{jk} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}$$

where n_k is the number of objects in cluster C_k

Within-clusters dispersion



We can focus on the dispersion within each cluster

Within-clusters dispersion

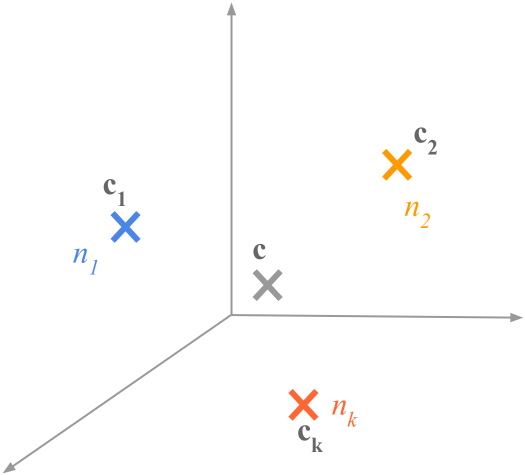
Each cluster will have an associated amount of spread from its centroid, that is a *Cluster Sum of Squared distances* (CSS):

$$\mathsf{CSS}_k = \sum_{i \in C_k} d^2(\mathbf{x_i}, \mathbf{c_k})$$

We can combine the cluster sum-of-squared distances to obtain the Within-clusters sum of squared distances W:

$$\mathsf{W} = \sum_{k=1}^{K} \sum_{i \in C_k} d^2(\mathbf{x_i}, \mathbf{c_k})$$

Global and Group Centroids



What if we focus on just the centroids?

Dispersion between clusters

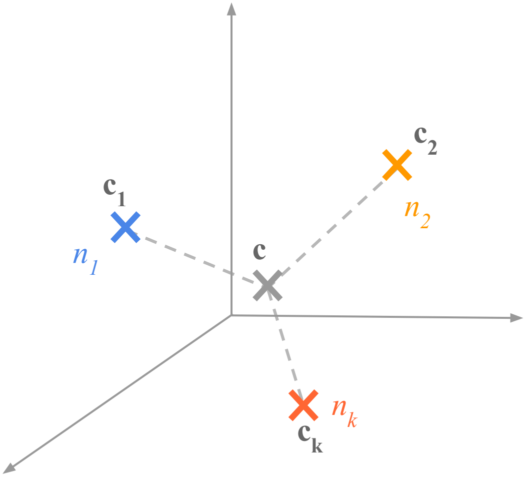
Focusing on just the centroids, we can get the Between-cluster sum of squared distances

$$\mathsf{B} = \sum_{k=1}^{K} n_k \, d^2(\mathbf{c_k}, \mathbf{c})$$

Note that the global centroid $\mathbf{c}$ can be expressed as a weighted average of the cluster centroids:

$$\mathbf{c} = \frac{n_1}{n}\mathbf{c_1} + \frac{n_2}{n}\mathbf{c_2} + \cdots + \frac{n_K}{n}\mathbf{c_K} = \sum_{k=1}^{K} \left(\frac{n_k}{n}\right) \mathbf{c_k}$$

Between-clusters dispersion



Dispersion between the centroids

Dispersion Decomposition

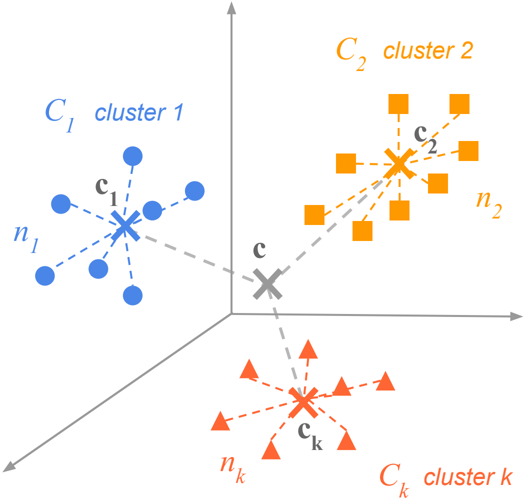
Let's recap. We have three types of sums-of-squared distances:

- T: Total sum of squared distances
- W: Within-clusters sum of squared distances
- B: Between-clusters sum of squared distances

It can be shown (Huygens theorem) that T = B + W:

$$\sum_{i=1}^{n} d^2(\mathbf{x_i}, \mathbf{c}) = \sum_{k=1}^{K} n_k \, d^2(\mathbf{c_k}, \mathbf{c}) + \sum_{k=1}^{K} \sum_{i \in C_k} d^2(\mathbf{x_i}, \mathbf{c_k})$$

Dispersion Decomposition



$$T = B + W$$

Dispersions and Clustering

Cluster analysis aims to minimize the within-clusters sum of squared distances W, to a fixed number of cluster K.

A cluster becomes more homogenous as its sum of squared distances decreases, and the clustering of the data set becomes better as W decreases.

Also, as B increases, the separation between clusters also increases, indicating satisfactory clustering.

Dispersion Decomposition

Thus there are two criteria for “correct” clustering: B should be large and W should be small.

From the decomposition formula, $T = B + W$, for a fixed number of cluster K, minimizing W or maximizing B are equivalent.

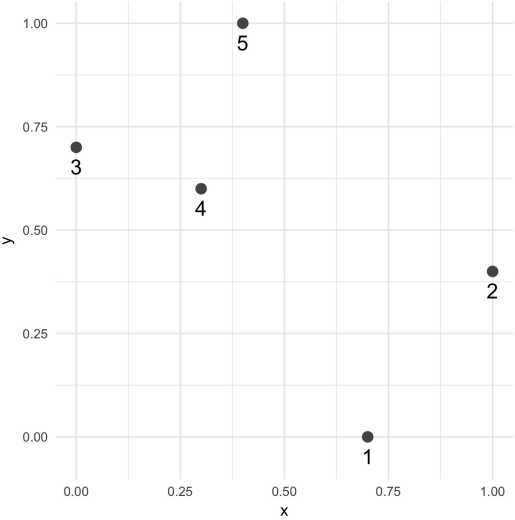
Complexity in Cluster Analysis

Toy Data Set

Consider a toy data set with $n = 5$ observations and 2 variables:

```
    x   y
1 0.7 0.0
2 1.0 0.4
3 0.0 0.7
4 0.3 0.6
5 0.4 1.0
```

Toy example



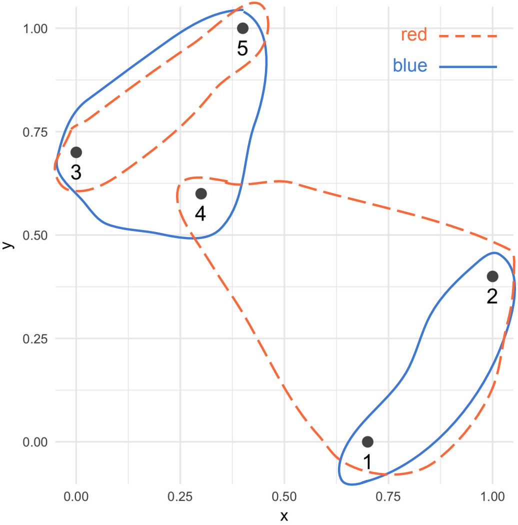
Distance Matrix

Euclidean squared distance between objects i and l
(i.e. $\mathbf{x_i}$ and $\mathbf{x_l}$)

$$d_{il}^2 = d^2(\mathbf{x_i}, \mathbf{x_l}) = \sum_{j=1}^{p} (x_{ij} - x_{lj})^2 = \|\mathbf{x_i} - \mathbf{x_l}\|_2^2$$

```
     1    2    3    4    5
1 0.00
2 0.25 0.00
3 0.98 1.09 0.00
4 0.52 0.53 0.10 0.00
5 1.09 0.72 0.25 0.17 0.00
```

Two possible clustering configurations



Within-cluster dispersions

Here $n = 5$ and $K = 2$

Red clustering:

$$W_{red} = \frac{0.25 + 0.53 + 0.52}{3} + \frac{0.25}{2} = 0.56$$

Blue clustering:

$$W_{blue} = \frac{0.25}{2} + \frac{0.10 + 0.17 + 0.25}{3} = 0.30$$

Clustering Partitions

Smaller W is better, so why don't we just directly find the clustering partition C that minimizes W?

To find the partition with smallest W would require trying all possible assignments of the n objects into K groups.

How many possible assignments of n objects into K groups?

Number of clustering partitions

The number of possible assignments of n objects into K groups is given by:

$$A(n, K) = \frac{1}{K!} \sum_{k=1}^{K} (-1)^{(K-k)} \binom{K}{k} k^n$$

Note that:

- $A(10, 4) = 34105$, and
- $A(25, 4) \approx 5 \times 10^{13} \ldots$ huge!.

Let's see another view of the complexity of a clustering problem.

Complexity of Clustering

Consider a data set of four objects: $\{a, b, c, d\}$. The possible (non-overlapping) partitions for $n = 4$ objects are:

- 1 partition with 1 cluster:
  $(abcd)$
- 7 partitions with 2 clusters:
  $(ab, cd), (ac, bd), (ad, bc), (a, bcd), (b, acd), (c, abd)(d, abc)$
- 6 partitions with 3 clusters:
  $(a, b, cd), (a, c, bd), (a, d, bc), (b, c, ad), (b, d, ac), (c, d, ab)$
- 1 partition with 4 clusters:
  (a, b, c, d)

Complexity of Clustering

With four objects, we could form clusters of sizes 1, 2, 3, and 4. Their corresponding number of assignments are:

- $A(4,1) = 1$
- $A(4,2) = 7$
- $A(4,3) = 6$
- $A(4,4) = 1$

Hence, the different number of (non-overlapping) partitions for different cluster sizes is:

$$A(4,1) + A(4,2) + A(4,3) + A(4,4) = 15$$

Complexity of Clustering

The number of (non-overlapping) partitions for n objects is the so-called **Bell** number B_n:

$$B_n = \frac{1}{e} \sum_{k=1}^{\infty} \frac{k^n}{k!}$$

For $n = 4$ objects, we have $B_4 = 15$

For $n = 30$ objects, we have $B_{30} = 8.47 \times 10^{23} = \ldots$ a huge number greater than Avogadro's number (6.022×10^{23}), which is the number of molecules in one mole of any gas!

As a general rule, $B_n > \exp(n)$.

Complexity of Clustering

Because of the complexity of clustering problems, we'll have to settle for an approximation.

Otherwise they would be intractable.

Clustering Idea

- We seek a partition of the data into distinct groups.
- We want the observations within each group to be quite similar to each other.
- We must define what it means for two or more observations to be similar or different.
- This is often a domain-specific consideration that must be made based on knowledge of the data being studied.

Assumptions

We will assume that the rows of the data matrix correspond to the individuals to be clustered (although you could also cluster variables).

We will assume that the individuals are embeded in a Euclidean space (e.g. quantitative variables, or output of a dimension reduction method)

Clustering Approaches

Clustering Paradigms

2 main paradigms:

Crisp (hard) -vs- Fuzzy (soft)

Clustering Approaches

Crisp (hard) approaches:

- **Direct Partitioning**: Clusters are always separated, and their number is (usually) defined *a priori*.
- **Hierarchical**: clusters are either separate or embeded, defining a hierarchy of partitions.

Fuzzy (soft) approaches

- **Overlapping**: two clusters can have a number of objects in common (overlapping clusters).
- **Fuzzy**: each object has a certain probability of belonging to a given cluster.

We will focus of Crisp methods (won't cover soft approaches).

Direct Partitioning Methods

With partitioning methods, two clusters are always separated; and the number of clusters is generally defined a priori. The main methods are:

- K-means
- K-medoids
- K-prototypes
- Methods based on a concept of density
- Kohonen networks (or maps)

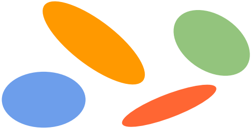
Hierarchical Methods

Two cluster are separate or one contains the other: the methods used here are hierarchical ascendant methods (known as “agglomerative”) or descendant (known as “divisive”). The main agglomerative methods are:

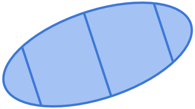
- Single linkage
- Average linkage
- Complete linkage
- Ward’s method

Data Structures

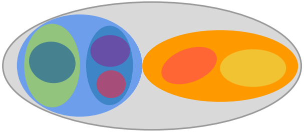
Several Populations



One Population



Hierarchical Structure



Methodological Considerations

Methodological Considerations

- The definition of natural clusters is a tricky matter.
- Clusters are not always as obvious as presented in textbooks.
- Cluster configurations differ according to the employed algorithm.
- The determination of the “real” number of clusters is emperical as much as theoretical.

Methodological Considerations

- Clusters are often required to be readable and interpretable, rather than theoretically optimal.
- e.g. the data are naturally clustered into three groups, but four clusters are requested.
- Practical matters must be taken into account when choosing the number of clusters.
- In practice, sometimes analysts mix/combine different clustering methods to obtain *consolidations*.
- e.g. Perform hierarchical clustering on a sample, determine number of clusters, and then apply direct partitioning on the entire set.

Bibliography

- **Modern Multivariate Statistical Techniques** by Julian Izenman (2008). *Chapter 12: Cluster Analysis*. Springer.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 9: Cluster Analysis*. Wiley.

French literature

- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2006). *Chapter 11: Methodes de classification*. Editions Technip, Paris.
- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 2, section 2.1: Agregation autour des centres mobiles*. Dunod, Paris.
- **Approche pragmatique de la classification** by Nakache and Confais (2005). *Chapter 4: Classification par partition*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 9: Analyses des proximites, des preferences et typologie*. Editions Technip, Paris.