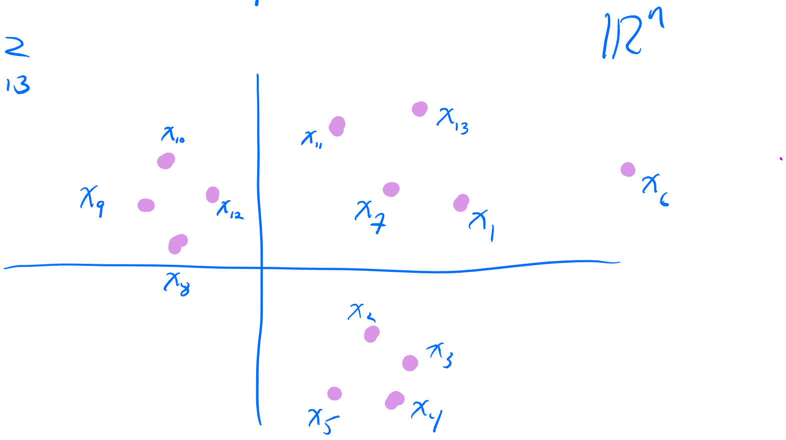
Hierarchical Trees & Clustering

Data $\underset{n \times p}{X} = [x_1, \dots, x_p]$

$x_i \in \mathbb{R}^n$

Ex $n=2$, $p=13$



#1 Key Idea Distance between observations

$$\boxed{dist(x_i, x_j)}$$

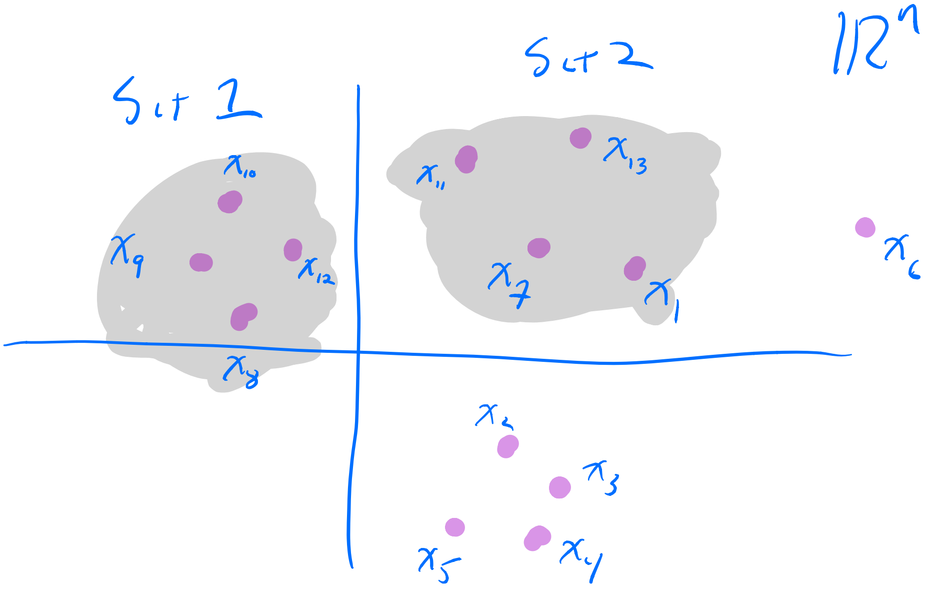
Examples

- Euclidean Distance in $\mathbb{R}^n$
- Correlation-based

correlation ≈ 1 ⇒ distance small

correlation ≈ 0 ⇒ distance large

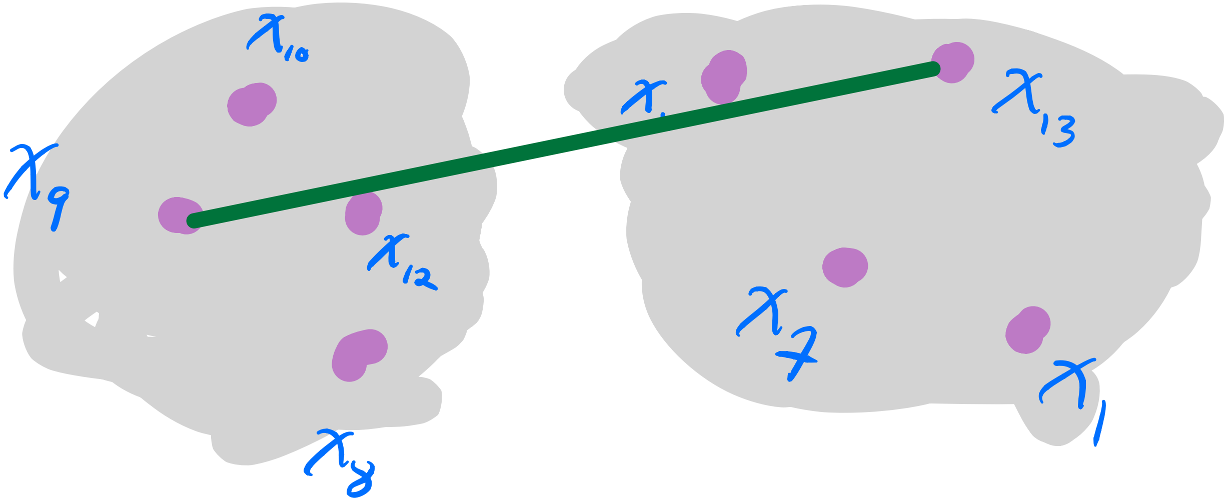
#2 Key Idea Distance between sets of observations



dist (Set 1, Set 2) ?

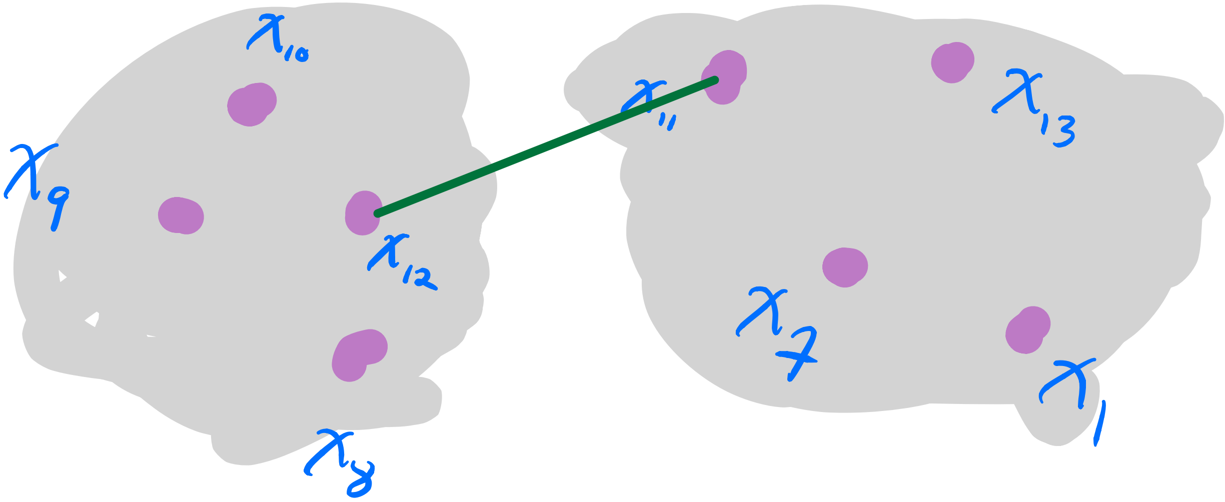
Complete

Maximal Pair-wise Distance



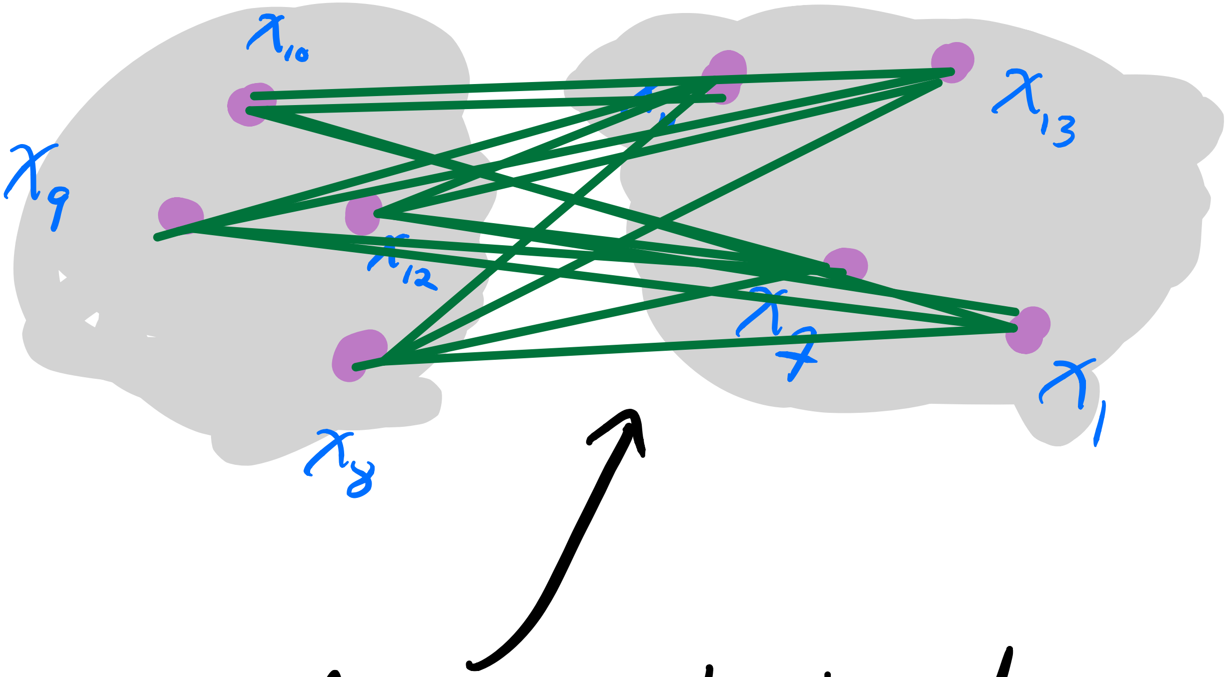
Single

Minimal Pair-wise Distance



Average

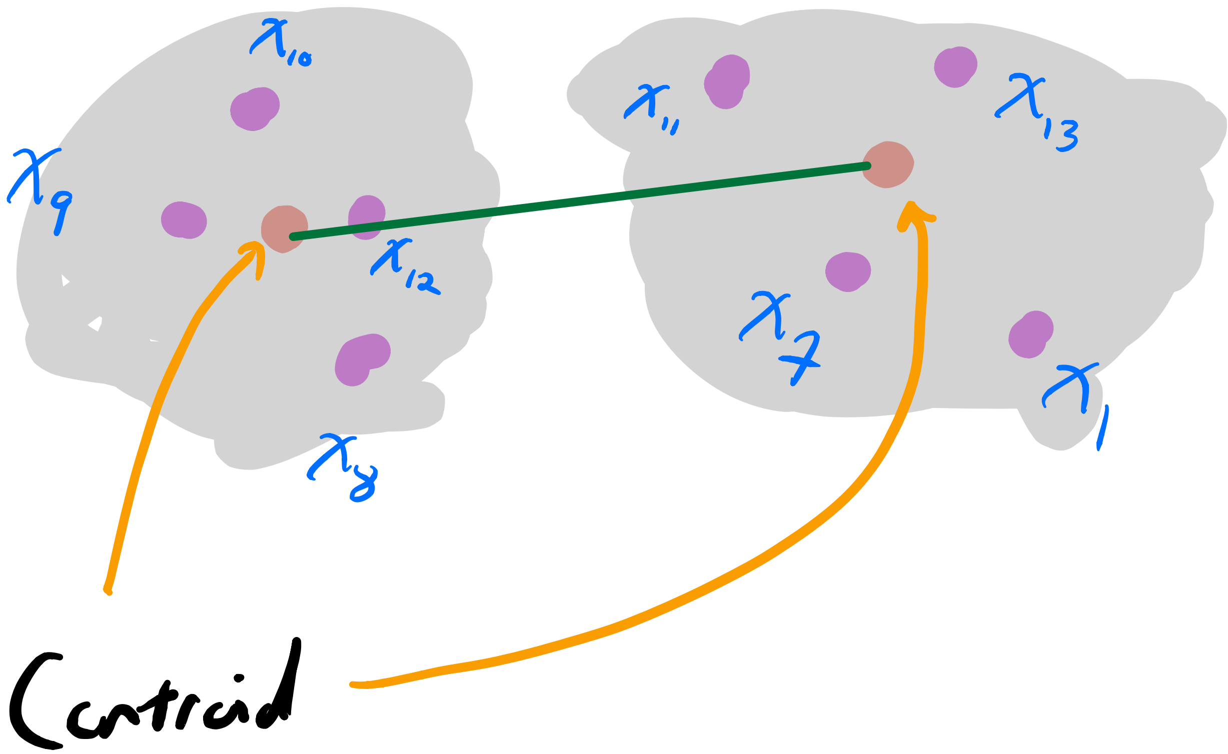
Average of Pair-wise Distances



Average all these!

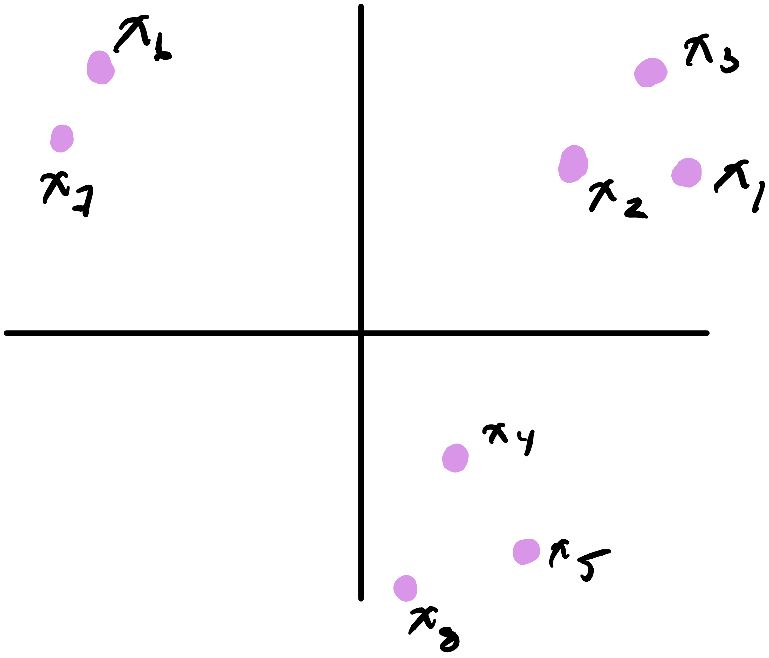
Centroid

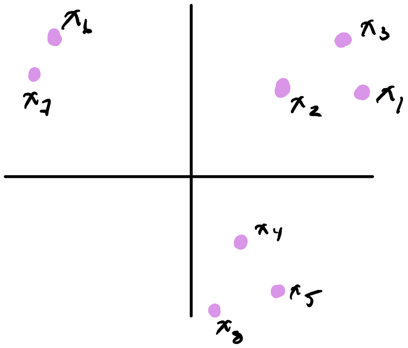
Distance between Centroids



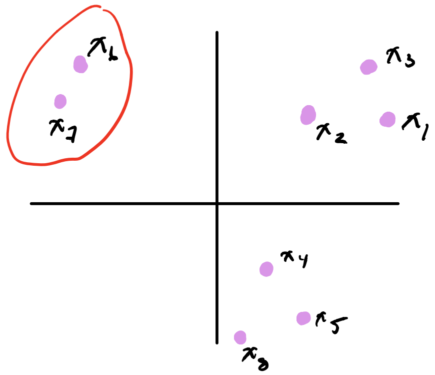
Building Hierarchical Tree

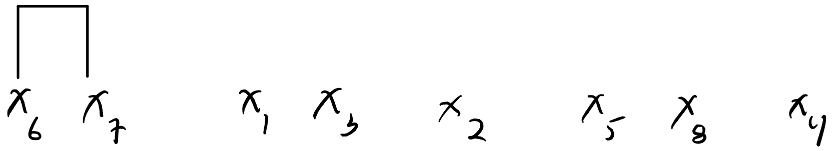
Euclidian Distance
Central Method

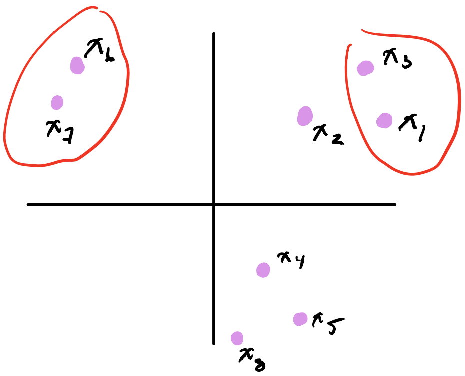


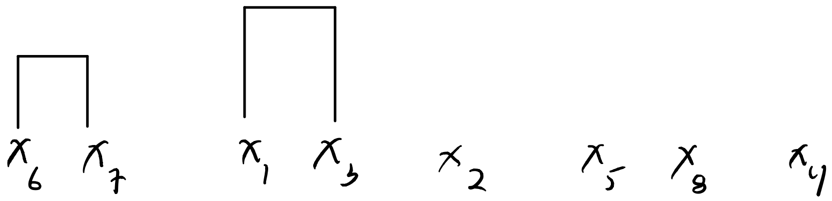


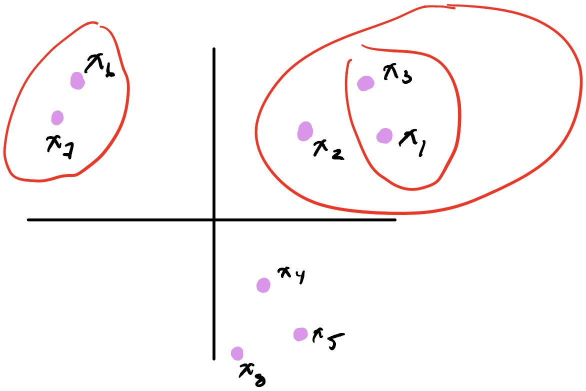
x_6 x_7 x_1 x_3 x_2 x_5 x_8 x_4

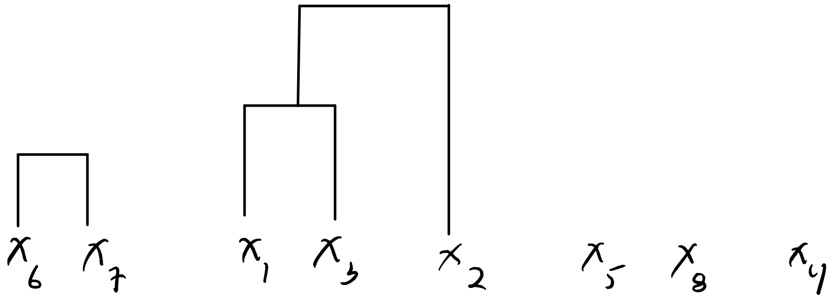

x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8

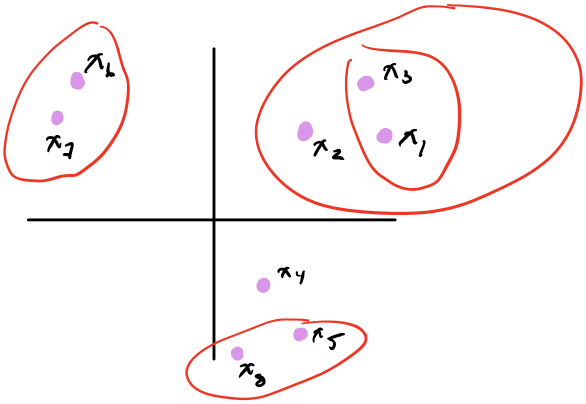

x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4

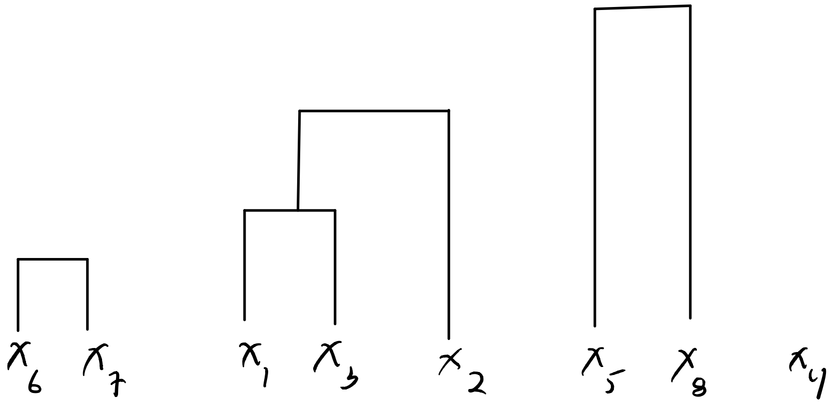

x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8

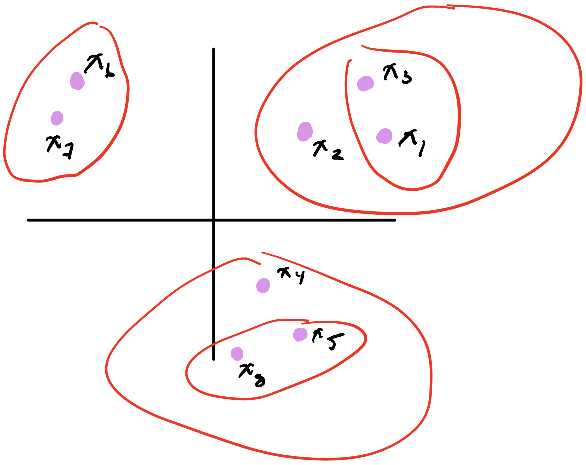

x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4

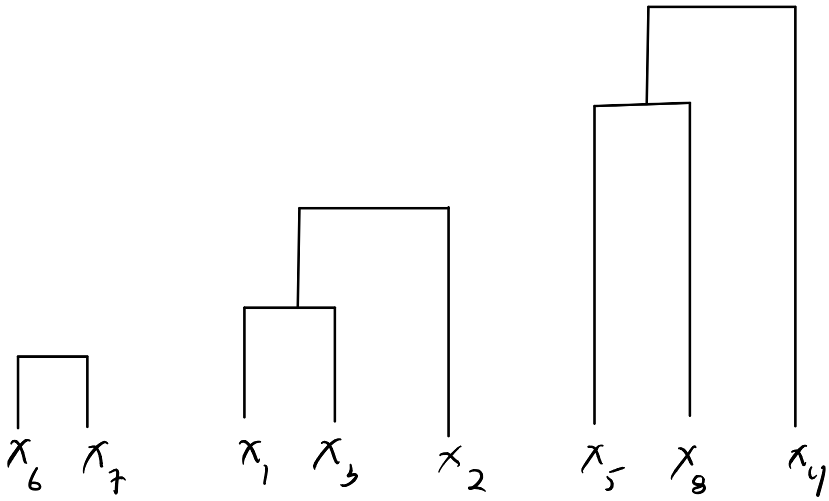

x_6
x_3
x_7
x_2
x_1
x_4
x_5
x_8

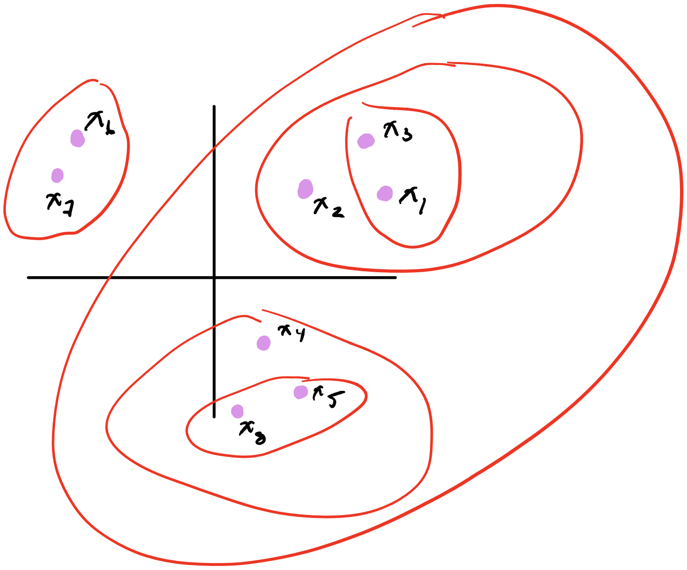

x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4


x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8

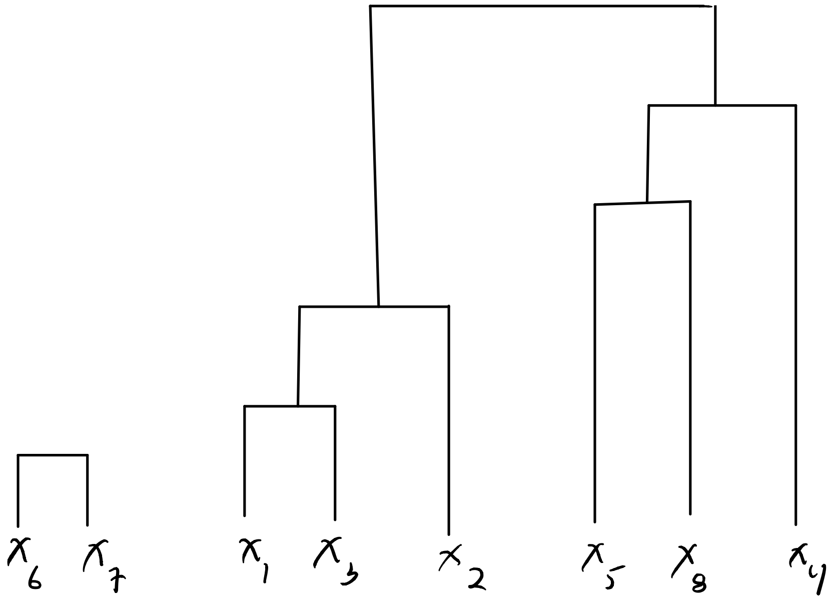

x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4


x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8

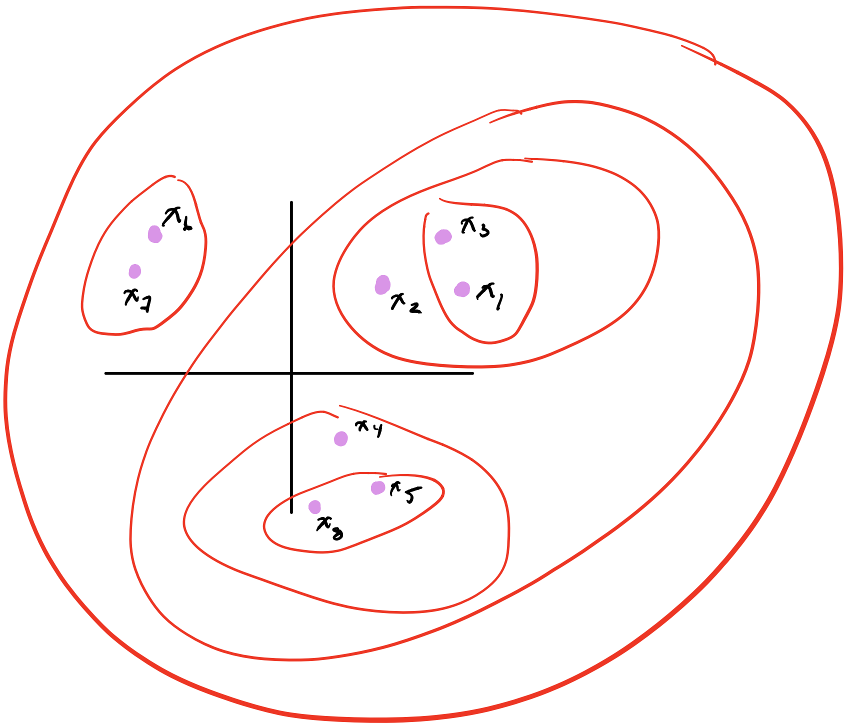

x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4

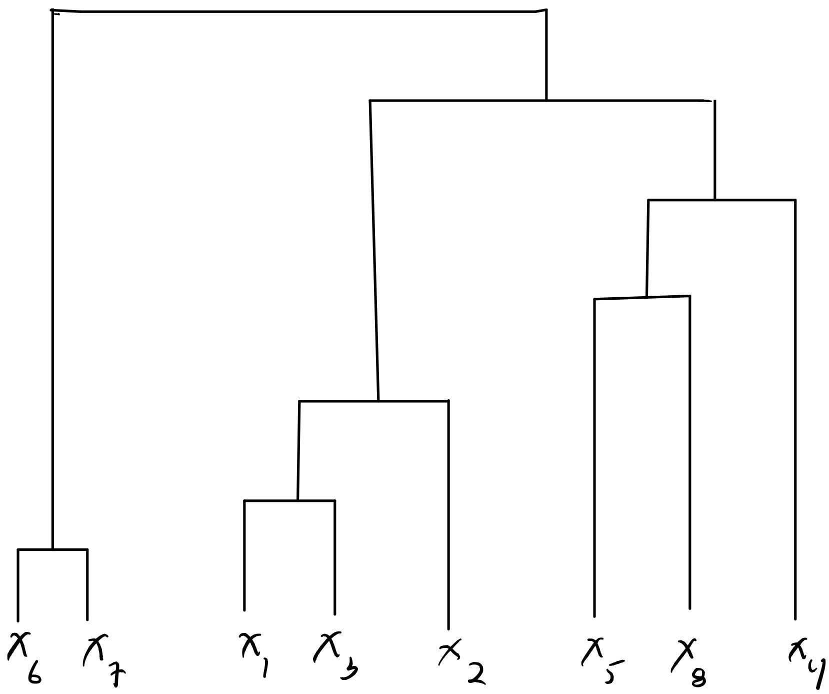


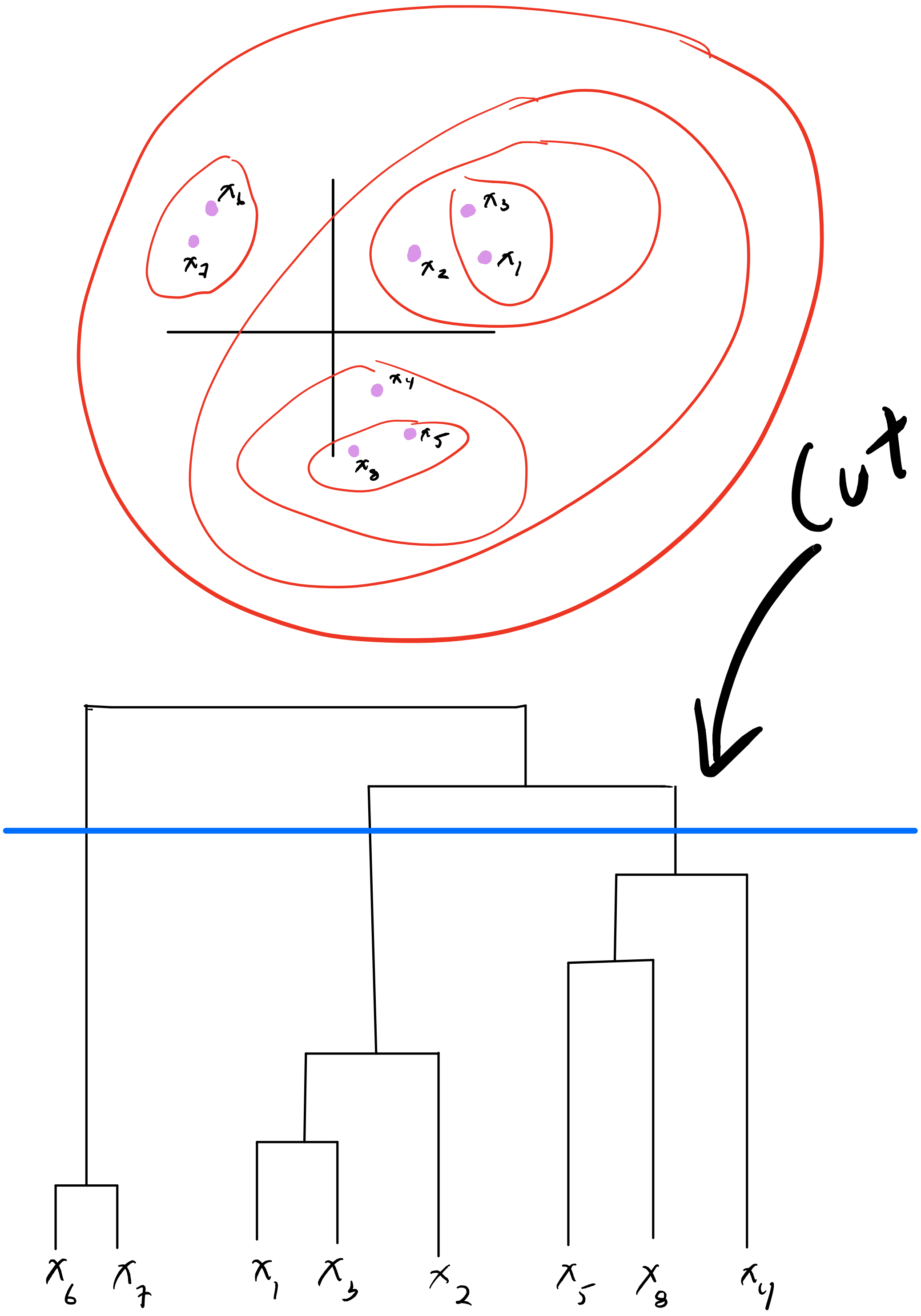
x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8

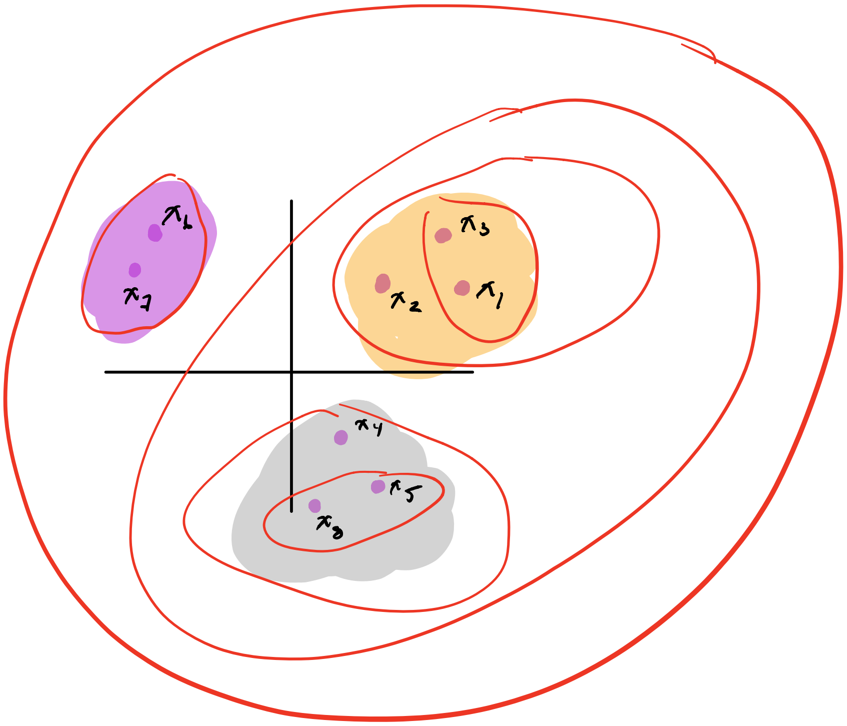


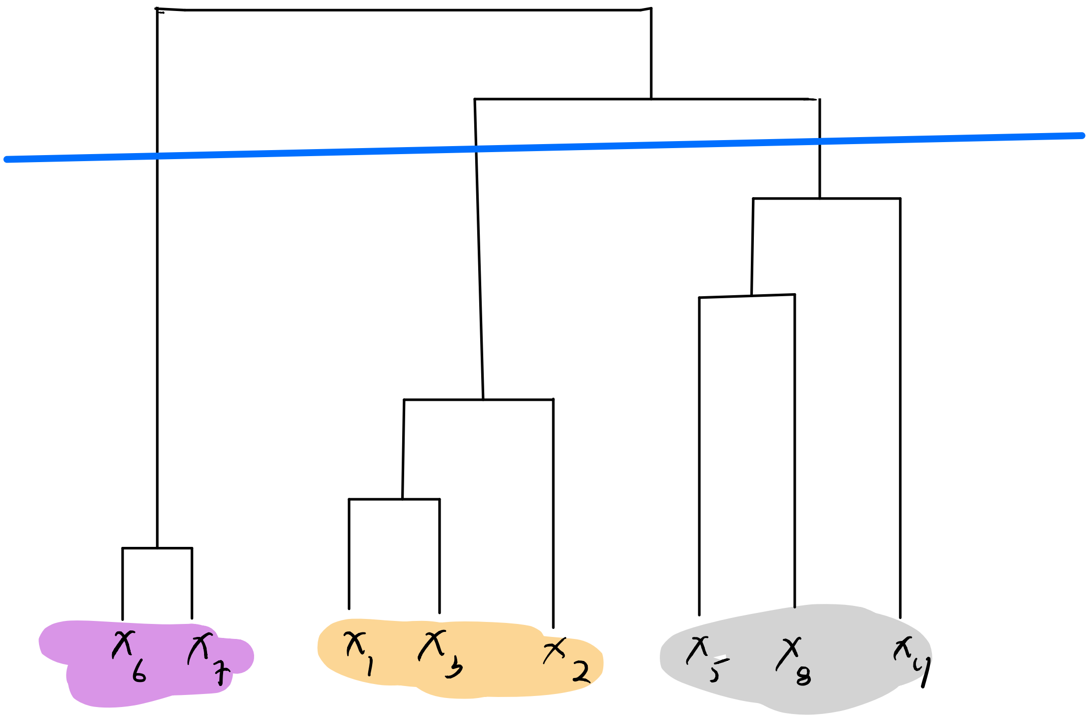
x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4


x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8


x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4


x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8
Cut
x_6 x_7 x_1 x_3 x_2 x_5 x_8 x_4


x_6
x_7
x_3
x_2
x_1
x_4
x_5
x_8


x_6
x_7
x_1
x_3
x_2
x_5
x_8
x_4

Hierarchical Clustering

- Intuitive
- Visual

  Works in any dimension
- Sensitive to choices
  - Distance
  - Cluster type (e.g. Centroid vs Average)