STAT 391

5/30/23

Lecture 19

Clustering
- kmeans

L9 Clustering
HW 6
Sols 5, 6

Lecture Notes IX – Clustering

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

May, 2023

Paradigms for clustering

Parametric clustering algorithms (K given)

- Cost based / hard clustering
- K-means clustering and the quadratic distortion
- Model based / soft clustering

Issues in parametric clustering

- Selecting K

Reading: Ch. 18

What is clustering? Problem and Notation

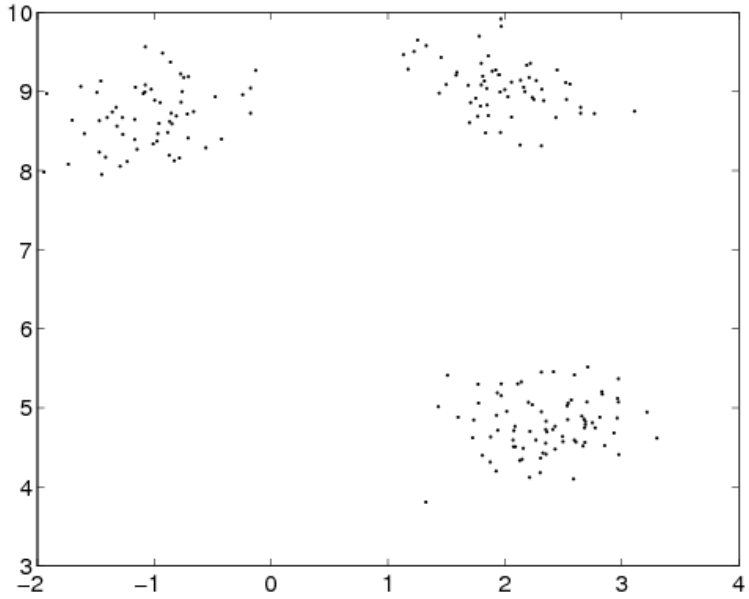
- **Informal definition** **Clustering** = Finding groups in data
- **Notation**

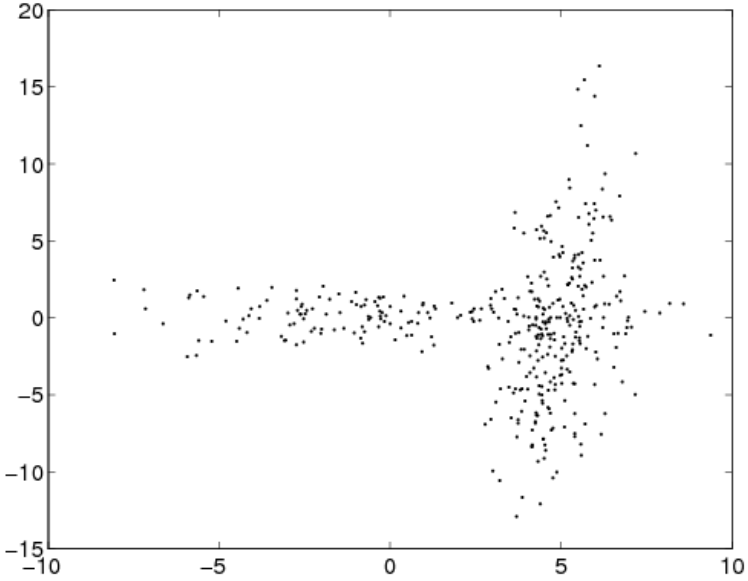
$\mathcal{D}$	=	$\{x_1, x_2, \ldots x_n\}$ a **data set**
n	=	number of **data points**
K	=	number of **clusters** ($K << n$)
Δ	=	$\{C_1, C_2, \ldots, C_K\}$ a partition of $\mathcal{D}$ into disjoint subsets
$k(i)$	=	the **label** of point i
$\mathcal{L}(\Delta)$	=	cost (loss) of Δ (to be minimized)

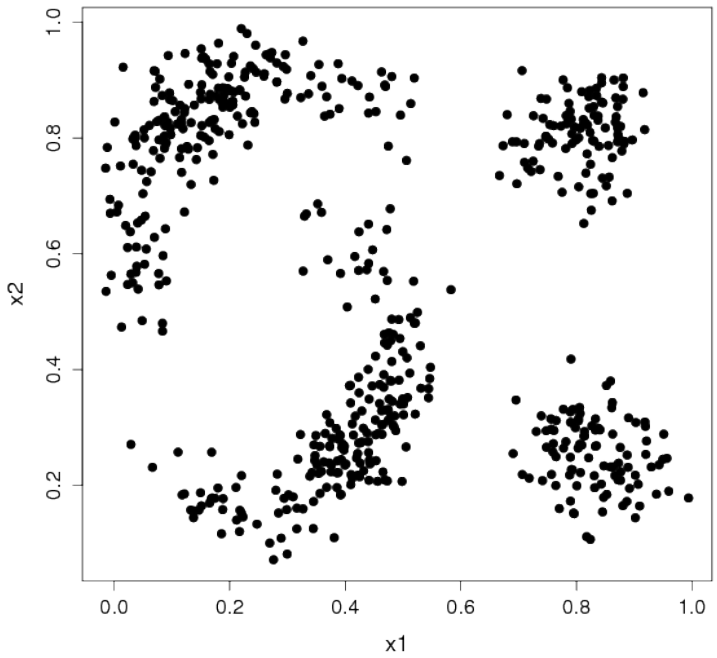
- **Second informal definition** **Clustering** = given n **data points**, separate them into K **clusters**
- Hard vs. soft clusterings
  - **Hard** clustering Δ: an item belongs to only 1 cluster
  - **Soft** clustering $\gamma = \{\gamma_{ki}\}_{k=1:K}^{i=1:n}$
    γ_{ki} = the **degree of membership** of point i to cluster k

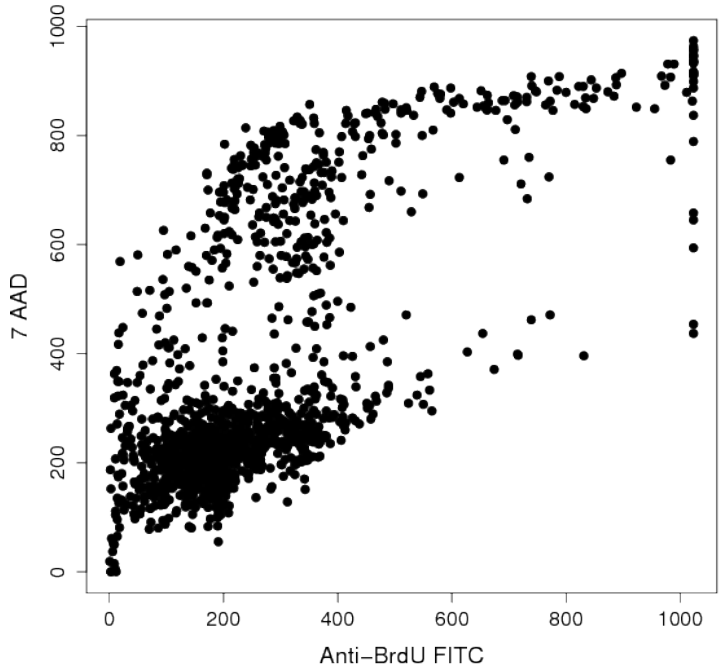
$$\sum_k \gamma_{ki} = 1 \quad \text{for all } i$$

(usually associated with a probabilistic model)


10
9
8
7
6
5
4
3
−2
−1
0
1
2
3
4

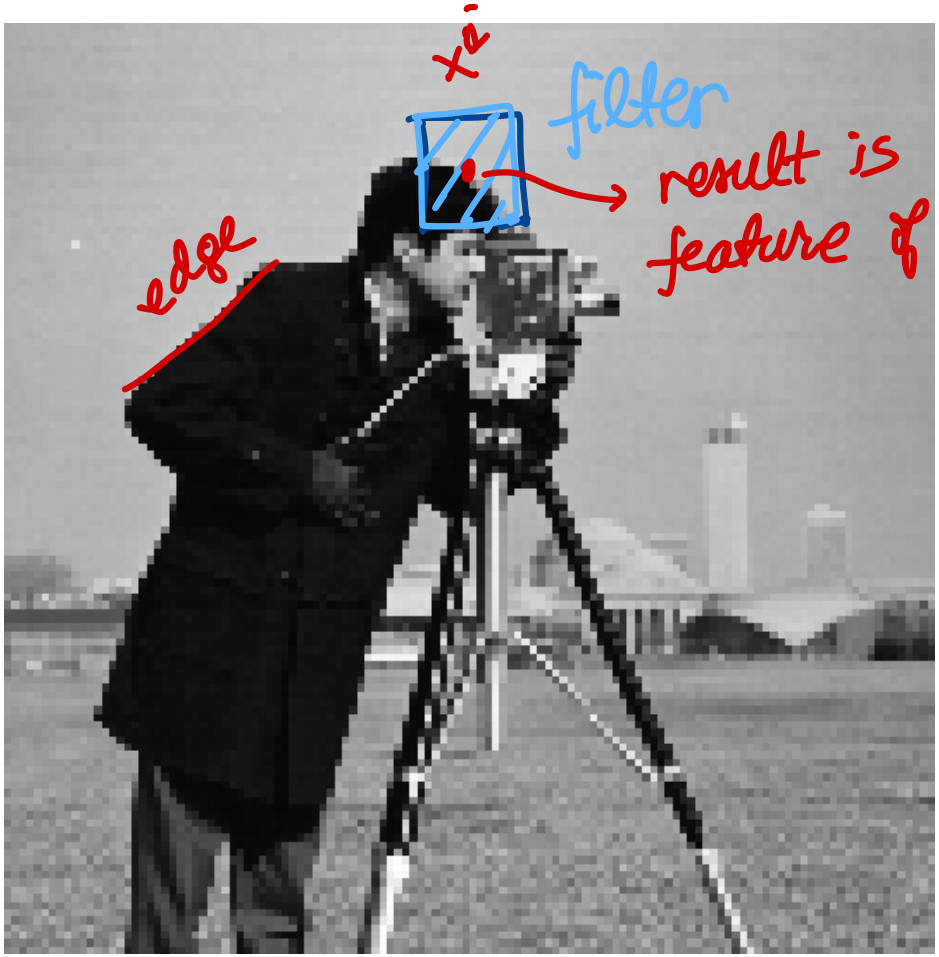

20
15
10
5
0
−5
−10
−15
−10
−5
0
5
10


x2
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
x1


7 AAD
1000
800
600
400
200
0
0
200
400
600
800
1000
Anti−BrdU FITC

(from)

from Carreira-Perpinan, 2006



step 0

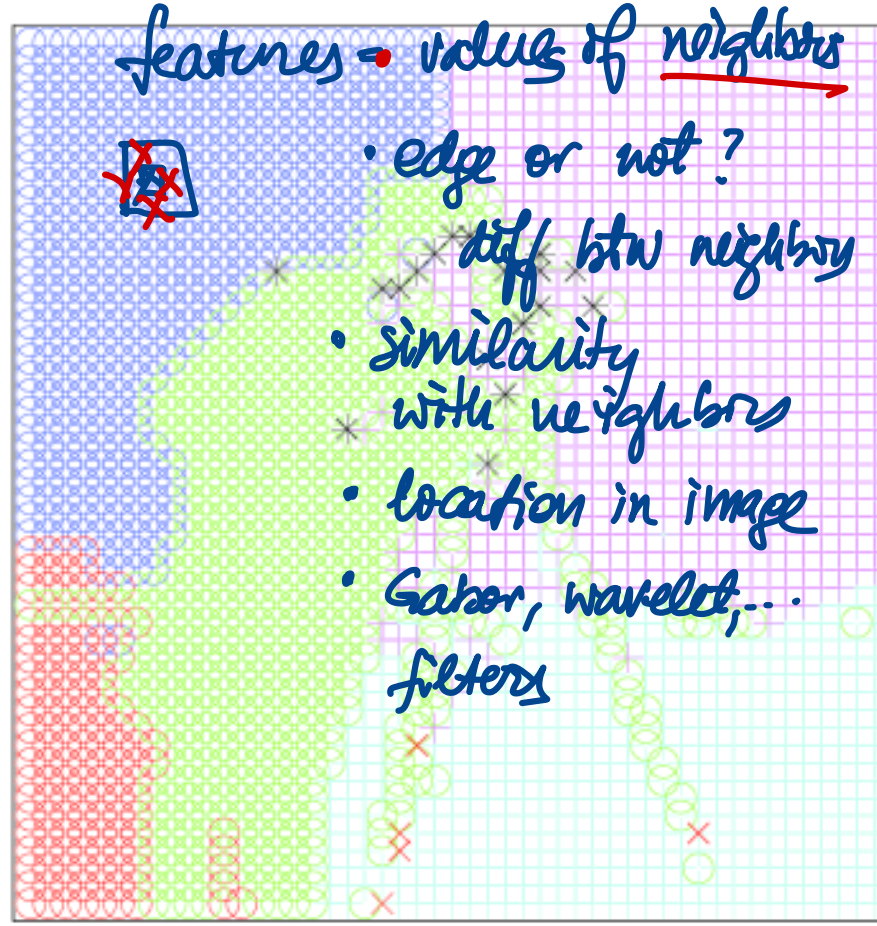


image segmentation

n = #pixels

$K = 5$

x^i = pixel i = ? 1) grey value $\in [0, 255]$

or 2) other features

Paradigms

Depend on type of data, type of clustering, type of cost (probabilistic or not), and constraints (about K, shape of clusters)

- Data = vectors $\{x_i\}$ in $\mathbb{R}^d$

Parametric (K known)	Cost based [hard]	→ k-means
	Model based [soft]	→ Mixture models
Non-parametric (K determined by algorithm)	Dirichlet process mixtures [soft]	
	Information bottleneck [soft]	
	Modes of distribution [hard]	
	Gaussian blurring mean shift [hard]	

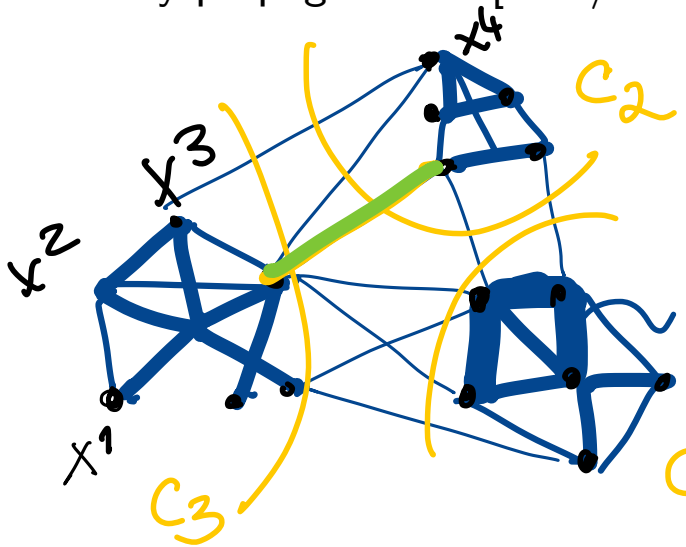
$x^i \in$ unique C^k

$x^i \in C_k$ probabilistically

- Data = similarities between pairs of points $[S_{ij}]_{i,j=1:n}$, $S_{ij} = S_{ji} \geq 0$ **Similarity based clustering**

Graph partitioning	spectral clustering [hard, K fixed, cost based]
	typical cuts [hard non-parametric, cost based]
Affinity propagation	[hard/soft non-parametric]

image segmentation



- similar points in same cluster
- robust to adding "noise" edges

Parametric clustering algorithms

- Cost based
  - Single linkage (min spanning tree)
  - Min diameter
    - Fastest first traversal (HS initialization)
  - K-medians
  - K-means
- Model based (cost is derived from likelihood)
  - EM algorithm
  - "Computer science"/"Probably correct" algorithms

[Supplement: Single Linkage Clustering]

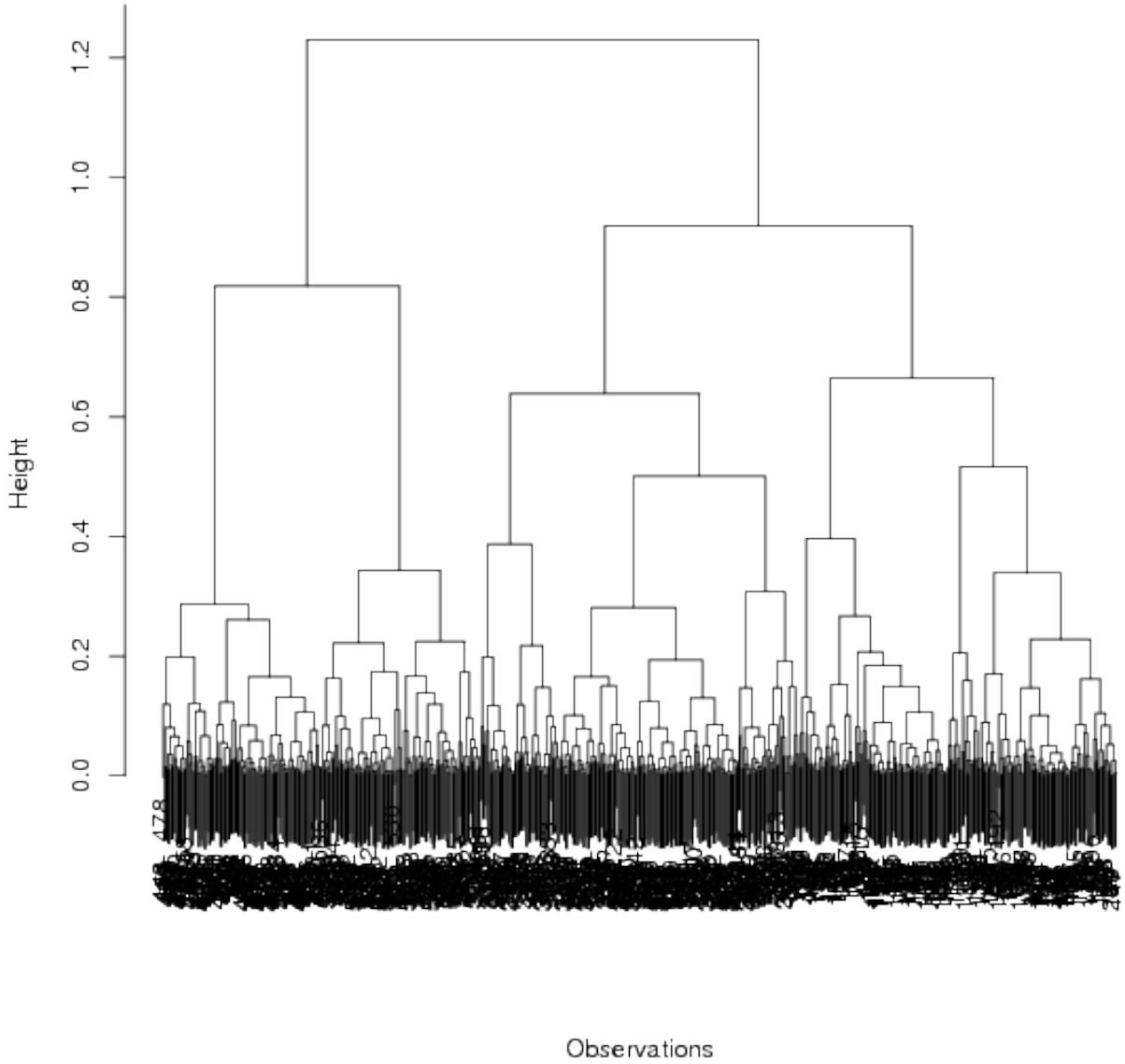
Algorithm Single-Linkage

Input Data $\mathcal{D} = \{x_i\}_{i=1:n}$, number clusters K

1. Construct the Minimum Spanning Tree (MST) of $\mathcal{D}$
2. Delete the largest $K-1$ edges

- **Cost** $\mathcal{L}(\Delta) = -\min_{k,k'} \text{distance}(C_k, C_{k'})$
  where $\text{distance}(A, B) = \underset{x \in A,\, y \in B}{\text{argmin}} ||x - y||$
- Running time $\mathcal{O}(n^2)$ one of the very few costs $\mathcal{L}$ that can be optimized in **polynomial** time
- Sensitive to outliers!

[Supplement: Single Linkage Clustering]



[Supplement: Minimum diameter clustering]

- **Cost** $\mathcal{L}(\Delta) = \max_k \underbrace{\max_{i,j \in C_k} ||x_i - x_j||}_{\text{diameter}}$
  - Mimimize the diameter of the clusters
  - Optimizing this cost is NP-hard
- **Algorithms**
  - **Fastest First Traversal** – a factor 2 approximation for the min cost

For every $\mathcal{D}$, FFT produces a Δ so that

$$\mathcal{L}^{opt} \leq \mathcal{L}(\Delta) \leq 2\mathcal{L}^{opt}$$

  - rediscovered many times

[Supplement: Minimum diameter clustering]

Algorithm Fastest First Traversal

Input Data $\mathcal{D} = \{x_i\}_{i=1:n}$, number clusters K

defines **centers** $\mu_{1:K} \in \mathcal{D}$

(many other clustering algorithms use centers)

1. pick μ_1 at random from $\mathcal{D}$
2. for $k = 2 : K$

$$\mu_k \leftarrow \operatorname*{argmax}_{\mathcal{D}} \text{distance}(x_i, \{\mu_{1:k-1}\})$$

3. for $i = 1 : n$ (assign points to centers)

$k(i) = k$ if μ_k is the nearest center to x_i

[Supplement: K-medians clustering]

- **Cost** $\mathcal{L}(\Delta) = \sum_k \sum i \in C_k ||x_i - \mu_k||$ with $\mu_k \in \mathcal{D}$
  - (usually) assumes centers chosen from the data points (analogy to median)

Exercise Show that in 1D $\underset{\mu}{\operatorname{argmin}} \sum_i |x_i - \mu|$ is the median of $\{x_i\}$

  - optimizing this cost is NP-hard

  - has attracted a lot of interest in theoretical CS (general from called "Facility location"

K-means clustering

Algorithm K-Means

Input Data $\mathcal{D} = \{x_i\}_{i=1:n}$, number clusters K

tialize centers $\mu_1, \mu_2, \ldots \mu_K \in \mathbb{R}^d$ at random ← initial guess

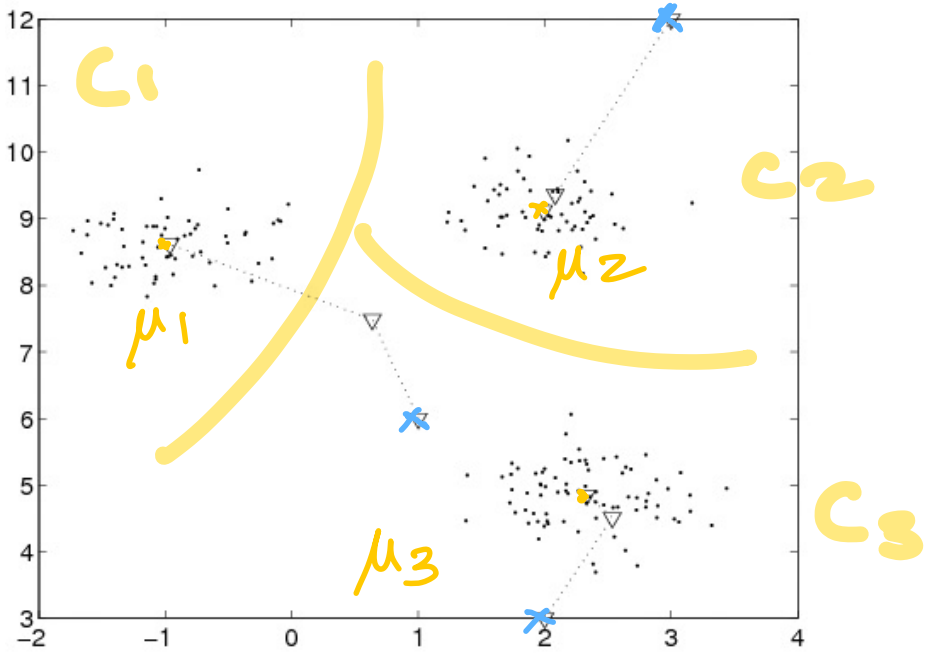
terate until convergence

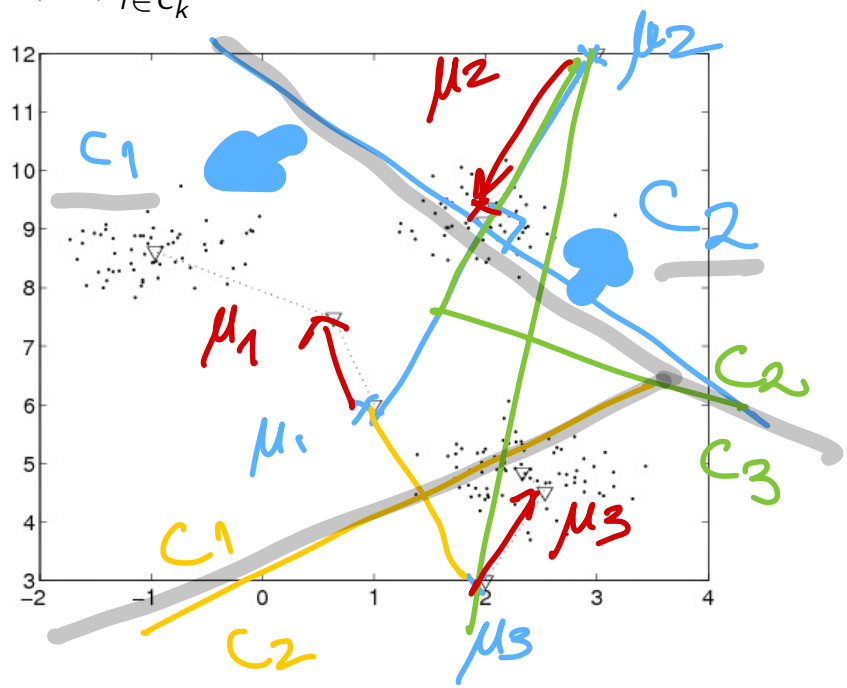
1. for $i = 1 : n$ (assign points to clusters ⇒ new clustering)

$$k(i) = \operatorname*{argmin}_k ||x_i - \mu_k|| \quad \leftarrow \text{assign } x_i \text{ to nearest } \mu_k$$

2. for $k = 1 : K$ (recalculate centers)

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \tag{1}$$





K-means clustering

Algorithm K-Means

Input Data $\mathcal{D} = \{x_i\}_{i=1:n}$, number clusters K

tialize **centers** $\mu_1, \mu_2, \ldots \mu_K \in \mathbb{R}^d$ at random

terate until convergence

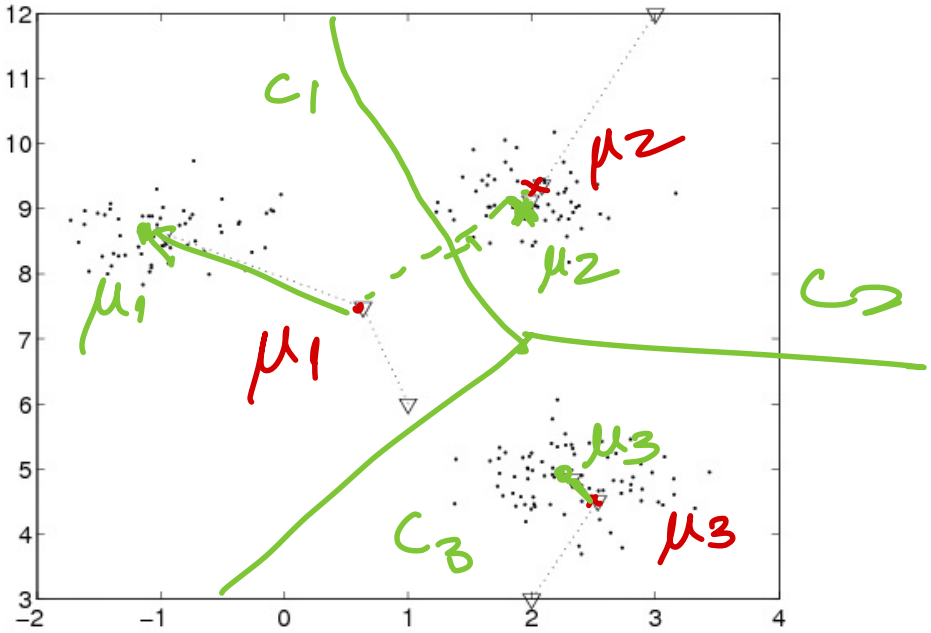
1. for $i = 1 : n$ (assign points to clusters $\Rightarrow$ new clustering)

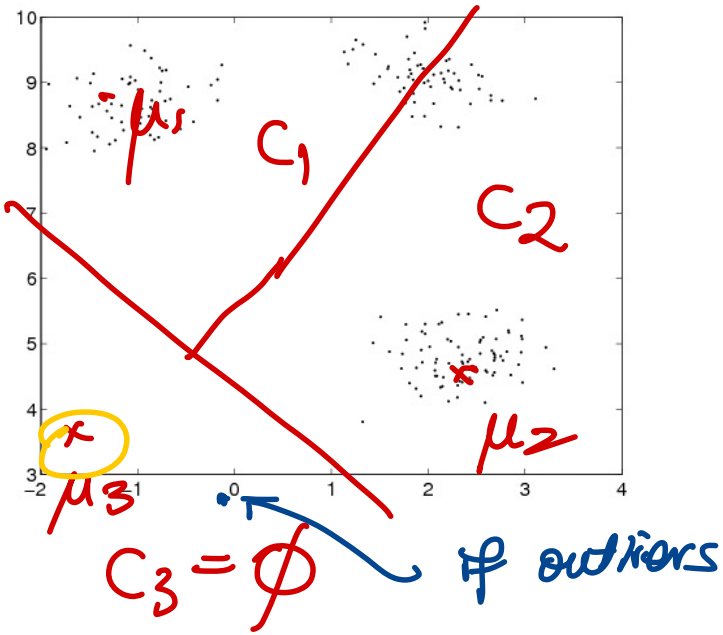
$$k(i) = \operatorname*{argmin}_k ||x_i - \mu_k||$$

2. for $k = 1 : K$ (recalculate centers)

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

- convergence if assignments don't change (1)





The K-means cost = loss (= "risk") • how good is Δ? • $\mathcal{L}(\Delta)$ smaller is better

$$\mathcal{L}(\Delta) = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2 \tag{2}$$

clustering

$\Delta = \{C_1, \dots C_K\}$

$n_k = |C_k|$

$n_1 + \dots n_K = n$

$\sum_k n_k \operatorname{Var} C_k$

- K-means solves a **least-squares** problem
- the cost $\mathcal{L}$ is called **quadratic distortion** = loss

Proposition The K-means algorithm decreases $\mathcal{L}(\Delta)$ at every step.

Sketch of proof

- step 1: reassigning the labels can only decrease $\mathcal{L}$
- step 2: reassigning the centers μ_k can only decrease $\mathcal{L}$ because μ_k as given by (1) is the solution to

$$\mu_k = \min_{\mu \in \mathbb{R}^d} \sum_{i \in C_k} ||x_i - \mu||^2 \tag{3}$$

Prop $\mathcal{L}(\Delta^t) \leq \mathcal{L}(\Delta^{t-1})$

Δ^t = clustering at iteration t

$t=0$ initialization

Algorithm

Init $\mu_1^0, \mu_2^0, \dots \mu_k^0 \Rightarrow \Delta^1 : \|x^i - \mu_k^0\|^2 \leq \|x^i - \mu_{k'}^0\| \iff x_i \in C_k$

$t=1$: $\mu_1^1 \leftarrow \frac{\sum_{i \in C_k} x_i}{n_1}$ $\quad \sum_{i \in C_1} \|x_i - \mu_1\|^2 \leq \sum_{i \in C_1} \|x_i - a\|^2$ for any a

($a = \mu_1^0$)

(recalculate $\mu_{1:k}$)

$$\mathcal{L}(\Delta^1) = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k^1\|^2 \leq \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu^0\|^2 = \mathcal{L}(\Delta^1, \mu_{1:k}^0)$$

$\mathcal{L}(\Delta^1, \mu_{1:k}^1)$

Δ^1: assignments; $\mu_{1:k}^0$: centers

$t=2$ assign $x_{1:n}$ to clusters $(\mu_{1:k}^1) \Rightarrow$ obtain $\Delta^2 \leftarrow$ from $\mu_{1:k}^1$

for x_i : same C_k $\quad \|x_i - \mu_k\|^2 = \min_{k'=k} \|x_i - \mu_{k'}\|^2$

Δ^1: $x_i \in C_k$ $\quad$ another $C_{k'} \iff \|x_i - \mu_{k'}'\|^2 \leq \|x_i - \mu_k'\|^2$

$\mu_k' \neq \mu_k^0$ $\quad k' = k$

$k(i)$ cluster of x_i

decreases!

if $\Delta^1 \to \Delta^2$ $\quad \mathcal{L}(\Delta^2, \mu_{1:k}^1) = \sum_{\substack{i \text{ don't} \\ \text{change cluster}}} \|x_i - \mu_{k(i)}^1\|^2 + \sum_{\substack{i \text{ change} \\ \text{cluster } C_k}} \|x_i - \mu_{k(i)}\|^2$

$\Rightarrow \quad \mathcal{L}(\Delta^2, \mu^1_{1:k}) \leq \mathcal{L}(\Delta^1, \mu^1_{1:k})$

$< \text{ if } \Delta^1 \neq \Delta^2$

$t=3$ recalculate μ's : $\mu^2_{1:k} \leftarrow \ldots$

$\mathcal{L}(\Delta^2, \mu^2_{1:k}) \leq \mathcal{L}(\Delta^2, \mu^1_{1:k})$

recalculate Δ

$\mathcal{L}(\Delta^3, \mu^3_{1:k}) \leq \mathcal{L}(\Delta^2, \mu^2_{1:k})$

...

$\Rightarrow$ Convergence $\Leftrightarrow \mathcal{L}(\Delta, \mu)$ doesn't change

- This is LOCAL MIN

Δ is best only w.r.t small change

↑ at convergence

- How to choose Δ after several initializations?

for $b = 1:B$ Initialize $\mu_{1:K}$

  Run K-means algorithm $\Rightarrow \Delta^{(b)}, \mu_{1:K}^{(b)}$

Output $\underset{b=1:B}{\operatorname{argmin}} \mathcal{L}(\Delta^{(b)}, \mu_{1:K}^{(b)})$

How to guess $\mu_{1:K}^{0}$ initially?

Choose K ← heuristics

[Supplement: Equivalent and similar cost functions]

- The distortion can also be expressed using intracluster distances

$$\mathcal{L}(\Delta) = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} ||x_i - x_j||^2 \tag{4}$$

- **Correlation clustering** is defined as optimizing the related criterion

$$\mathcal{L}(\Delta) = \sum_{k=1}^{K} \sum_{i,j \in C_k} ||x_i - x_j||^2$$

- This cost is equivalent to the (negative) sum of (squared) intercluster distances

$$\mathcal{L}(\Delta) = -\sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \notin C_k} ||x_i - x_j||^2 + \text{constant} \tag{5}$$

Proof of (6) Replace μ_k as expressed in (1) in the expression of $\mathcal{L}$, then rearrange the terms

Proof of (5) $\sum_k \sum_{i,j \in C_k} ||x_i - x_j||^2 = \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{n} ||x_i - x_j||^2}_{\text{independent of } \Delta} - \sum_k \sum_{i \in C_k} \sum_{j \notin C_k} ||x_i - x_j||^2$

[Supplement: The K-means cost in matrix form – the assignment matrix]

- $\mathcal{L}$ as sum of squared intracluster distances

$$\mathcal{L}(\Delta) = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} ||x_i - x_j||^2 \tag{6}$$

-
- Define the **assignment matrix** associated with Δ by $Z(\Delta)$
  Let $\Delta = \{C_1 = \{1,2,3\},\ C_2 = \{4,5\}\}$

$$Z^{unnorm}(\Delta) = \begin{matrix} \\ \text{point } 1 \\ \text{point } 2 \\ \text{point } 3 \\ \text{point } 4 \\ \text{point } 5 \end{matrix} \begin{matrix} C_1 & C_2 \end{matrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad Z(\Delta) = \begin{matrix} C_1 & C_2 \end{matrix} \begin{bmatrix} 1/\sqrt{3} & 0 \\ 1/\sqrt{3} & 0 \\ 1/\sqrt{3} & 0 \\ 0 & 1/\sqrt{2} \\ 0 & 1/\sqrt{2} \end{bmatrix}$$

Then Z is an orthogonal matrix (columns are orthornormal) and

$$\mathcal{L}(\Delta) = \operatorname{trace} Z^T D Z \qquad \text{with } D_{ij} = ||x_i - x_j||^2 \tag{7}$$

$$\text{Let } \mathcal{Z} = \{ Z \in \mathbb{R}^{n \times K},\ K \text{ orthonormal} \}$$

Proof of (7) Start from (2) and note that $\operatorname{trace} Z^T A Z = \sum_k \sum_{i,j \in C_k} Z_{ik} Z_{jk} A_{ij} = \sum_k \sum_{i,j \in C_k} \frac{1}{|C_k|} A_{ij}$

[Supplement: The K-means cost in matrix form – the co-ocurrence matrix]

$$n = 5,\ \Delta = (1,1,1,2,2), \qquad X(\Delta) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

1. $X(\Delta)$ is symmetric, positive definite, ≥ 0 elements
2. $X(\Delta)$ has row sums equal to 1
3. trace $X(\Delta) = K$

$$\|X(\Delta)\|_F^2 = \langle X, X \rangle = K$$
$$X(\Delta) = Z(\Delta)Z^T(\Delta)$$

$$2\mathcal{L}(\Delta) = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} ||x_i - x_j||^2 = \frac{1}{2}\langle D, X(\Delta) \rangle$$

with $D_{ij} = ||x_i - x_j||^2$

[Supplement: Symmetries between costs]

- K-means cost $\mathcal{L}(\Delta) = \min_{\mu_{1:K}} \sum_k \sum_{i \in C_k} ||x_i - \mu_k||^2$
- K-medians cost $\mathcal{L}(\Delta) = \min_{\mu_{1:K}} \sum_k \sum_{i \in C_k} ||x_i - \mu_k||$

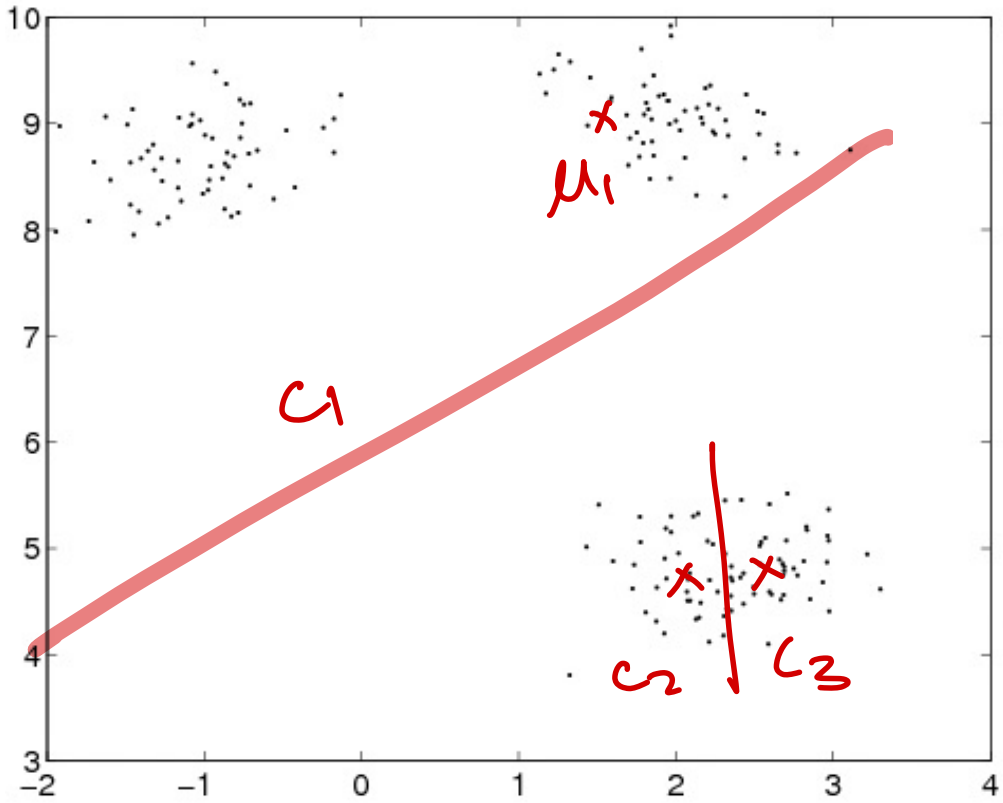
- Correlation clustering cost $\mathcal{L}(\Delta) = \sum_k \sum_{i,j \in C_k} ||x_i - x_j||^2$
- min Diameter cost $\mathcal{L}^2(\Delta) = \max_k \max_{i,j \in C_k} ||x_i - x_j||^2$

Initialization of the centroids $\mu_{1:K}$

- Idea 1: start with K points at random

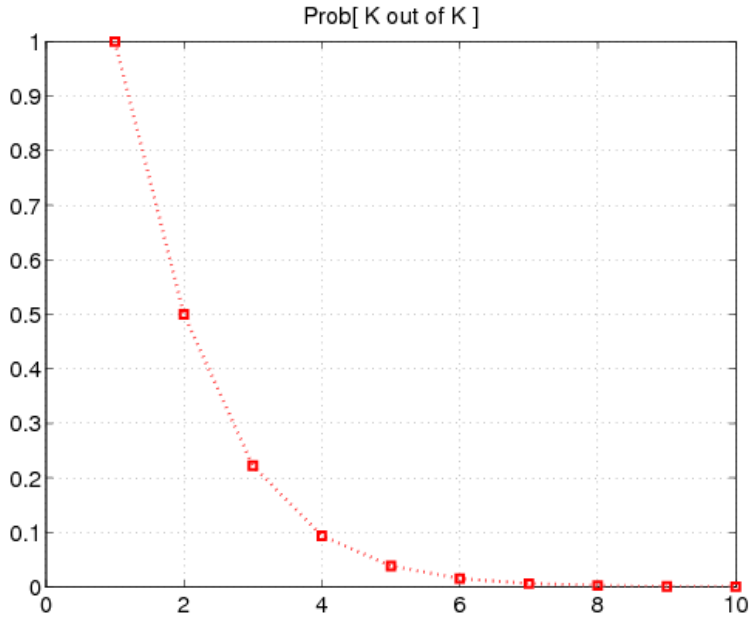
Initialization of the centroids $\mu_{1:K}$

- Idea 1: start with K points at random ← Pb: μ_k^0 may be far away from data
- Idea 2: start with K data points at random



Initialization of the centroids $\mu_{1:K}$

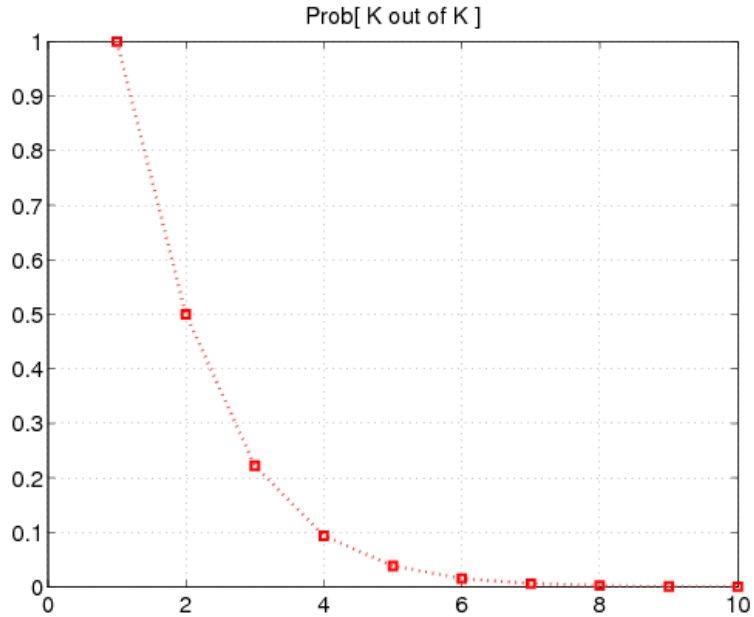
- Idea 1: start with K points at random
- Idea 2: start with K data points at random
  What's wrong with chosing K data points at random?



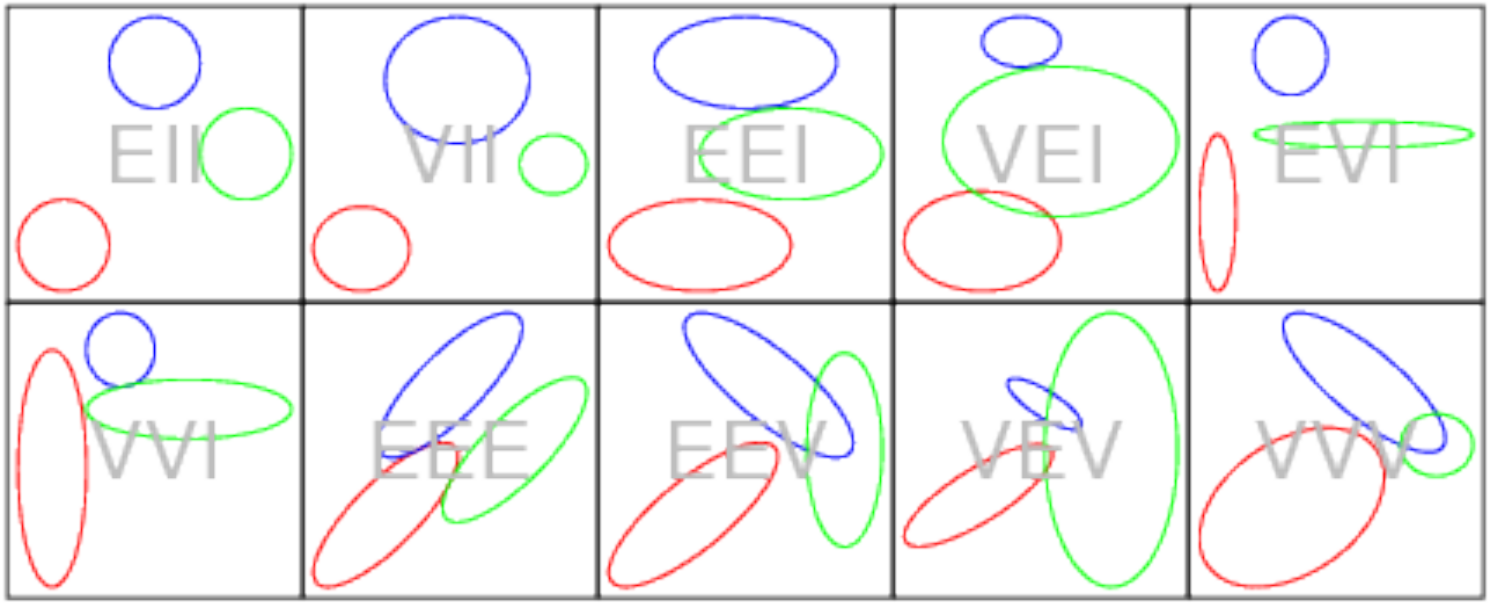
The probability of hitting all K clusters with K samples approaches 0 when $K > 5$

Initialization of the centroids $\mu_{1:K}$

- Idea 1: start with K points at random
- Idea 2: start with K data points at random
  What's wrong with chosing K data points at random?



  The probability of hitting all K clusters with K samples approaches 0 when $K > 5$
- Idea 3: start with K data points using Fastest First Traversal (greedy simple approach to spread out centers)



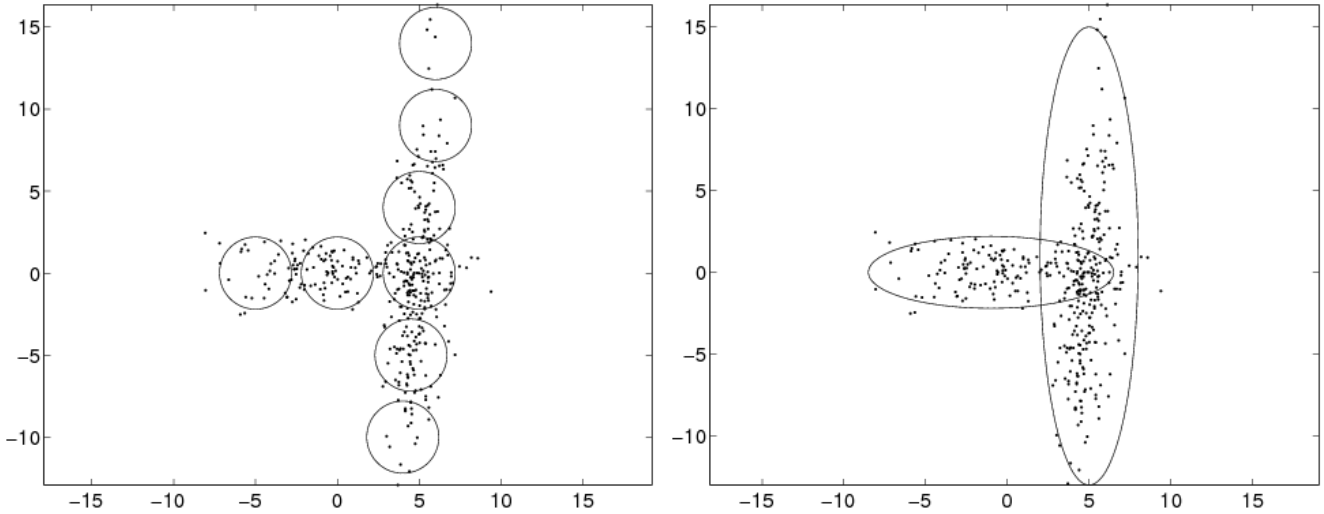
More special cases introduce the following description for a covariance matrice in terms of *volume, shape, alignment with axes* (=determinant, trace, e-vectors). The letters below mean: I=unitary (shape, axes), E=equal (for all k), V=unequal

- EII: equal volume, round shape (spherical covariance)
- VII: varying volume, round shape (spherical covariance)
- EEI: equal volume, equal shape, axis parallel orientation (diagonal covariance)
- VEI: varying volume, equal shape, axis parallel orientation (diagonal covariance)
- EVI: equal volume, varying shape, axis parallel orientation (diagonal covariance)
- VVI: varying volume, varying shape, equal orientation (diagonal covariance)
- EEE: equal volume, equal shape, equal orientation (ellipsoidal covariance)
- EEV: equal volume, equal shape, varying orientation (ellipsoidal covariance)
- VEV: varying volume, equal shape, varying orientation (ellipsoidal covariance)
- VVV: varying volume, varying shape, varying orientation (ellipsoidal covariance)

(from)

EM versus K-means

- Alternates between cluster assignments and parameter estimation
- Cluster assignments γ_{ki} are probabilistic
- Cluster parametrization more flexible



- Converges to local optimum of **log-likelihood**
  Initialization recommended by K-logK method

- **Modern algorithms with guarantees** (for e.g. mixtures of Gaussians)
  - Random projections
  - Projection on principal subspace
  - Two step EM (=K-logK initialization + one more EM iteration)

[Supplement: A two-step EM algorithm]

Similar to K-logK initialization for K-means

Assumes K spherical gaussians, separation $||\mu_k^{true} - \mu_{k'}^{true} \geq C\sqrt{d}\sigma_k$

1. Pick $K' = \mathcal{O}(K \ln K)$ centers μ_k^0 at random from the data
2. Set $\sigma_k^0 = \frac{d}{2} \min_{k \neq k'} ||\mu_k^0 - \mu_{k'}^0||^2$, $\pi_k^0 = 1/K'$
3. Run one E step and one M step $\Longrightarrow \{\pi_k^1, \mu_k^1, \sigma_k^1\}_{k=1:K'}$
4. Compute "distances" $d(\mu_k^1, \mu_{k'}^1) = \frac{||\mu_k^1 - \mu_{k'}^1||}{\sigma_k^1 - \sigma_{k'}^1}$
5. Prune all clusters with $\pi_k^1 \leq 1/4K'$
6. Run Fastest First Traversal with distances $d(\mu_k^1, \mu_{k'}^1)$ to select K of the remaining centers. Set $\pi_k^1 = 1/K$.
7. Run one E step and one M step $\Longrightarrow \{\pi_k^2, \mu_k^2, \sigma_k^2\}_{k=1:K}$

eorem For any $\delta, \varepsilon > 0$ if d large, n large enough, separation $C \geq d^{1/4}$ the **Two step EM** algorithm obtains centers μ_k so that

$$||\mu_k - \mu_k^{true}|| \leq ||\text{mean}(C_k^{true}) - \mu_k^{true}|| + \varepsilon\sigma_k\sqrt{d}$$

Selecting K

- Run clustering algorithm for $K = K_{min} : K_{max}$
  - obtain $\Delta_{K_{min}}, \ldots \Delta_{K_{max}}$ or $\gamma_{K_{min}}, \ldots \gamma_{K_{max}}$
  - choose best Δ_K (or γ_K) from among them
- Typically increasing $K \Rightarrow$ cost $\mathcal{L}$ decreases
  - ($\mathcal{L}$ cannot be used to select K)
  - Need to "penalize" $\mathcal{L}$ with function of number parameters

Selecting K for mixture models

The BIC (Bayesian Information) Criterion

- let θ_K = parameters for γ_K
- let $\#\theta_K$=number independent parameters in θ_K
  - e.g for mixture of Gaussians with full Σ_k's in d dimensions

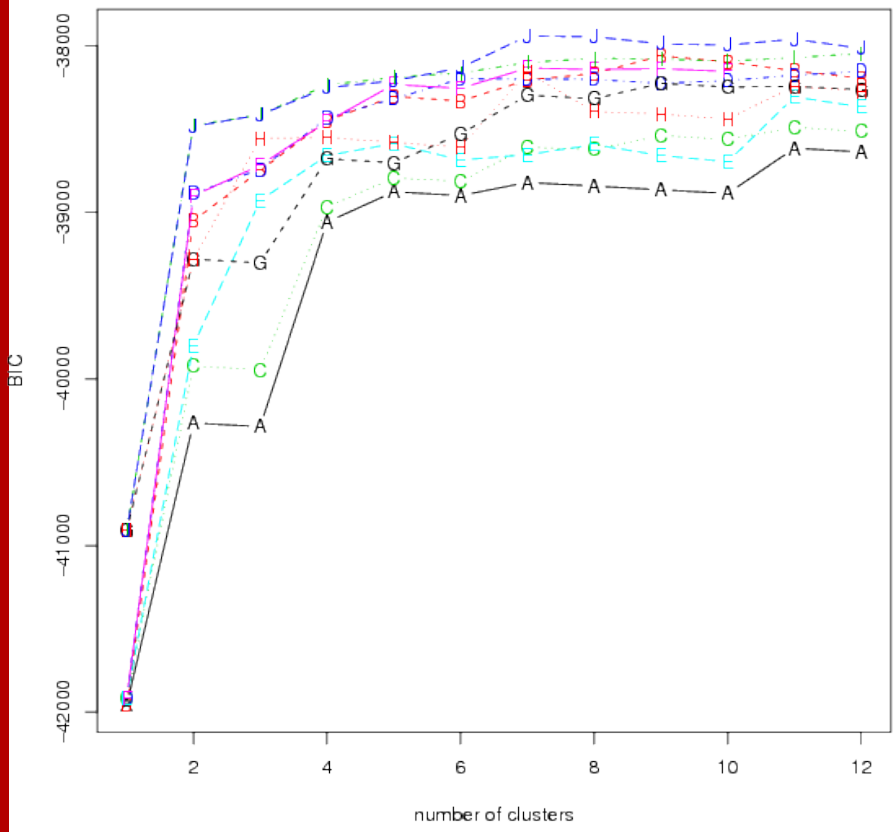
$$\#\theta_K = \underbrace{K-1}_{\pi_{1:K}} + \underbrace{Kd}_{\mu_{1:K}} + \underbrace{Kd(d-1)/2}_{\Sigma_{1:K}}$$

- define

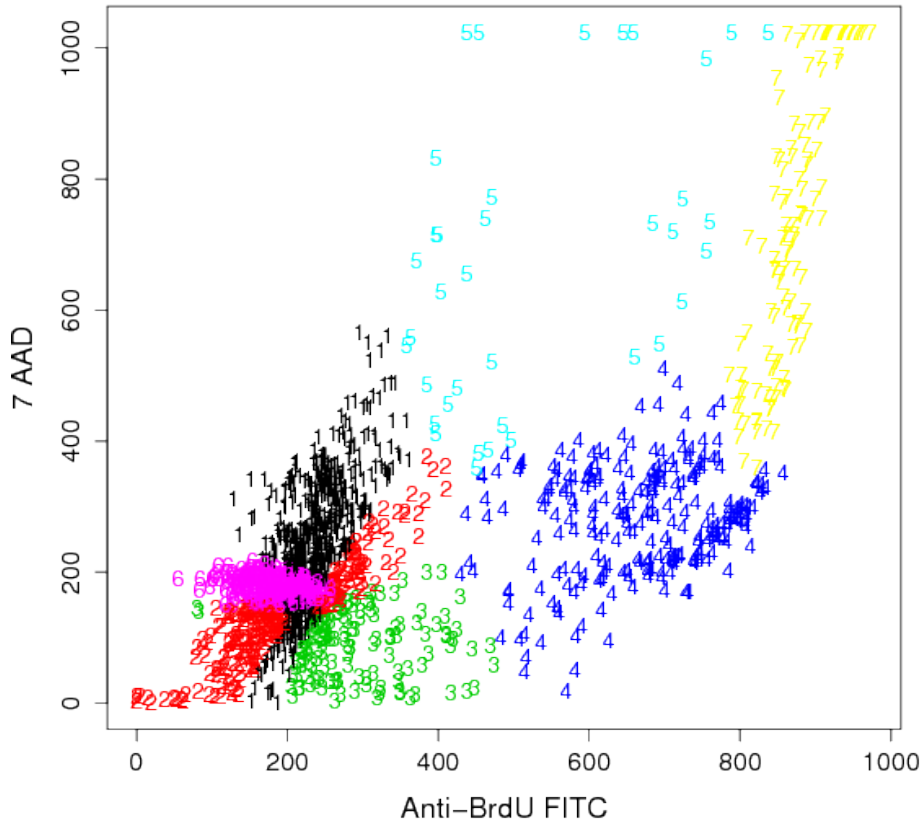
$$BIC(\theta_K) = l(\theta_K) - \frac{\#\theta_K}{2} \ln n$$

- **Select** K that maximizes $BIC(\theta_K)$
- selects true K for $n \to \infty$ and other technical conditions (e.g parameters in compact set)
- but theoretically not justified (and overpenalizing) for finite n

Number of Clusters vs. BIC EII (A), VII (B), EEI (C), VEI (D), EVI (E), VVI (F), EEE (G), EEV (H), VEV (I), VVV (J)

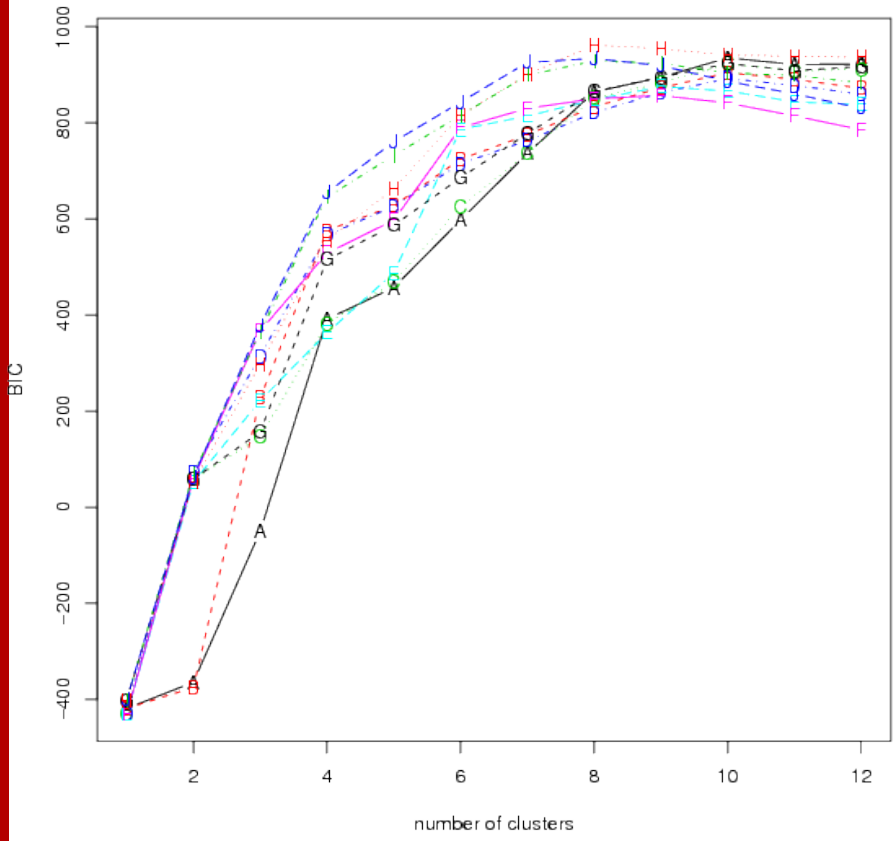


EEV, 8 Cluster Solution

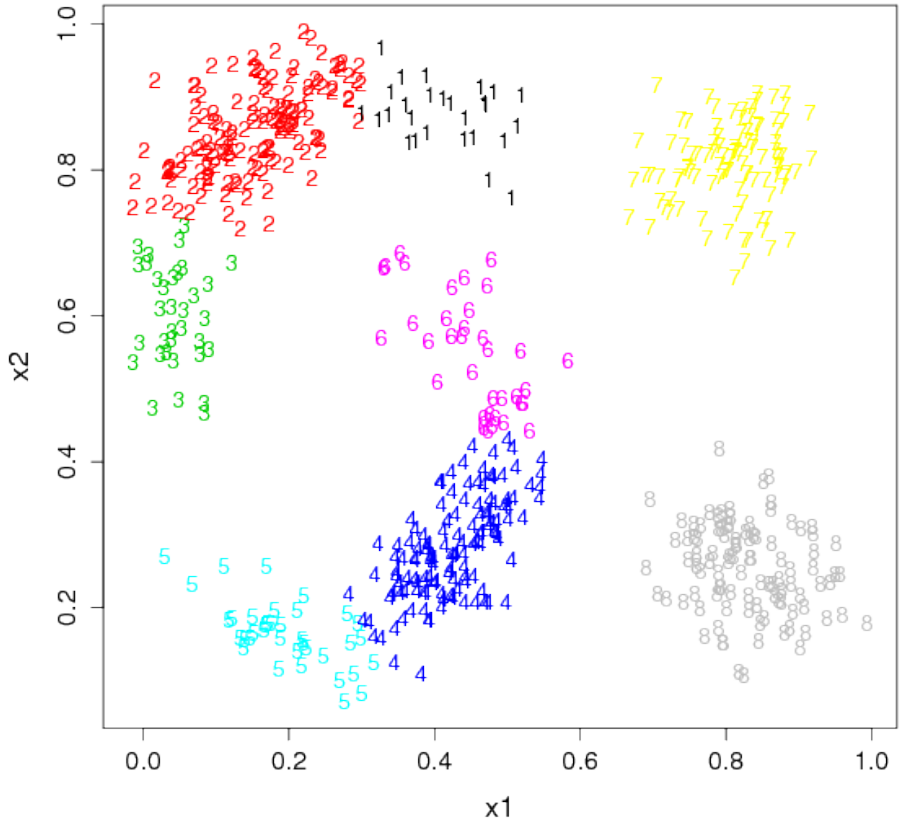


(from)

Number of Clusters vs. BIC EII (A), VII (B), EEI (C), VEI (D), EVI (E), VVI (F), EEE (G), EEV (H), VEV (I), VVV (J)



EEV, 8 Cluster Solution



(from)

[Supplement: Stability methods for choosing K]

- like bootstrap, or crossvalidation
- **Idea** (implemented by)

  for each K
  1. perturb data $\mathcal{D} \rightarrow \mathcal{D}'$
  2. cluster $\mathcal{D}' \rightarrow \Delta'_K$
  3. compare Δ_K, Δ'_K. Are they similar?
     If yes, we say Δ_K is **stable to perturbations**

Fundamental assumption If Δ_K is stable to perturbations then K is the correct number of clusters

- these methods are supported by experiments (not extensive)
- **not YET supported by theory** . . . see for a summary of the area

Clustering with outliers

- What are outliers?
- let p = proportion of outliers (e.g 5%-10%)
- Remedies
  - mixture model: introduce a $K+1$-th cluster with large (fixed) Σ_{K+1}, bound Σ_k away from 0
  - K-means and EM
    - **robust** means and variances
      e.g eliminate smallest and largest $pn_k/2$ samples in mean computation (**trimmed mean**)
    - K-medians
    - replace Gaussian with a heavier-tailed distribution (e.g. Laplace)
  - single-linkage: do not count clusters with $< r$ points

Is K meaningful when outliers present?

- alternative: non-parametric clustering