Lecture 20

Selecting K
- Mixtures
- K-means

PCA

Q3 3/26 4:00

HW7 - correction

HW6 - edits

↑ LVI VAE

LVIII PCA

Lecture VII: Clustering: K-means and Mixtures of Gaussians

Marina Meilă
mmp@uwaterloo.ca

With Thanks to Pascal Poupart & Gautam Kamath
Cheriton School of Computer Science
University of Waterloo

March, 2026

Paradigms for clustering

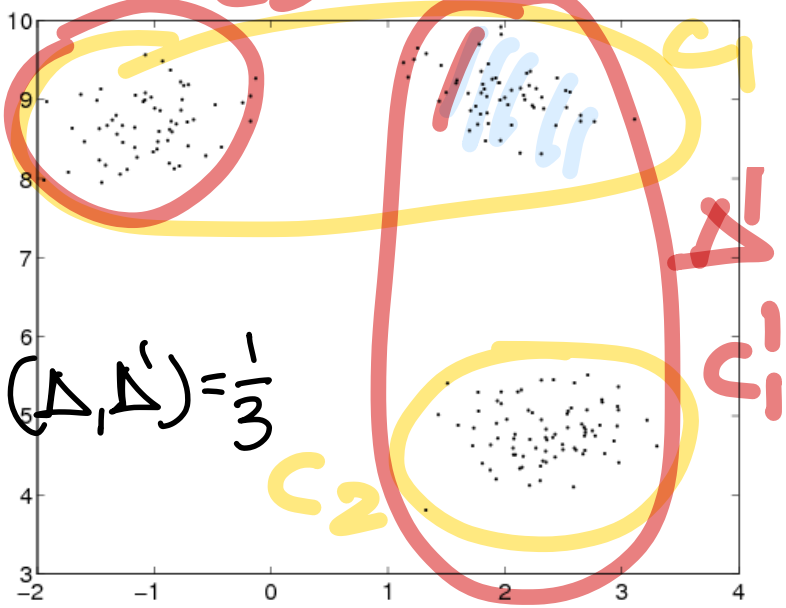
K-means clustering

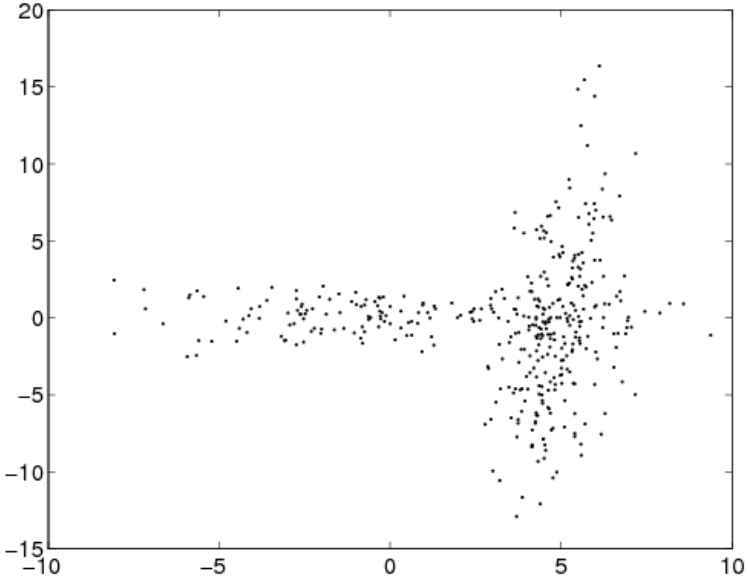
Mixtures of Gaussians and the EM algorithm

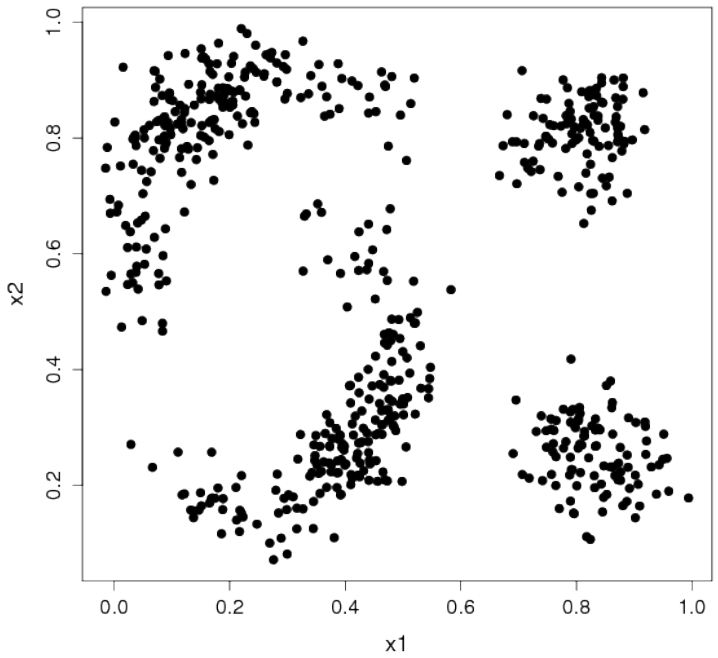
Special topics in clustering

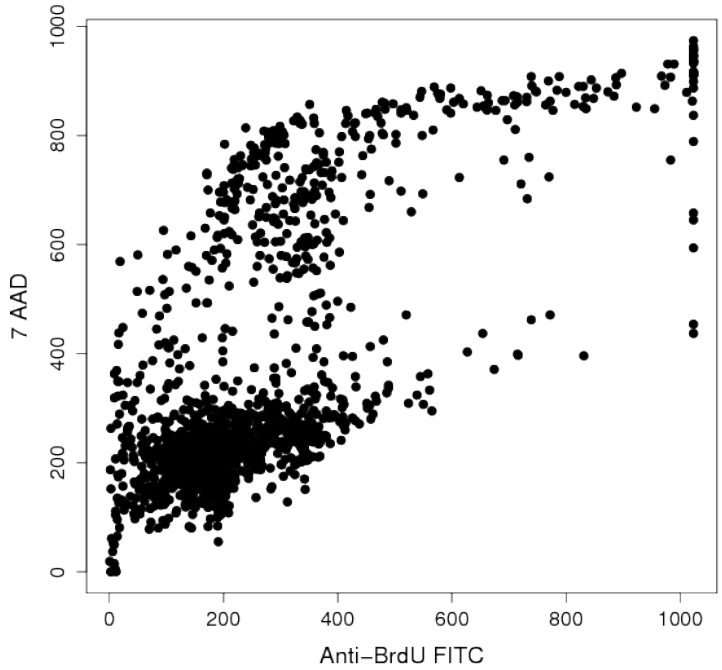
Reading HTF Ch.: 14.3, Murphy Ch.: Ch 11.[1], 11.2.1-3, 11.3, Ch 25, Bach Ch.:

$K=2$


C_2
C_1
Δ
Δ'
C_1'
C_2
$\text{dist}(\Delta,\Delta')=\frac{1}{3}$




x1
x2


Anti−BrdU FITC
7 AAD

(from)

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

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

penalty ↗ with K

- **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

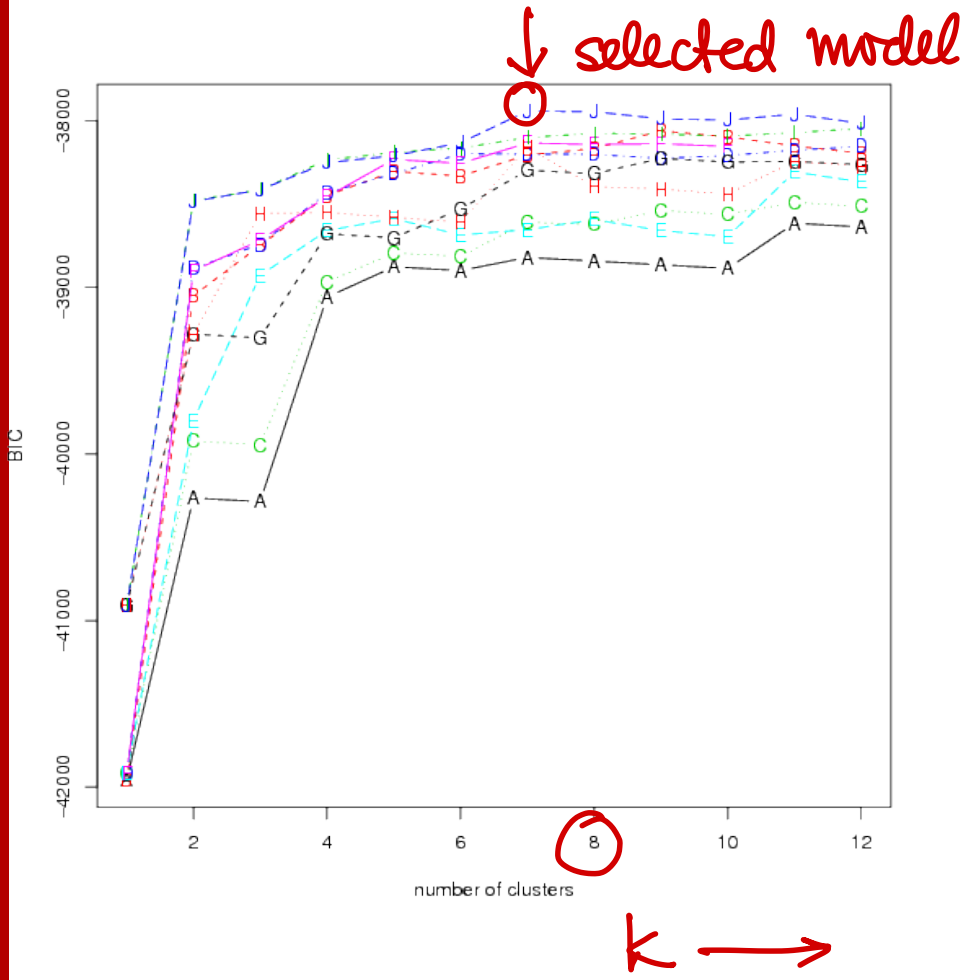
#params = K − 1 + KD + S

Selecting K

1. estimate Δ_K (or θ_K^{ML}) for $K = 2, 3, \dots K_{max}$
2. select K^* by comparing $\Delta_{1:K_{max}}$ $\theta_{1:K_{max}}$

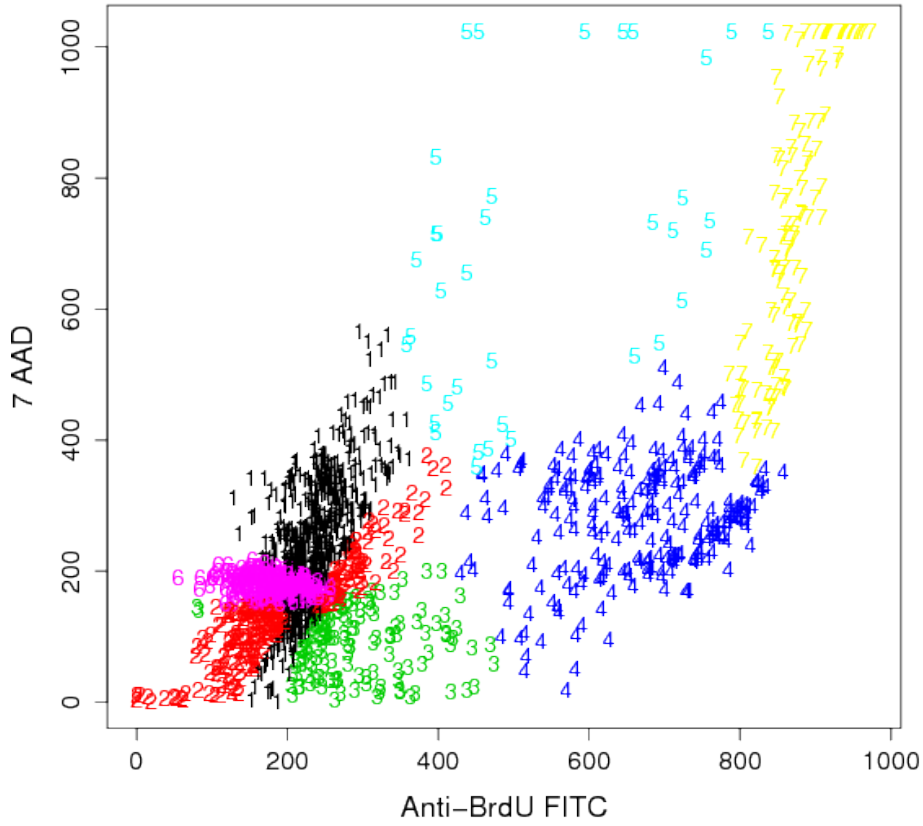
- **BIC** - model based paradigms (heuristic)
- **'elbow'** - any, heuristic
- **stability** - (many paradigms)
  - heuristic — detect instability
  - guarantees - K means, etc optimizations

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)

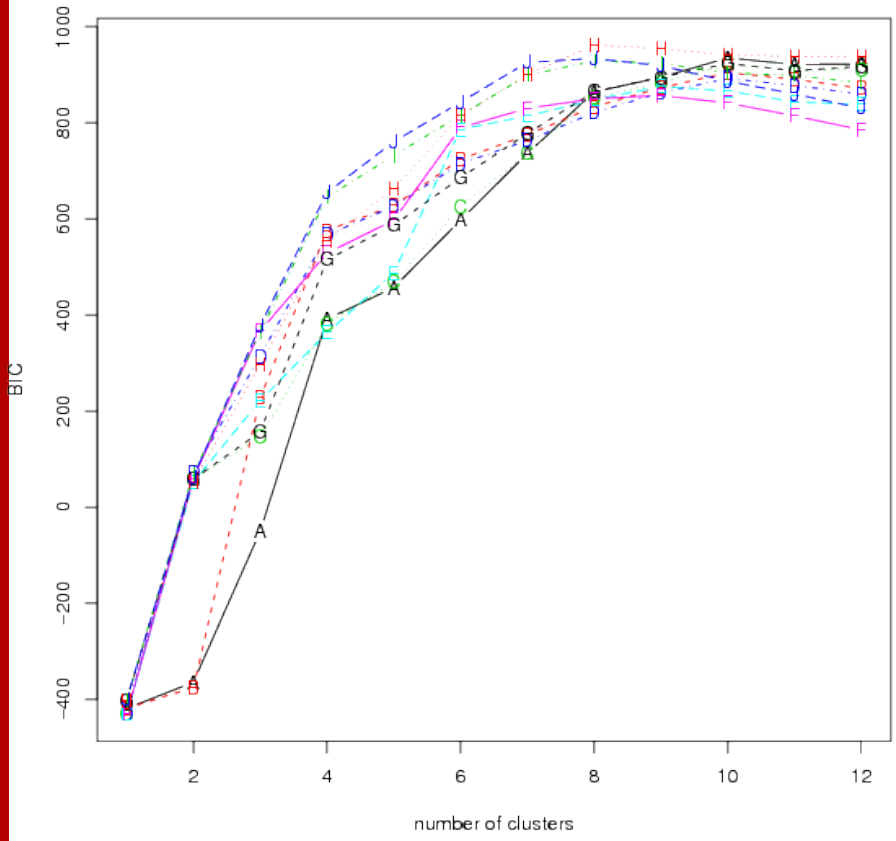


(from)

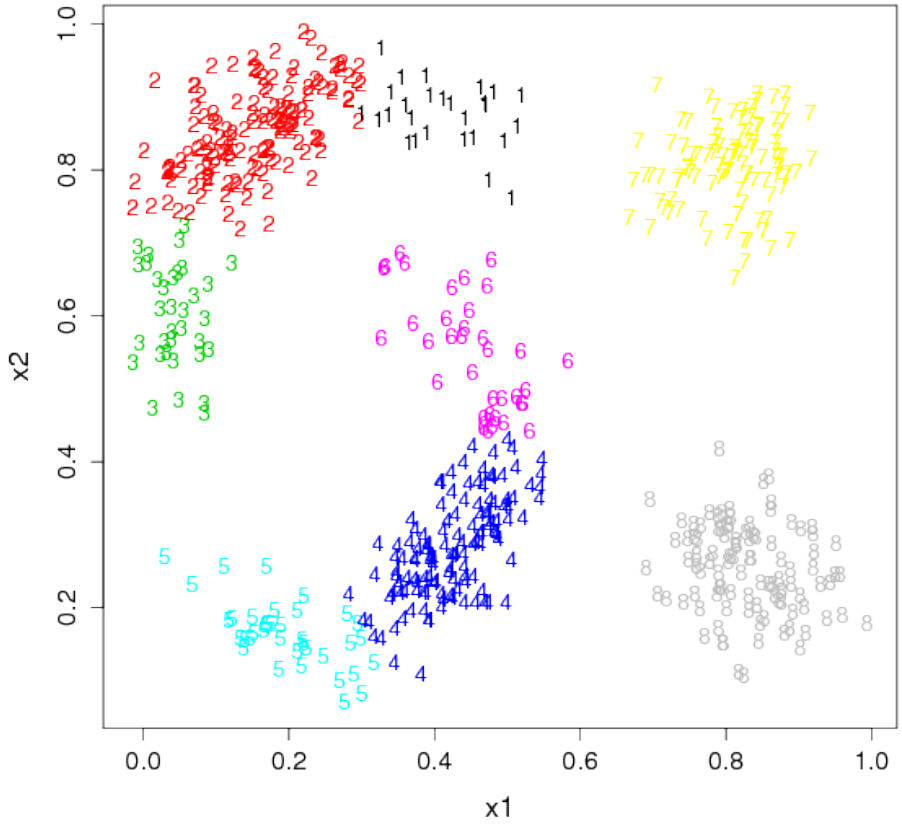
EEV, 8 Cluster Solution



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)

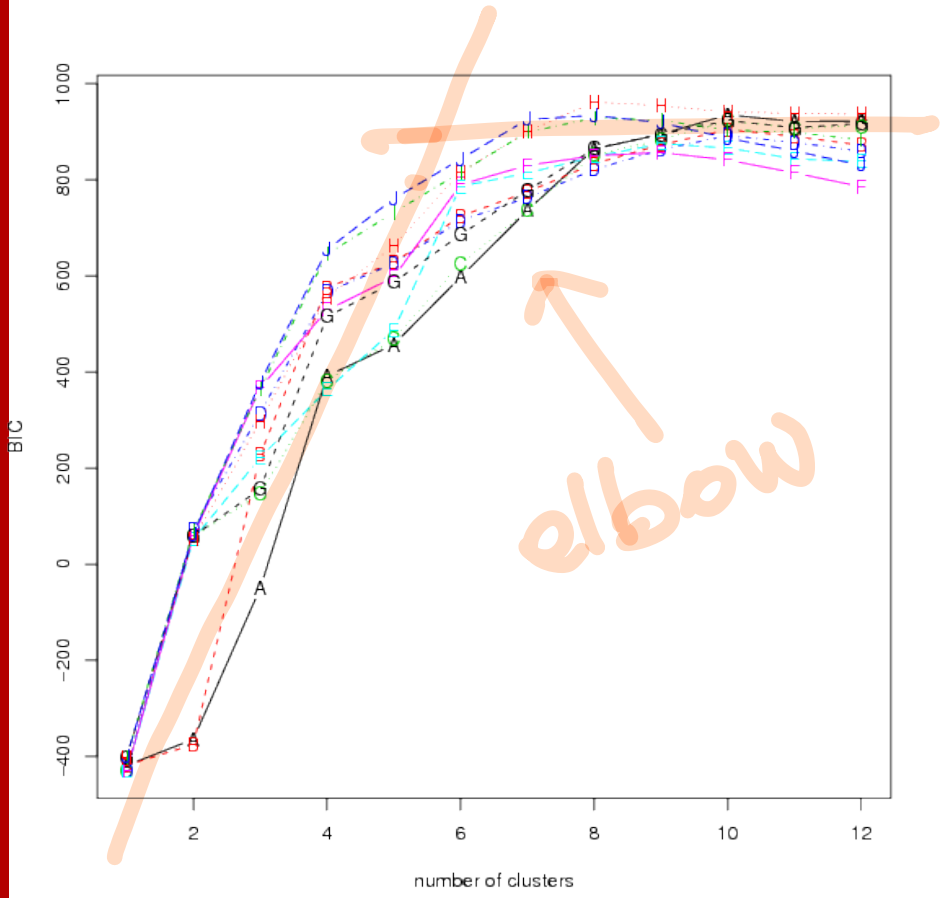


EEV, 8 Cluster Solution

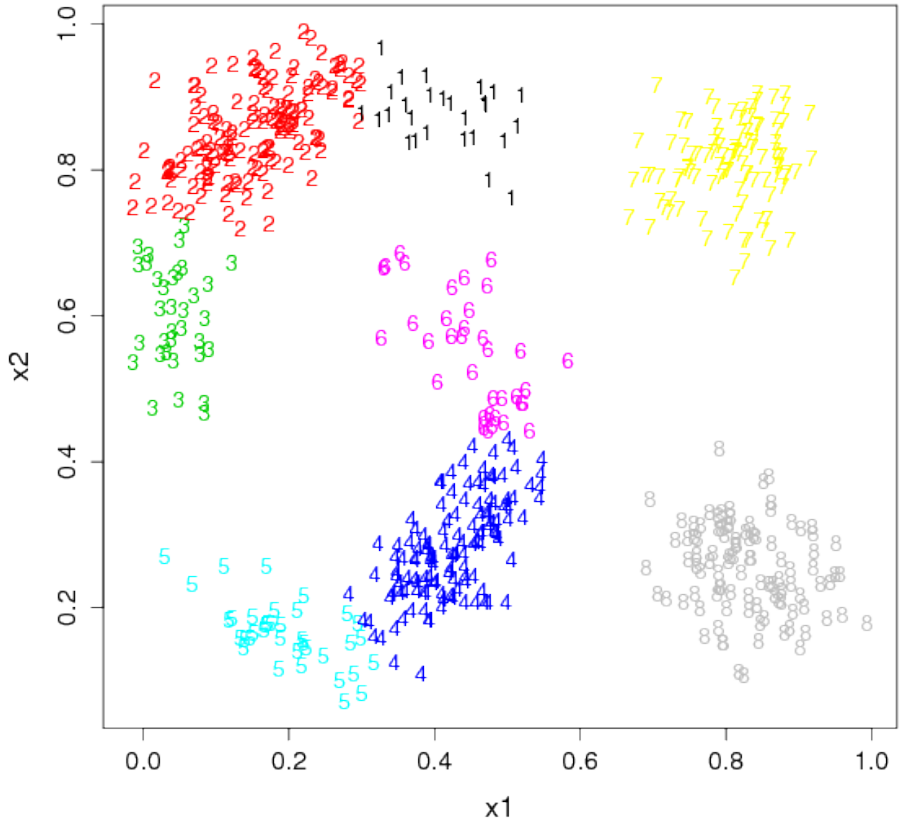


(from)

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



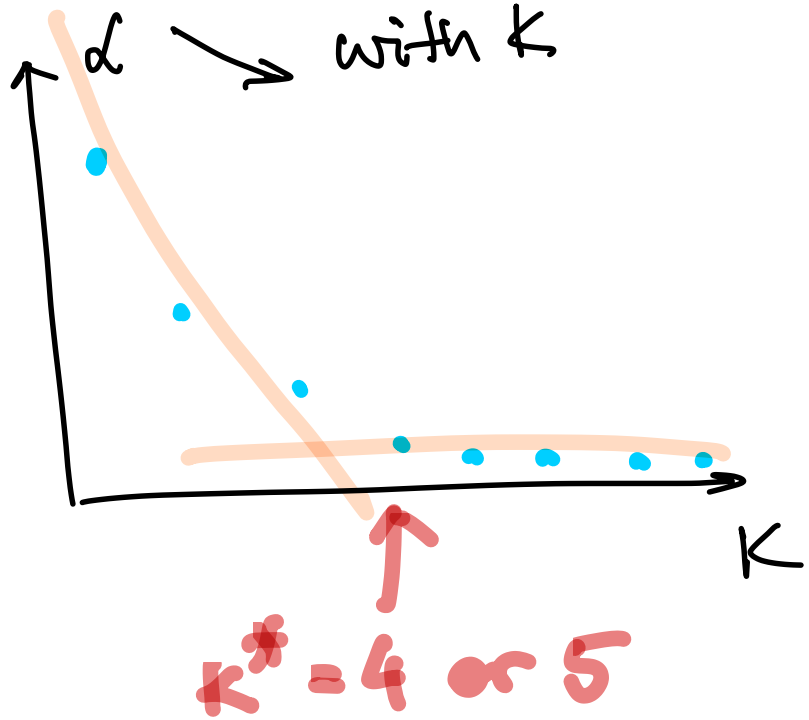
(from)

Selecting K for hard clusterings

ELBOW
any cost based clustering

- based on statistical testing: the **gap** statistic (Tibshirani, Walther, Hastie, 2000)
- **X-means** heuristic: splits/merges clusters based on statistical tests of Gaussianity
- Stability methods
  - Empirical – prove instability
  - Optimization based – prove stability

heuristic



Empirical Stability methods for choosing K

heuristic

- 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**

1. $\mathcal{D}, K \rightarrow \Delta$
2. perturb $\mathcal{D}$ } algo $\rightarrow \Delta'$
3. dist(Δ, Δ') = ? large?

x B times &

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 directly supported by theory** . . . see for a summary of the area

"Misclassification" Error dist(Δ, Δ') = $\frac{\#\text{points to relabel}}{n}$

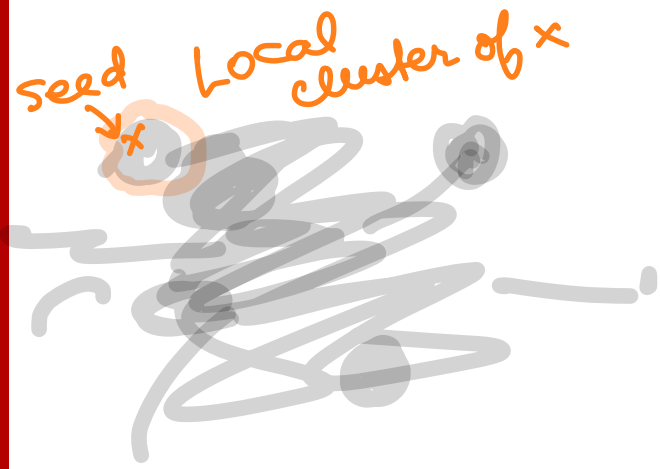
$\in [0, 1)$

Stability with guarantees

A Hierarchical clustering! $K=4$ $K=2$

What I didn't talk about

- Hierarchical clustering
- Subspace clustering (or clustering on subsets of attributes)
- Bi-clustering (and multi-way-clustering)
- Partial clustering
- Non-parametric clustering
- Ensembles of clusterings, consensus clustering, and clustering clusterings



Hierarchical clustering

- **Divisive** (top down)
  - starts with all data in one cluster, divides recursively into 2 (or more) clusters
  - Example: spectral clustering, min diameter
- **Agglomerative** (bottom up)
  - starts n cluster containing 1 item, merges 2 clusters recursively
  - Example: Ward algorithm, single linkage
- **Hierarchical Dirichlet processes**
- Remarks
  - Any cost based clustering paradigm can produce a hierarchical clustering
  - Any non-parametric level-sets paradigm can produce a hierarchical clustering
  - Mixture models (finite or not) can also be defined hierarchically. Issues of identifyability appear

The Ward agglomerative algoritm

- Cost = same as K-means
- Algorithm idea:
  - Start with n single point clusters
  - Merge the two clusters that increase $\mathcal{L}$ the least, until K clusters left
- **Greedy**, recursive algorithm, $\mathcal{O}(n^3)$ operations

Subspace clustering

- Problem: each cluster is defined by a subset of relevant attributes (features)
  - Examples: user modeling (clusters of users vs clusters of products/services), gene expression data
- Known as **Clustering on Subsets of Attributes (COSA) Biclustering (and Multiway Clustering), Subspace clustering**
- Amounts to clustering both the data exemplars and the data features
- Approaches
  - **COSA** cost based, $+$ additional entropy term. Alternate minimization algorithm.
  - Dirichlet process mixtures approach. Each $f(.;\theta_k)$ samples a set of relevant features. Estimated by MCMC
  - **Multivariate Information Bottleneck** Information theory based. Estimation by alternate (KL-divergence) projections.
  - many others. . . see IEEE TKDE

Partial clustering

- Problem: Given a node, find its cluster
- Premise: the data set is extremely large, there are many small clusters, possibly $\mathcal{O}(n)$
- **Nibble** algorithm of

  Given: a graph, by its Markov transition matrix P
  Start with node i, tolerance ε, number steps t
  Initialize $p \in \mathbb{R}^n$ with $p_i = 1$, $p_j = 0$ for $j \neq i$
  - Iterate for t steps
    1. $p \leftarrow Pp$
    2. for $j = 1 : n$, if $p_j < \varepsilon$ set $p_j = 0$

  Output $C(i) = \{ j \mid p_j > 0 \}$
- $C(i)$ is the set of items attainable from i by a "likely" path
- Original algorithm has sparsest cut guarantees
  Used as subroutine by other algorithms.

Methods based on non-parametric density estimation

Idea The clusters are the isolated peaks in the (empirical) data density

- group points by the peak they are under
- some outliers possible
- $K = 1$ possible(no clusters)
- shape and number of clusters K determined by algorithm
- **structural parameters**
  - **smoothness** of the **density estimate**
  - what is a peak

Algorithms

- peak finding algorithms Mean-shift algorithms
- level sets based algorithms
  - Nugent-Stuetzle, Support Vector clustering
- Information Bottleneck

Lecture Notes VII – Principal Component Analysis

Marina Meilă
mmp@uwaterloo.ca

With Thanks to Pascal Poupart & Gautam Kamath
Cheriton School of Computer Science
University of Waterloo

March 19, 2026

Principal Component Analysis (PCA)

$X = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{bmatrix}$ } n, width D

$x^{1:n} \in \mathbb{R}^D$

mean $x^{1:n} = 0$

$n > D$

1. $\mathrm{Cov}(X) = \Sigma = \frac{1}{n} X^T X \in \mathbb{R}^{D \times D}$

E-value decomposition

$$\Sigma = V \tilde{\Sigma} V^T \cdot \frac{1}{n}$$

V orthogonal

$\tilde{\Sigma} = \begin{bmatrix} \tilde{\sigma}_1^2 & & \\ & \ddots & \\ & & \tilde{\sigma}_D^2 \end{bmatrix}$

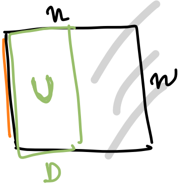
$V = [v_1 \cdots v_D]$ basis for $\mathbb{R}^D$

2. Gram matrix

$$G = XX^T = \tilde{\tilde{U}}\, \tilde{\tilde{\Sigma}}\, \tilde{\tilde{U}}^T = U \tilde{\Sigma} U^T$$

$\tilde{\tilde{\Sigma}} = \begin{bmatrix} \tilde{\sigma}_1^2 & & & \\ & \ddots & & \\ & & \tilde{\sigma}_D^2 & \\ & & & 0 \ \ddots \ 0 \end{bmatrix} \in \mathbb{R}^{n \times n}$

$\tilde{\tilde{U}} =$



3. Representing X, x^i

$X = U \sqrt{\Sigma} V^T$ $\quad n \times D$ $\quad \leftarrow SVD(X)$

$u^i_j = u_{ij} \tilde{\sigma}_j$

$\Downarrow$

$x^i = v_1 \cdot u^i_1 + v_2 \cdot u^i_2 + \cdots + v_D u^i_D$ $\quad \rightarrow$ scalar coefficients

in basis V

$U = [u_{ij}]_{\substack{i=1:n \\ j=1:D}}$

$x^i \leftrightarrow (u^i_1 \cdots u^i_D)$

$\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \tilde{\sigma}_d \geq \cdots \tilde{\sigma}_D$

$v_1 \quad v_2 \quad v_d \quad v_d$

Principal e-values vectors

Example – Gaussian data 2D