Lecture 20

Selecting k

PCA

Q3 :
HW6 → notation
LVI VAE ↗
HW7 correction
LVII

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 ✓

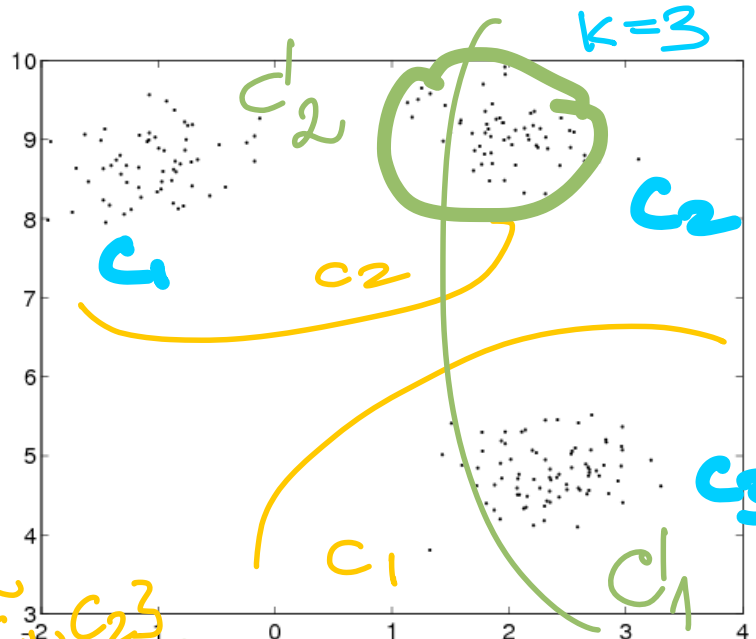
K-means clustering ✓

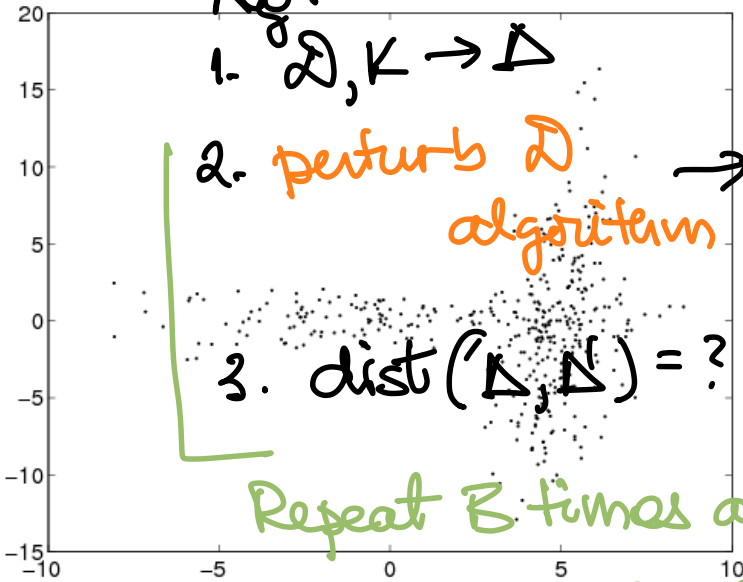
Mixtures of Gaussians and the EM algorithm ✓

Special topics in clustering
- choice of K
- ...

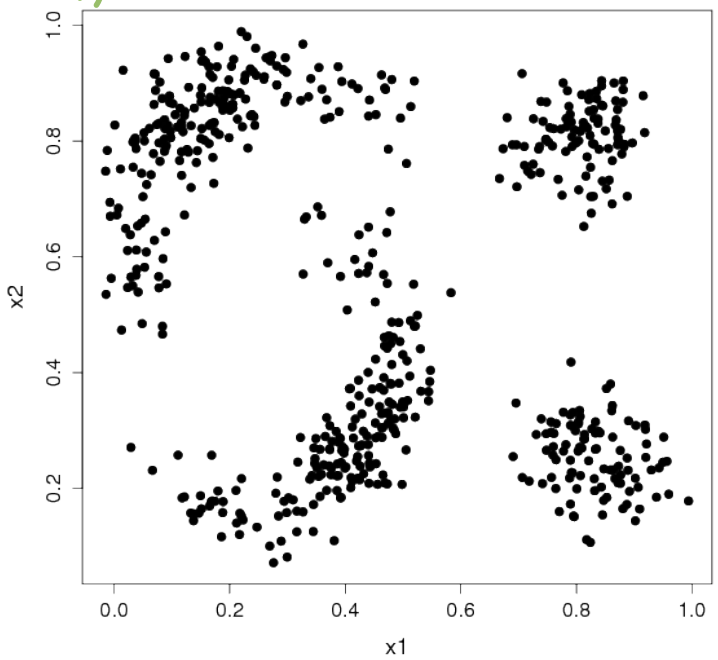
PCA ←

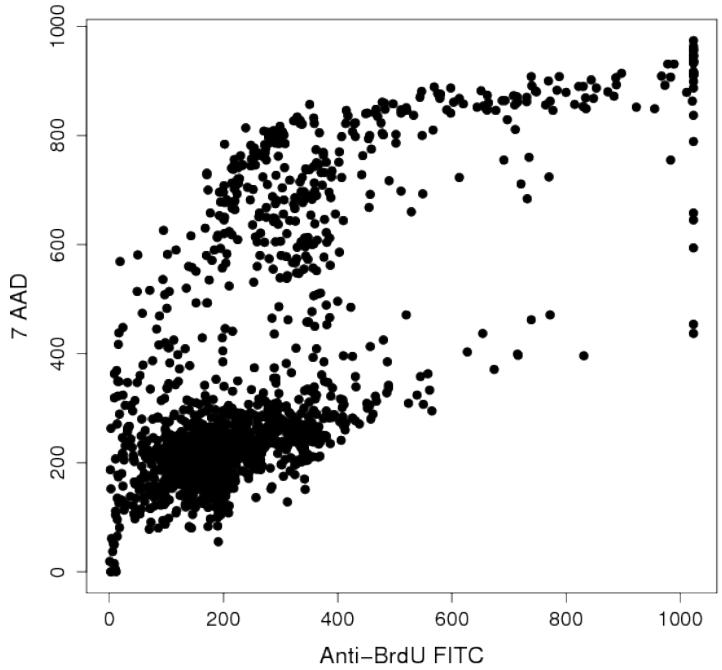
Reading HTF Ch.: 14.3, Murphy Ch.: Ch 11.[1], 11.2.1-3, 11.3, Ch 25, Bach Ch.:





K=2
D={C1, C2}
Δ'={C'1, C'2}





(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

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

- **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

Selecting K for mixture models

1. Run clustering algo for $K=2,3,\ldots K_{max}$
$\Rightarrow \Delta_2, \Delta_3, \ldots \Delta_{K_{max}}$
2. Select K^* based on Δ_K $K=2:K_{max}$

The BIC (Bayesian Information) Criterion

HEURISTIC

- 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

want: max score

$$BIC(\theta_K) = \underbrace{l(\theta_K)}_{\uparrow \text{ with } K} - \frac{\#\theta_K}{2} \ln n$$

- **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

$\#\theta$ = nr params $\uparrow$ with K

K selection

- BIC ← model-based ~ likelihood ✓
- "elbow" heuristics — cost-based ✓
- stability —"— — all clustering paradigms ✓

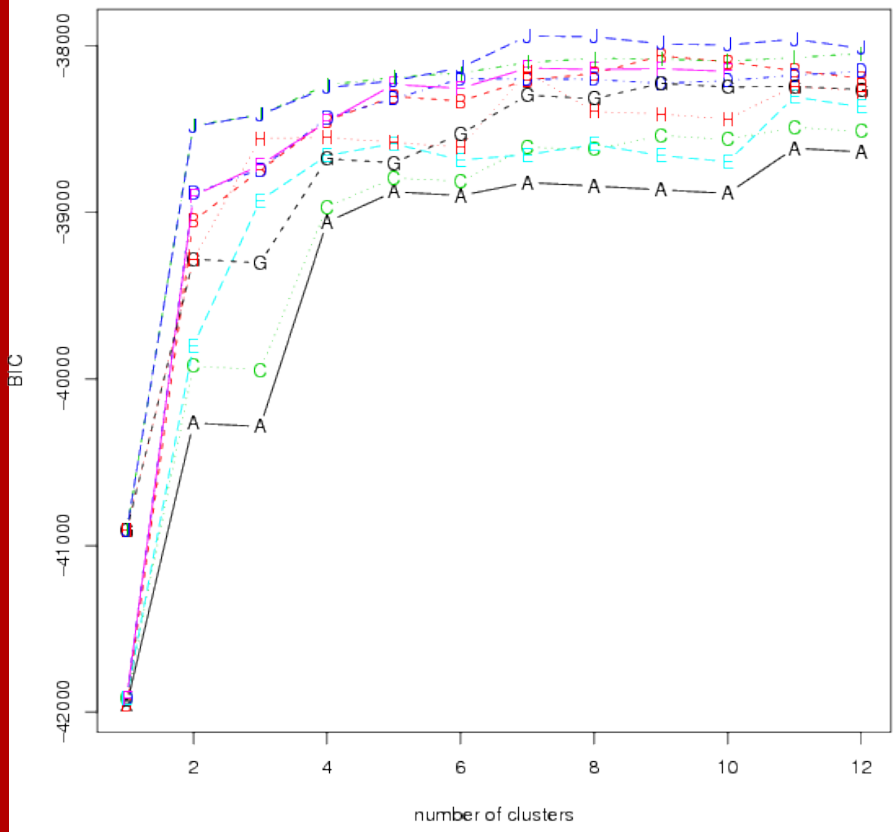
stability → some clustering (K-mean (EM) mixtures

citations in LVIII

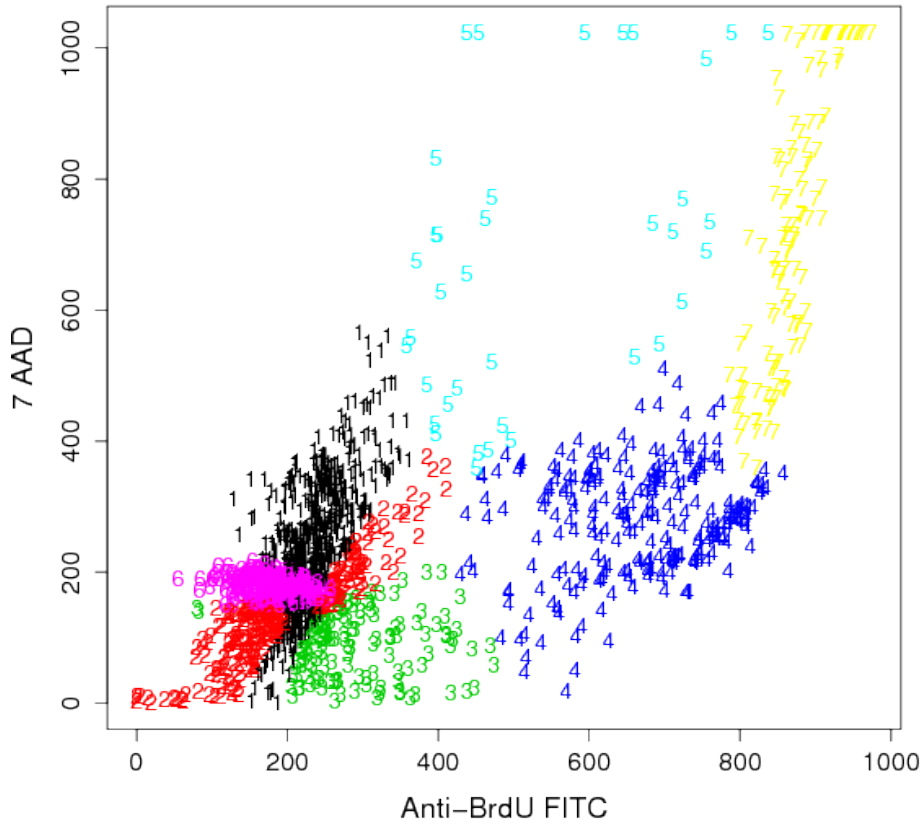
Rigorous, based on optimization

If well separated clusters ∃

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)

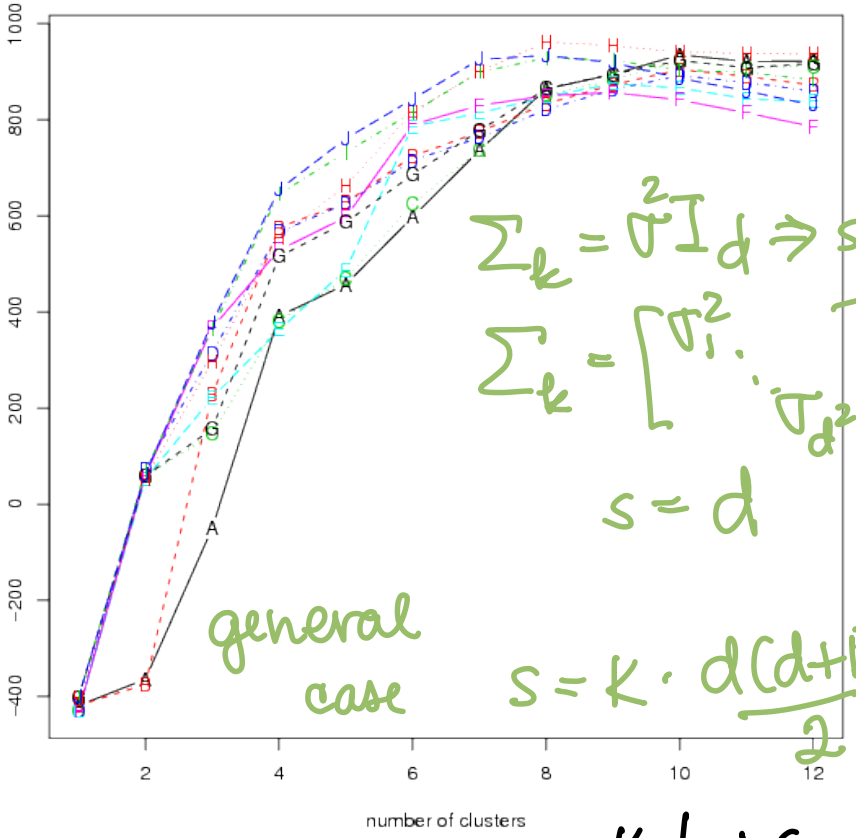


EEV, 8 Cluster Solution



(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)



$\Sigma_k = \sigma^2 I_d \Rightarrow s = 1$

$\Sigma_k = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2 \end{bmatrix}$ $\forall k$

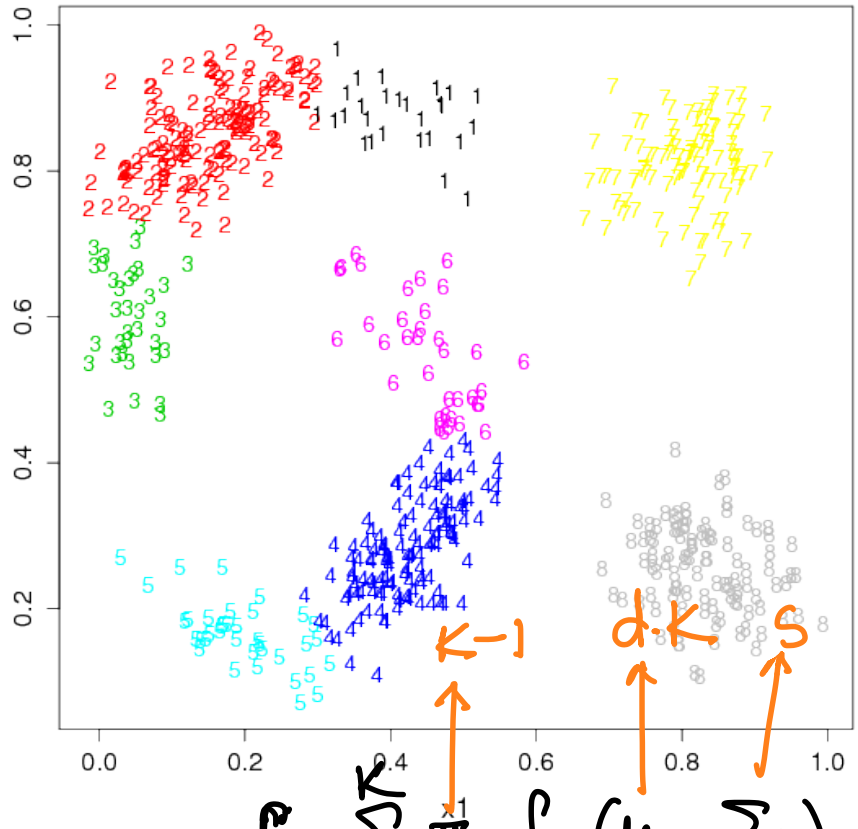
$s = d$

general case $s = K \cdot \frac{d(d+1)}{2}$

$\#\text{params} = K - 1 + Kd + s$

(from)

EEV, 8 Cluster Solution



$K-1$ $\quad d \cdot K \quad s$

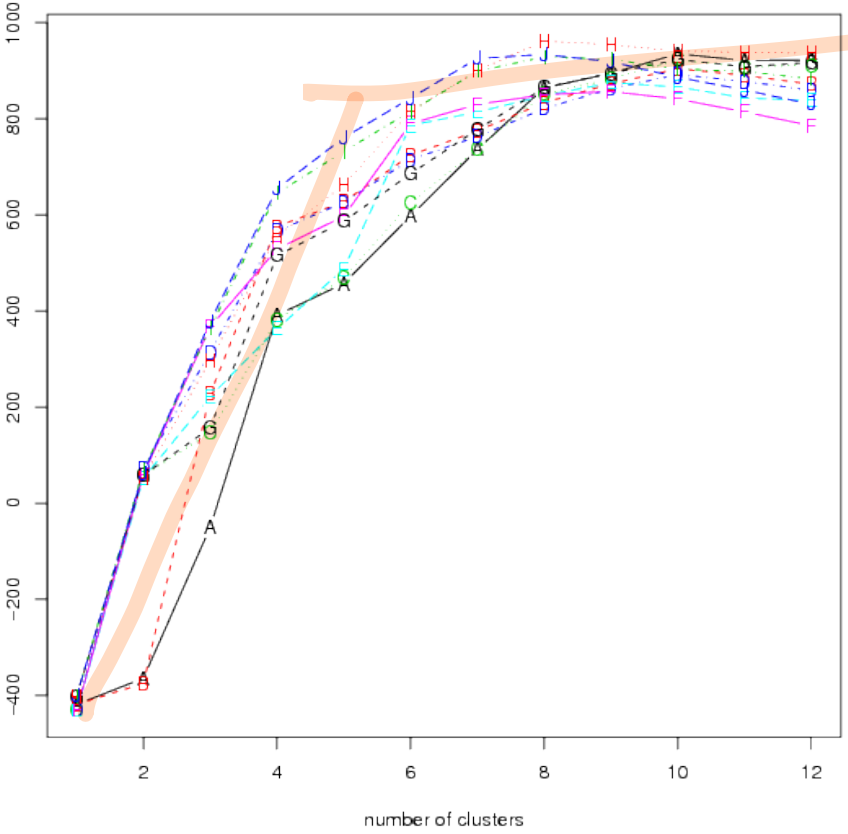
$f = \sum_{k=1}^{K} \pi_k f_k(\mu_k, \Sigma_k)$

$x, \mu_{1:K} \in \mathbb{R}^D$

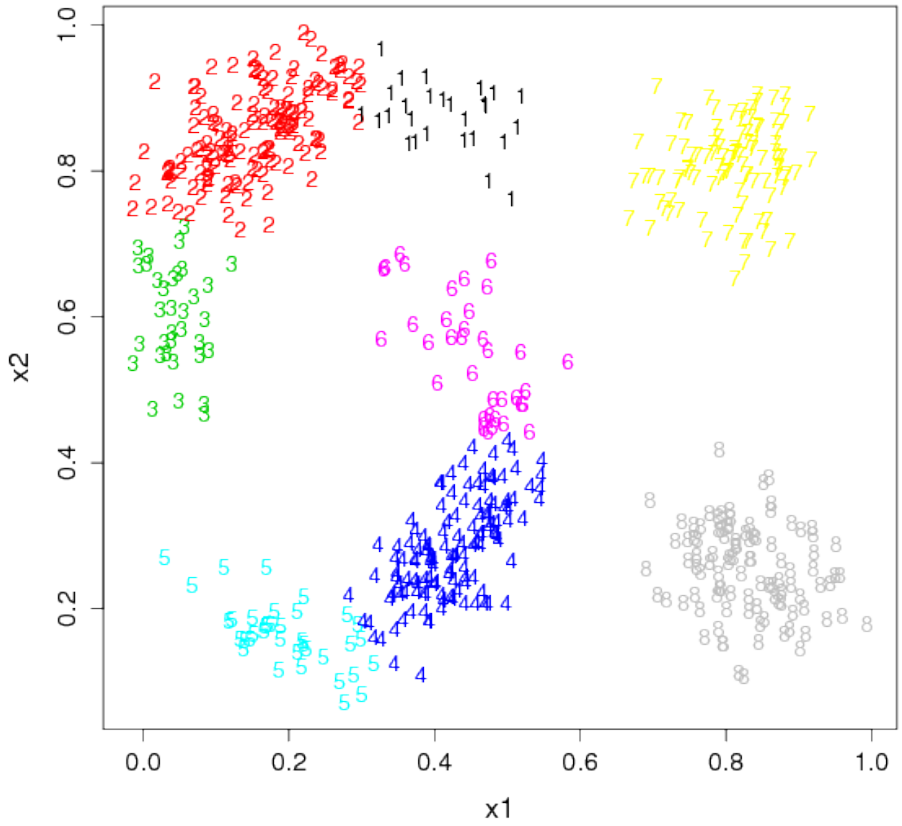
$s = \#\text{params}(\Sigma_k)$ same for $k = 1{:}K$

"elbows" = change in slope for loss

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)



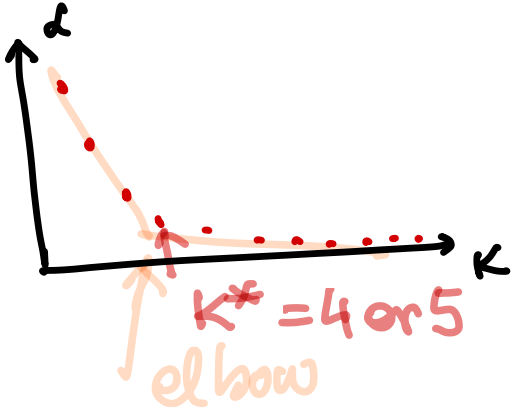
EEV, 8 Cluster Solution



(from)

elbow for k-means

$\mathcal{L}(\Delta_K)$ = loss of best clustering found for K



Stability heuristics

Idea: good clusters ⇒ stable to perturbations

good Δ

Selecting K for hard clusterings

- 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

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**

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

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

Eigendecompositions of Variance

0. D as matrix (centered)

$$X = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{bmatrix} \in \mathbb{R}^{n \times D}$$

1. $\text{Cov}\, X \equiv \Sigma = \frac{1}{n} X^T X$

D

$\text{mean}(X) = 0 \Leftrightarrow 1^T X = 0_d$

Eigendecomposition

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

orthogonal

$$\begin{bmatrix} \tilde{\sigma}_1^2 & & \\ & \ddots & \\ & & \tilde{\sigma}_D^2 \end{bmatrix}$$ e-value

$\tilde{\sigma}_j$ = singular values

$$V = [v_1 \; v_2 \cdots v_D] \in D \times D$$

orthogonal basis

Principal

$(\tilde{\sigma}_1, v_1), (\tilde{\sigma}_2, v_2), \ldots$

Principal e-values e-vector

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

Eigendecompositions of Variance

2. Gram Matrix

$G = XX^T \in n \times n$

$n > D$

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

$\tilde{U}$: orthogonal

$\tilde{\Sigma} = \begin{bmatrix} \tilde{\sigma}_1^2 & & & & \\ & \ddots & & & \\ & & \tilde{\sigma}_D^2 & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix}$

$\tilde{U} = [\, U \,]$ (D columns, $n-D$ columns)

$= U \tilde{\Sigma} U^T$

$n \times D$ $\quad D \times D$ $\quad D \times n$

SVD(X)

3.

$X = U \sqrt{\tilde{\Sigma}} V^T$

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

basis vectors

row i of $X = x^i = v_1 \cdot u_1^i + v_2 \cdot u_2^i + \cdots v_D \cdot u_D^i$

coefficients

$\rightarrow U = [u_{ij}]$. $\quad u_{ij} \tilde{\sigma}_j$

Eigendecompositions of Variance

4. Reduce dim of X

$D \longrightarrow d < D$

$\tilde{x} \in \mathbb{R}^d$

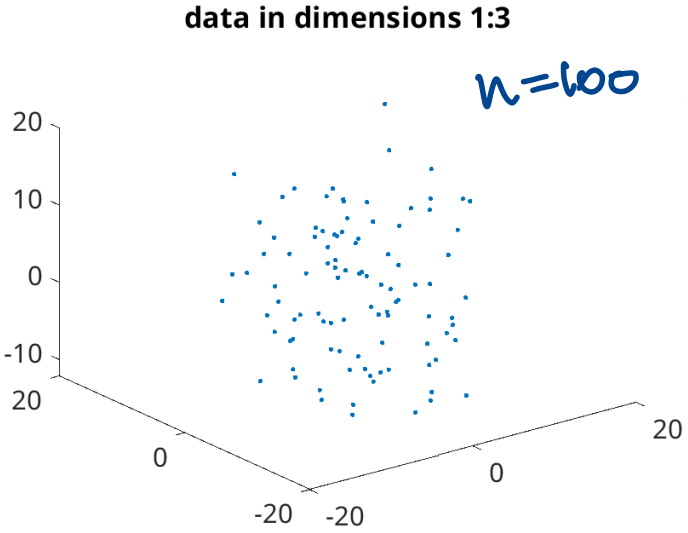
$\tilde{x}^i = u_1^i v_1 + \cdots u_d^i v_d$ optimal approximation

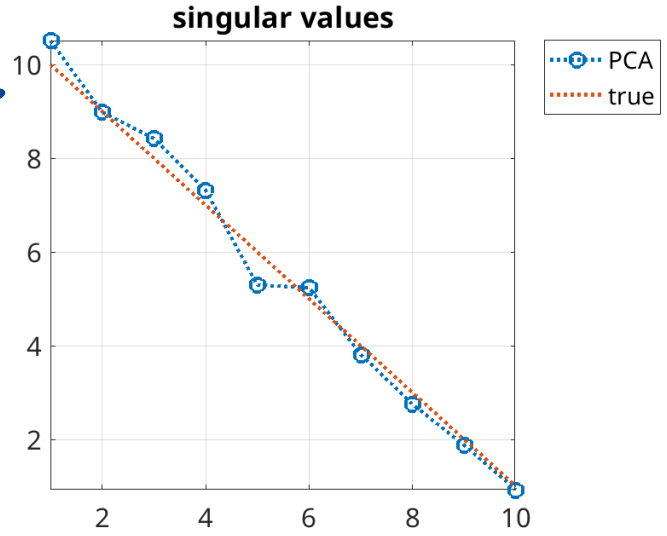
$\frac{1}{n}\sum_{i=1}^{n} \|x^i - \tilde{x}^i\|^2 = \tilde{\sigma}_{d+1}^2 + \cdots \tilde{\sigma}_D^2$ = min MSE over all possible bases B

residual (error)

mean squared error

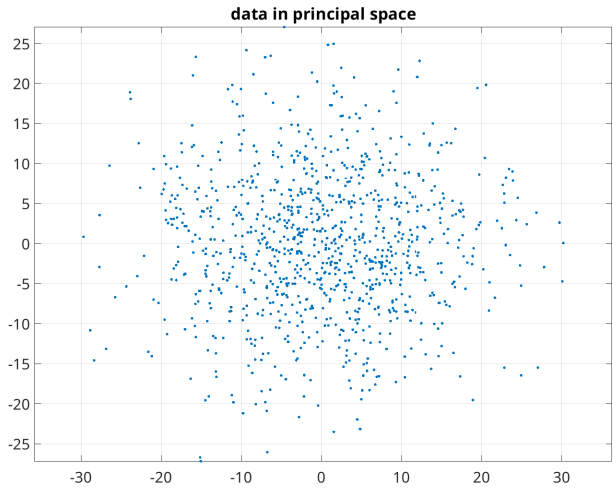
Example – Gaussian data

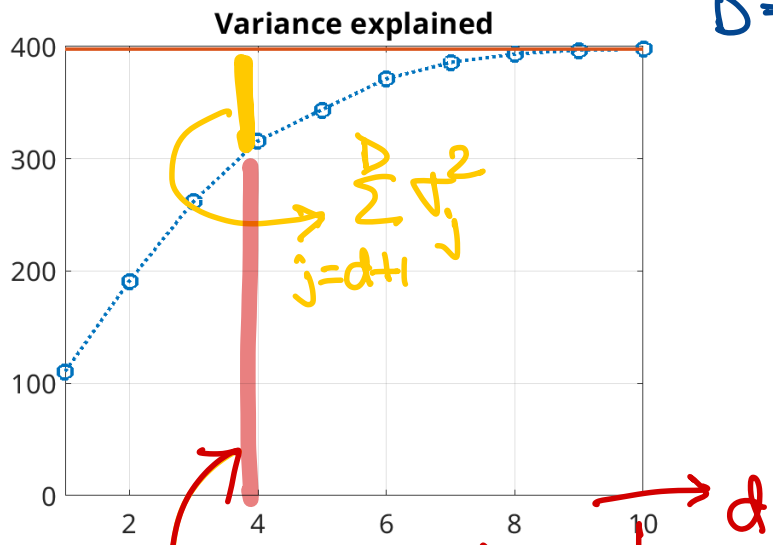




$\nabla_j = 11 - j$

$D = 10$





choosing d

var explained

D

Example – Gaussian data 2D