

L11: Choosing k

the # clusters

Data Mining
Sep 29, 2026
Jeff M. Phillips



4 Methods in choosing k

1. Draw Pictures (Plot)

2-dimensional

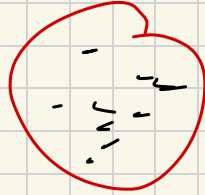
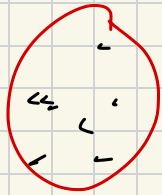
2. Elbow Method
Cost function

3. Silhouette Score
assume mean-based cost.

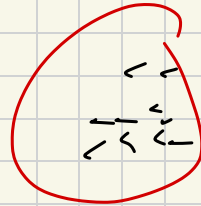
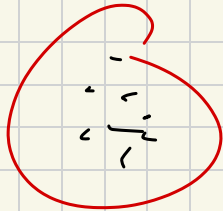
4. Bayesian Information Criterion (BIC)
- full likelihood function.

1. Choose k k_{sc} Plotting

Assum. data $X \in \mathbb{R}^z$



$k=4$



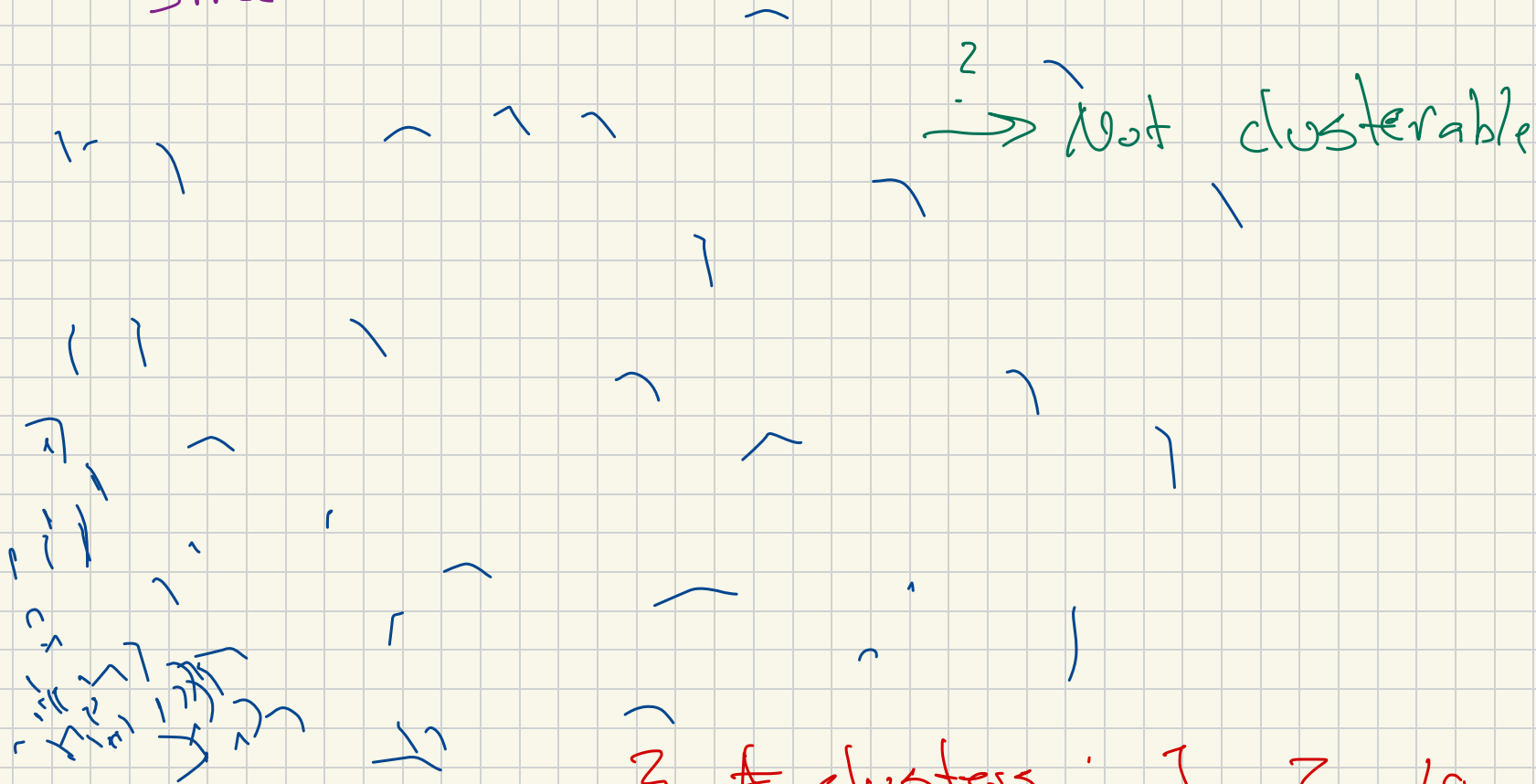
1 outlier

$d > z$

→ Dimensionality Reduction
PCA, MDS, DML
Spectral.

①

Smear



? clusters: 1, 2, 10

2 Elbow Method

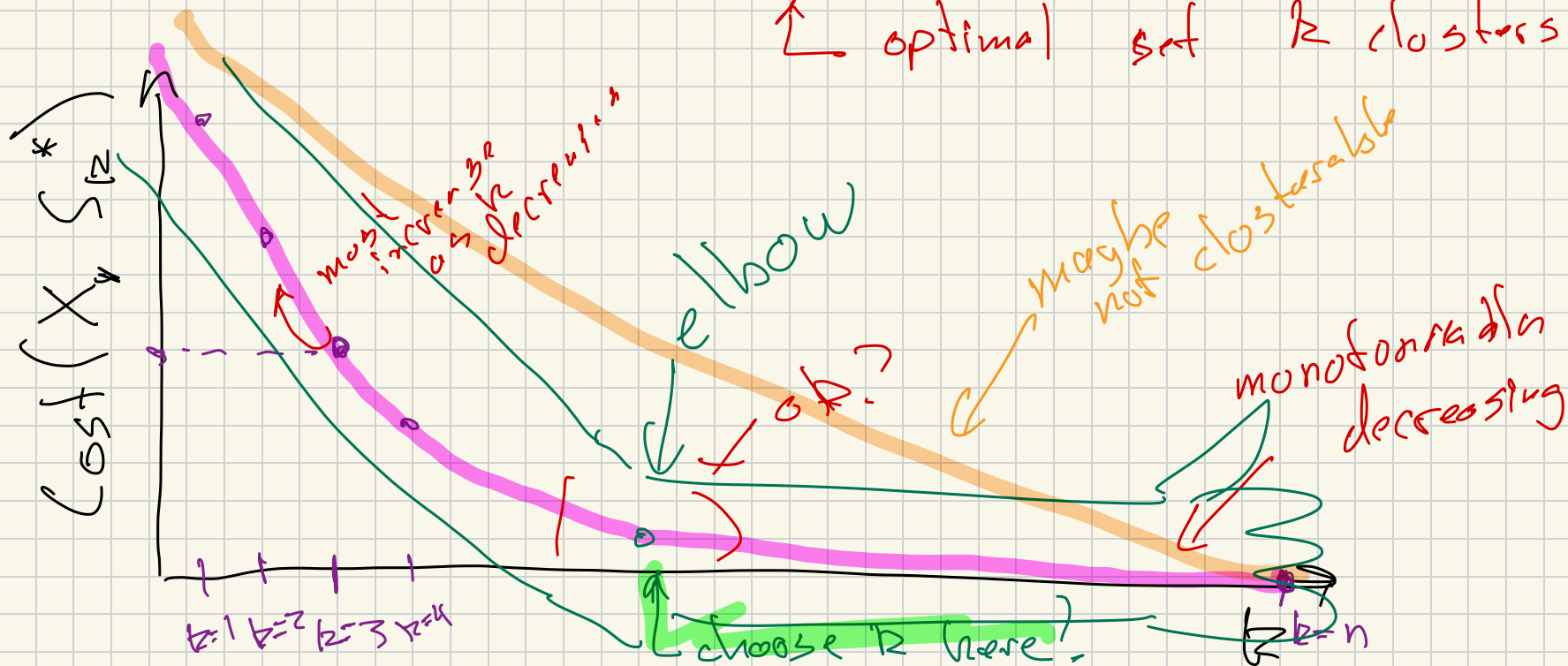
$X,$

$$\text{cost}(X, S_k^*)$$

ex:

$$\text{cost}(X, S_k^*) = \text{cost}_2(X, S_k^*) = \frac{1}{|X|} \sum_{x_i} \|x_i - \phi_{S_k^*}(x_i)\|^2$$

↑ optimal set k clusters



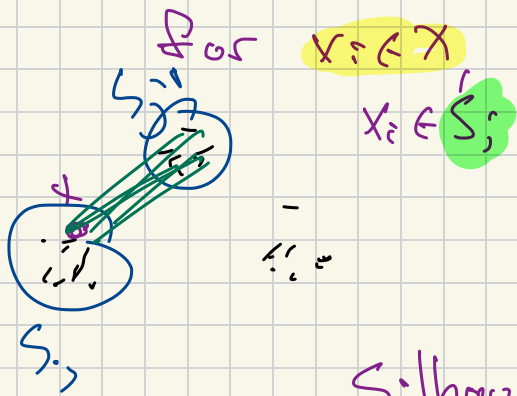
3. Silhouette Score

not single-link
HAC

— assume mean-based cluster

Goal: Score for each k , take max.

$$X \rightarrow S_1, S_2, \dots, S_k$$



$$a(i) = \frac{1}{|S_j| - 1} \sum_{\substack{x \in S_j \\ x \neq x_j}} D(x, x_i)$$

cluster score

$$b(i) = \min_{j' \neq j} \frac{1}{|S_{j'}|} \sum_{x \in S_{j'}} D(x, x_i)$$

replacement score

$$\text{Silhouette } s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

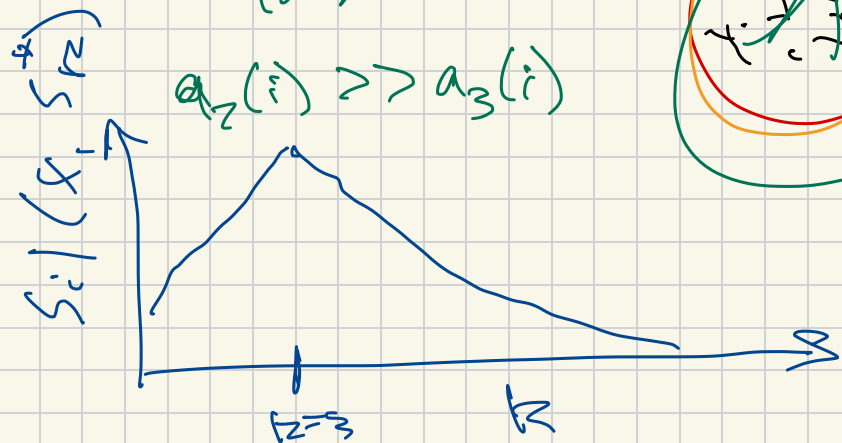
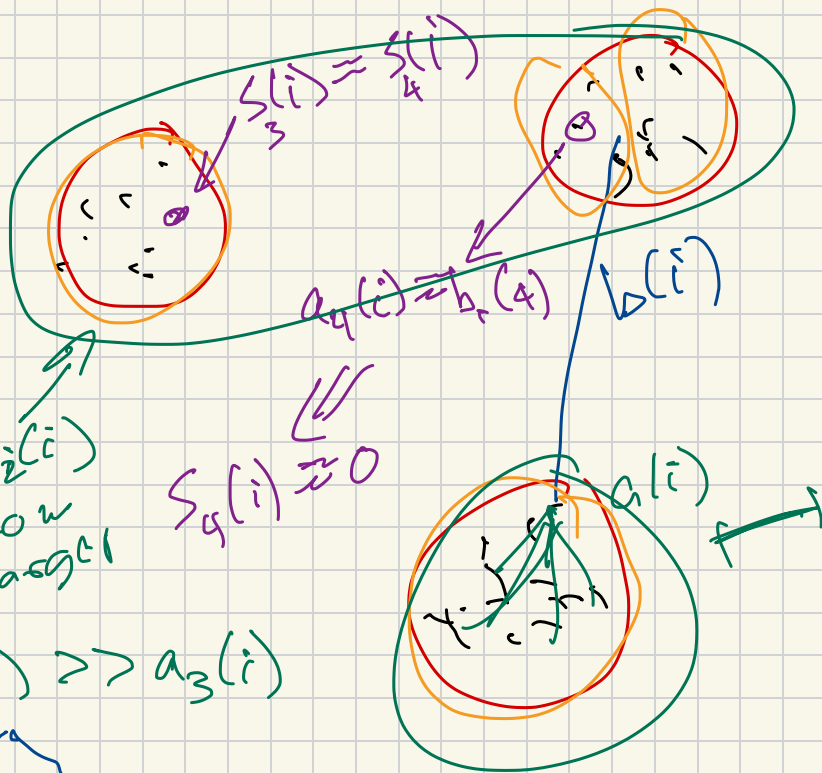
$\in [-1, 1]$
 ≥ 0

if in best cluster

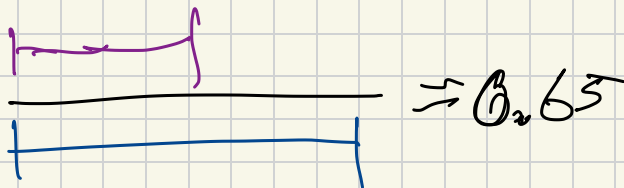
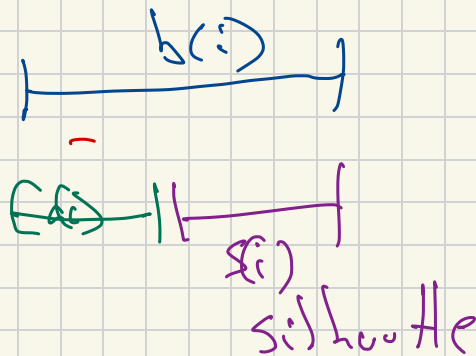
$$\text{Silhouette}(X, S) = \frac{1}{|X|} \sum_{x_i} s(i)$$

$k=4$

$k=2$

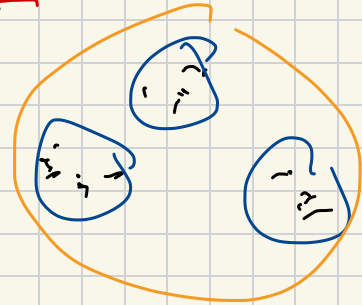
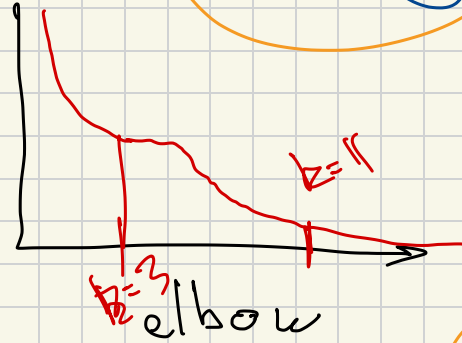
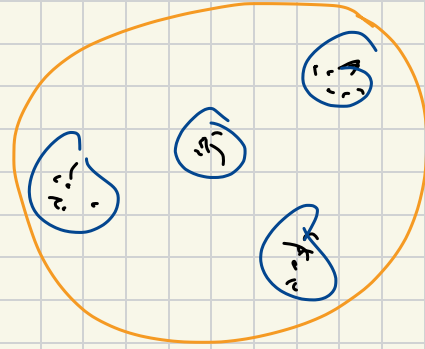
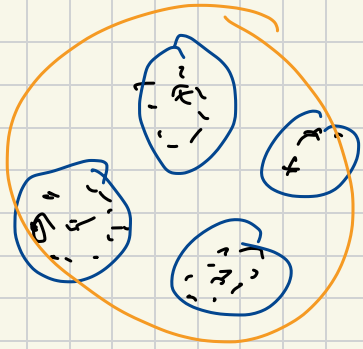


$k=3$



What if data looks like

$k=3$
 $k=11$



4. Bayes Information Criteria

Some likelihood model clusters

$$f(x_i)$$

eg. Mixture of Gaussians
 $\frac{1}{(\pi \sigma^2)^d} \exp\left(-\frac{\|x - s_i\|^2}{2\sigma^2}\right)$

$$f(x) = \prod_{x_i \in x} f(x_i)$$

choose k
to minimize

Information in model $\Rightarrow$ # parameters

σ is fixed k centers $\mathbb{R}^d \Rightarrow kd$ param

$$BIC(M) = -2 \ln(f(M)) + kd \ln(|X|)$$