STAT 391 6/1/23

Lecture 20 and last!

Clustering
- K-means : initialization
- Mixtures + EM algorithm
- other paradigms, non-parametric

Statistics beyond 391

Exam June 8, 10:30
E 125

Web: exam.html → past exam

Grading

List of topics with "links" t.b. posted
< Clustering NO
Mixtures YES

Review sessions
- when ?
- what ?

Grade
8% participation → t.b. posted
12% quiz
30% final
50% HW 1,3,5

No dropping

$$\frac{Q1 + Q2}{Q1^* + Q2^*} \cdot 12 + \frac{Hw1 + Hw5 + Hw3}{Hw1^* + \cdots} \cdot 50 + \cdots$$

drop min $\frac{Qi}{Qi^*}$ or min $\frac{Hw}{Hw^*}$

With drop max $\left\{ \begin{array}{l} \text{drop } Qi \\ \text{drop } Hwi \end{array} \right.$

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

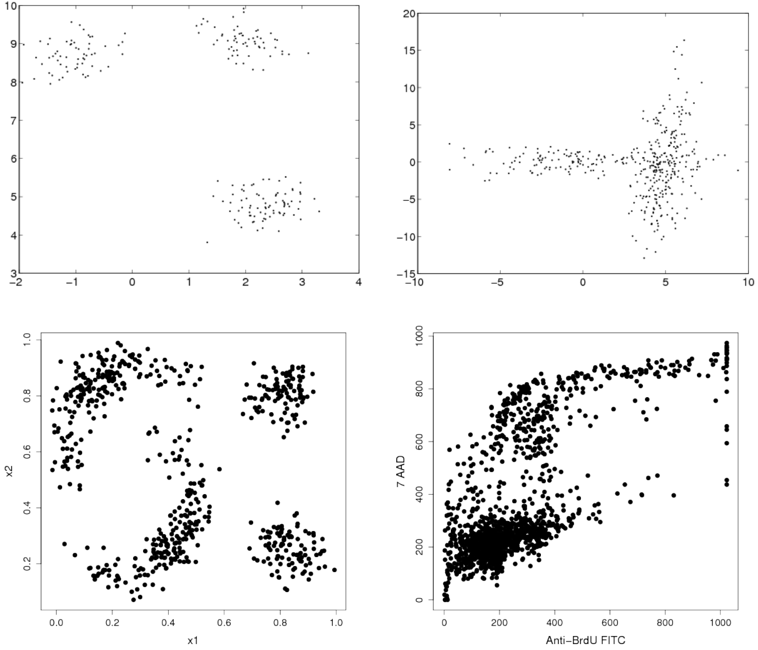
- **Informal definition** **Clustering** = Finding groups in data
- **Notation**

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

- **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)



(from)

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	Cost based [hard]
(K known)	Model based [soft] ← Mixtures of Gaussians
Non-parametric	Dirichlet process mixtures [soft]
(K determined	Information bottleneck [soft]
by algorithm)	Modes of distribution [hard]
	Gaussian blurring mean shift [hard]

- 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]

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

1 - Idea 1: start with K points at random ✗
- Idea 2: start with K data points at random

$G =$

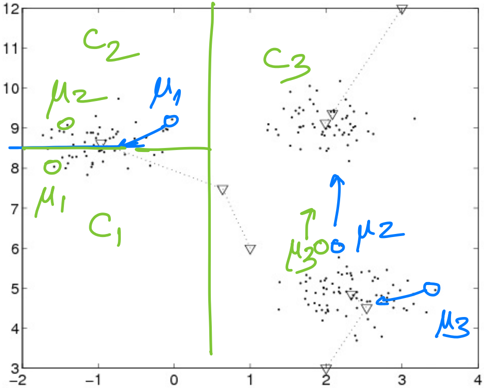
Good, safe: $\{\underbrace{\text{each } C_k}_{\text{true clusters}} \ni \underbrace{\mu_k^0}_{\text{initial centers}}\}$

⇕

URN with n balls in K colors

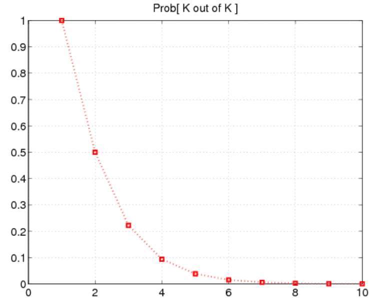
draw K balls = centers

$G = \{K$ balls have K distinct colors$\}$



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)
  selects μ_k^0 depending on $\mu^0_{1:k-1}$
- Idea 4: **k-means++** (randomized, theoretically backed approach to spread out centers)
- Idea 5: **"K-logK" Initialization** (start with enough centers to hit all clusters, then prune down to K)
  For EM Algorithm , for K-means

  selects $\mu^0_{1:K'}$ $K' = K \cdot \log_2 K$

The "K-logK" initialization

The K-logK Initialization (see also)

1. pick $\mu^0_{1:K'}$ at random from data set, where $K' = O(K \log K)$
   (this assures that each cluster has at least 1 center w.h.p)
2. run 1 step of K-means
3. remove all centers μ^0_k that have few points, e.g $|C_k| < \frac{n}{eK'}$
4. from the remaining centers select K centers by **Fastest First Traversal**
   4.1 pick μ_1 at random from the remaining $\{\mu^0_{1:K'}\}$
   4.2 for $k = 2 : K$, $\mu_k \leftarrow \underset{\mu^0_{k'}}{\operatorname{argmax}} \min_{j=1:k-1} ||\mu^0_{k'} - \mu_j||$, i.e next μ_k is furthest away from the already chosen centers
5. continue with the standard **K-means** algorithm

avg $n_k = \frac{n}{K'}$

init

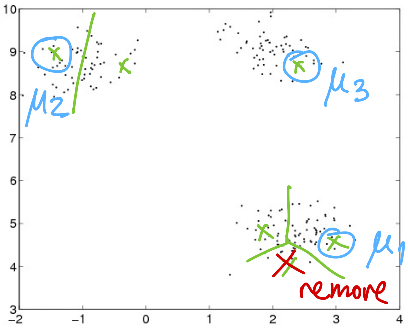
to convergence

$K=3$

$K' = 6 = \lceil 3 \cdot \log_2 3 \rceil$

$K'' = 5$

$K = 3$



K means ++

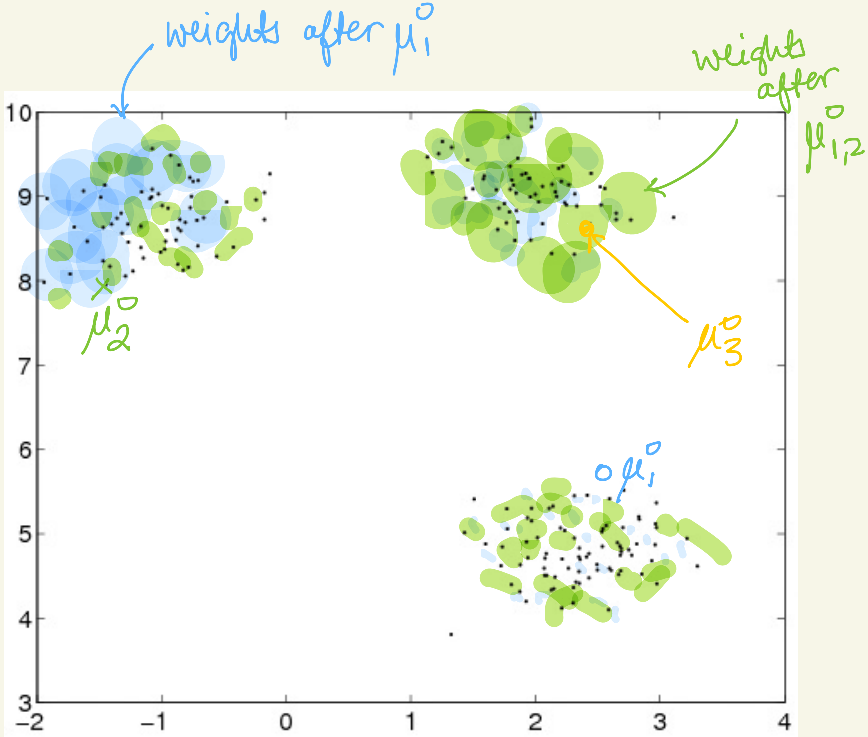
1. select μ_1^0 at random from $\mathcal{D}$

2. for $k = 2 : K$

   for $i = 1 : n$ not yet selected as centers

   $$w_i = \min_{k' = 1:k-1} \| x_i - \mu_{k'}^0 \|^2$$

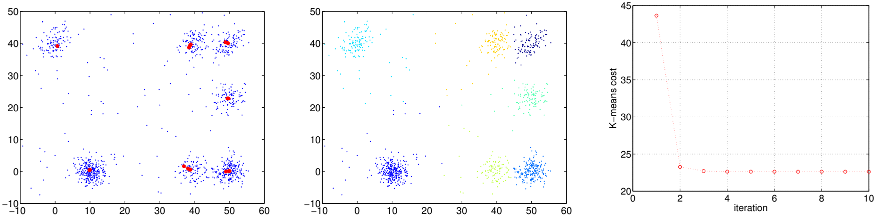
   $\mu_k^0 \leftarrow$ sampled $\propto w_i$



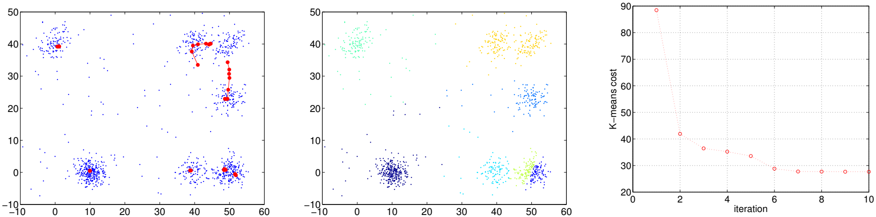
K-means clustering with K-logK Initialization

Example using a mixture of 7 Normal distributions with 100 outliers sampled uniformly

K-logK $K = 7,\ T = 100,\ n = 1100,\ c = 1$



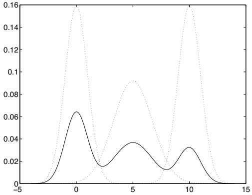
Naive $K = 7\ T = 100,\ n = 1100$



Model based clustering: Mixture models

Mixture in 1D

$K=3$



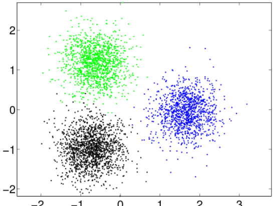
$\pi_1=\frac{1}{2}\quad \pi_2=\frac{1}{3},\ \pi_3=\frac{1}{6}$

- The **mixture density**

$$f(x) \;=\; \sum_{k=1}^{K} \pi_k f_k(x)$$

- $f_k(x)$ = the **components** of the mixture
  - each is a density
  - f called **mixture of Gaussians** if $f_k = Normal_{\mu_k, \Sigma_k}$
- π_k = the **mixing proportions**, $\sum_k = 1^K \pi_k = 1, \quad \pi_k \geq 0.$
- **model parameters** $\theta = (\pi_{1:K},\, \mu_{1:K},\, \Sigma_{1:K})$

Mixture in 2D



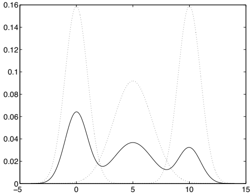
$K=3$

$d=2$

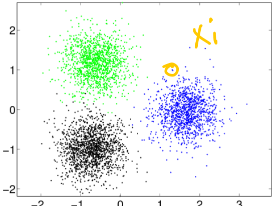
$\pi_1=\pi_2=\pi_3=\frac{1}{3}$

Model based clustering: Mixture models

Mixture in 1D



Mixture in 2D



- The **mixture density**

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

- $f_k(x)$ = the **components** of the mixture
  - each is a density
  - f called **mixture of Gaussians** if $f_k = Normal_{\mu_k, \Sigma_k}$
- π_k = the **mixing proportions**, $\sum_k = 1^K \pi_k = 1, \;\; \pi_k \geq 0.$
- **model parameters** $\theta = (\pi_{1:K}, \, \mu_{1:K}, \, \Sigma_{1:K})$
- The **degree of membership** of point i to cluster k

$$\gamma_{ki} \stackrel{\text{def}}{=} P[x_i \in C_k] = \frac{\pi_k f_k(x_i)}{f(x_i)} \text{ for } i = 1:n, k = 1:K \tag{8}$$

- depends on x_i and on the model parameters

$$\sum_{k=1}^{K} \gamma_{ki} = 1$$

Criterion for clustering: Max likelihood + estimating parameters

- denote $\theta = (\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K})$ (the parameters of the mixture model)
- Define **likelihood** $P[\mathcal{D}|\theta] = \prod_{i=1}^n f(x_i)$
- Typically, we use the **log likelihood**

$$l(\theta) = \ln \prod_{i=1}^n f(x_i) = \sum_{i=1}^n \ln \underbrace{\sum_k \pi_k f_k(x_i)}_{} \tag{9}$$

$N(\mu_k, \sigma_k^2)$

Σ_k

- denote $\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\, l(\theta)$
- θ^{ML} determines a soft clustering γ by (8)
- a soft clustering γ determines a θ (see later)
- Therefore we can write

$$\underset{\text{min}}{\mathcal{L}(\gamma)} = \underset{\text{max}}{-l(\theta(\gamma))}$$

- can't find max θ analytically
- maximize iteratively
- local optima $\exists$

Algorithms for model-based clustering

general alg
Hidden Markov Models
Parse trees – natl lang.
– genetic
...
– missing data

Maximize the (log-)likelihood w.r.t θ

- directly - (e.g by gradient ascent in θ)
- by the EM algorithm (very popular!)
- indirectly, w.h.p. by "computer science" algorithms

w.h.p = with high probability (over data sets)

The Expectation-Maximization (EM) Algorithm

$\Sigma^0_{1:k} = I_d$

$\pi^0_{1:k} = 1/K$

Algorithm Expectation-Maximization (EM)

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

Initialize parameters $\pi_{1:K} \in \mathbb{R}$, $\mu_{1:K} \in \mathbb{R}^d$, $\Sigma_{1:K} \in \mathbb{R}^{d \times d}$ at random[1]

Iterate until convergence

E step (Optimize clustering) for $i = 1 : n$, $k = 1 : K$

Expectation

$$\gamma_{ki} = \frac{\pi_k f_k(x)}{f(x)}$$

M step (Optimize parameters) set $\Gamma_k = \sum_{i=1}^n \gamma_{ki}$, $k = 1 : K$ (number of points in cluster k)

$$\pi_k = \frac{\Gamma_k}{n}, \quad k = 1 : K$$

$$\mu_k = \sum_{i=1}^{n} \frac{\gamma_{ki}}{\Gamma_k} x_i$$

weighted mean

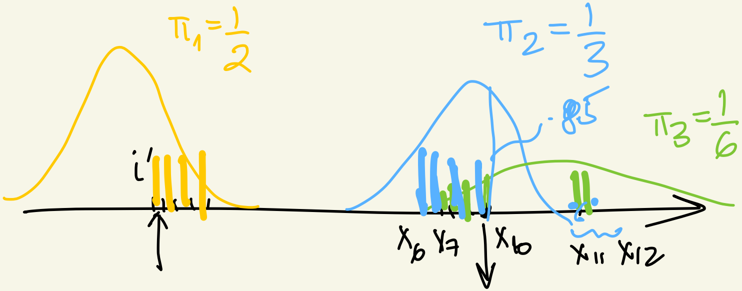
$$\Sigma_k = \frac{\sum_{i=1}^n \gamma_{ki}(x_i - \mu_k)(x_i - \mu_k)^T}{\Gamma_k}$$

weighted covariance

- $\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K}$ are the maximizers of $l_c(\theta)$ in (13)
- $\sum_k \Gamma_k = n$

"alternate maximization algorithm"

[1] Σ_k need to be symmetric, positive definite matrices



$\gamma_1(x_{i'}) \simeq 1$

$\gamma_1(x^i) \cong 0$

$\gamma_3(x^i) \cong 0.2$

$\gamma_2(x^i) \simeq 0.8$

$$\text{"}n_1\text{"} = \Gamma_1 \simeq 5$$

$$\text{"}n_2\text{"} = \Gamma_2 =$$

$$= 0.9 + 0.9 + 0.85 + 0.85 + 0.8 + 0.1 + 0.1 + \varepsilon$$

$$\text{"}n_3\text{"} = \Gamma_3 = 0.1 + 0.1 + 0.15 + 0.15 + 0.2 + 0.9 + 0.5 + \varepsilon$$

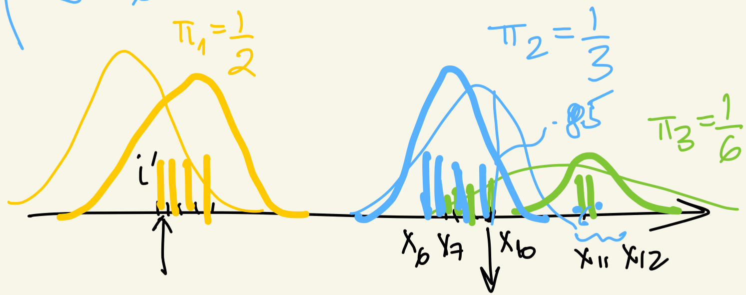
$$\Gamma_k = E[n_k] \qquad k = 1:K$$

$$\pi_2 = \frac{\text{"}n_2\text{"}}{n} = \frac{\Gamma_2}{n} = \frac{\sum \text{weights}}{n} \leftarrow \gamma_2(x^i)$$

$$\mu_2 = \frac{1}{\Gamma_2} \sum_{i=1}^{n} x_i \cdot \gamma_k(x^i) = \frac{1}{\Gamma_2}\left(0.9\, x_6 + 0.9\, x_7 + \cdots + 0.1\, x_{12}\right)$$

$\mu_3 \leftarrow \cdots$

$\Gamma_3^2 \leftarrow \cdots$



[Supplement: The EM Algorithm – Motivation]

- Define the **indicator variables**

$$z_{ik} = \begin{cases} 1 & \text{if } i \in C_k \\ 0 & \text{if } i \notin C_k \end{cases} \qquad (10)$$

  denote $\bar{z} = \{z_{ki}\}_{k=1:K}^{i=1:n}$
- Define the **complete log-likelihood**

$$l_c(\theta, \bar{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ki} \ln \pi_k f_k(x_i) \qquad (11)$$

- $E[z_{ki}] = \gamma_{ki}$
- Then

$$E[l_c(\theta, \bar{z})] = \sum_{i=1}^{n} \sum_{k=1}^{K} E[z_{ki}][\ln \pi_k + \ln f_k(x_i)] \qquad (12)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ki} \ln \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ki} \ln f_k(x_i)] \qquad (13)$$

- If θ known, γ_{ki} can be obtained by (8) **(Expectation)**
- If γ_{ki} known, π_k, μ_k, Σ_k can be obtained by separately maximizing the terms of $E[l_c]$ **(Maximization)**

Brief analysis of EM

$$Q(\theta, \gamma) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ki} \ln \underbrace{\pi_k f_k(x_i)}_{\theta}$$

- each step of EM increases $Q(\theta, \gamma)$
- Q converges to a local maximum ✔
- at every local maxi of Q, $\theta \leftrightarrow \gamma$ are fixed point
- $Q(\theta^*, \gamma^*)$ local max for $Q \Rightarrow l(\theta^*)$ local max for $l(\theta)$
- under certain regularity conditions $\theta \longrightarrow \theta^{ML}$
- the E and M steps can be seen as projections

- Exact maximization in **M step** is not essential.
  Sufficient to increase Q.
  This is called **Generalized EM**

Generalized to many situations

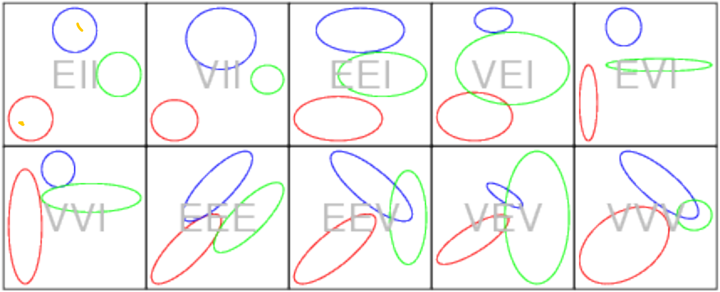
The M step in special cases

- Note that the expressions for μ_k, Σ_k = expressions for μ, Σ in the normal distribution, with data points x_i weighted by $\frac{\gamma_{ki}}{\Gamma_k}$

	M step
general case	$\Sigma_k = \sum_{i=1}^n \frac{\gamma_{ki}}{\Gamma_k}(x_i - \mu_k)(x_i - \mu_k)^T$
$\Sigma_k = \Sigma$ "same shape & size" clusters	$\Sigma \leftarrow \frac{\sum_{i=1}^n \sum_{k=1}^K \gamma_{ki}(x_i-\mu_k)(x_i-\mu_k)^T}{n}$
$\Sigma_k = \sigma_k^2 I_d$ "round" clusters	$\sigma_k^2 \leftarrow \frac{\sum_{i=1}^n \gamma_{ki}\|\|x_i-\mu_k\|\|^2}{d\Gamma_k}$
$\Sigma_k = \sigma^2 I_d$ "round, same size" clusters	$\sigma^2 \leftarrow \frac{\sum_{i=1}^n \sum_{k=1}^K \gamma_{ki}\|\|x_i-\mu_k\|\|^2}{nd}$

Exercise Prove the formulas above

- Note also that **K-means** is **EM** with $\Sigma_k = \sigma^2 I_d$, $\sigma^2 \to 0$ Exercise Prove it



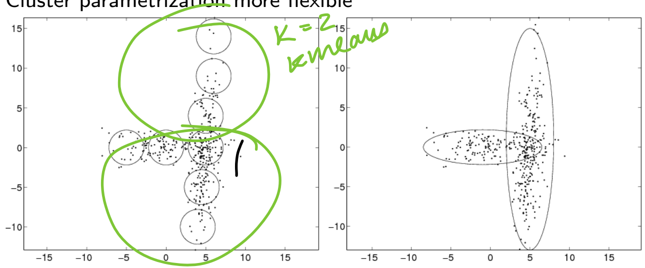
More special cases introduce the following description for a covariance matrice in terms of *volume, shape, alignment with axes* (=determinant, trace, e-vectors). The letters below mean: I=unitary (shape, axes), E=equal (for all k), V=unequal

- 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

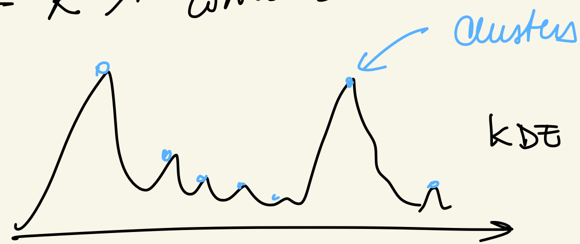


- Converges to local optimum of **log-likelihood**
  Initialization recommended by K-logK method

- **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)

Stats beyond 391

Non-parametric clustering — $k \nearrow$ with n



Dependent data

- streaming data
- sequences — language, text, speech
  - DNA, proteins
- networks
- curves

• Reinf Learning, Bandits = Adverts

• Causal inference

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 → Model Selection

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

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

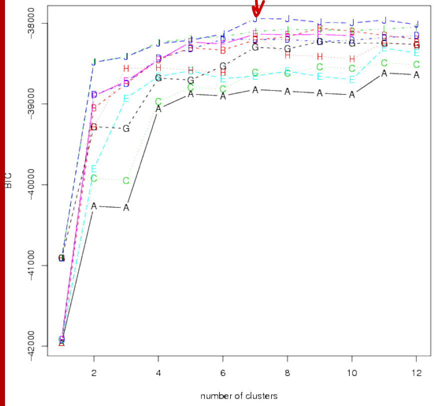
- define

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

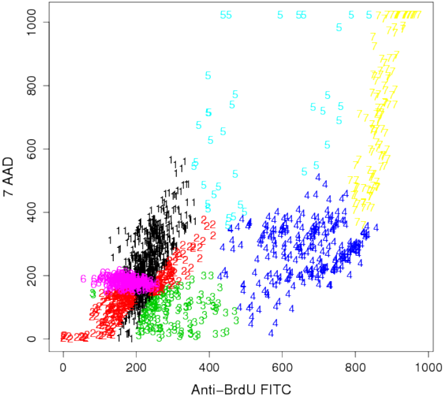
- **Select** K that maximizes $BIC(\theta_K)$
- selects true K for $n \to \infty$ and other technical conditions (e.g parameters in compact set)
- but theoretically not justified (and overpenalizing) for finite n

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)

Best model



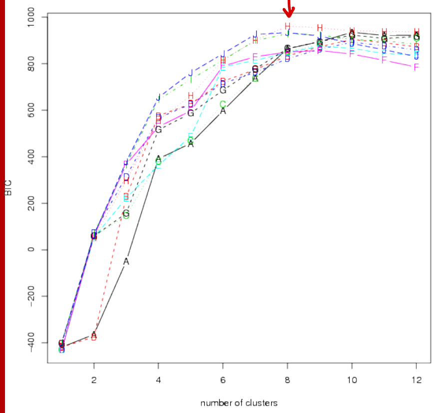
EEV, 8 Cluster Solution



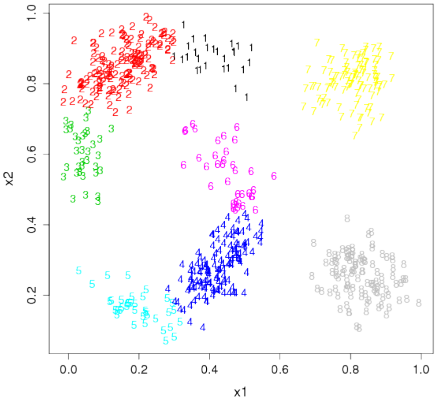
(from)

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)

Best model



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