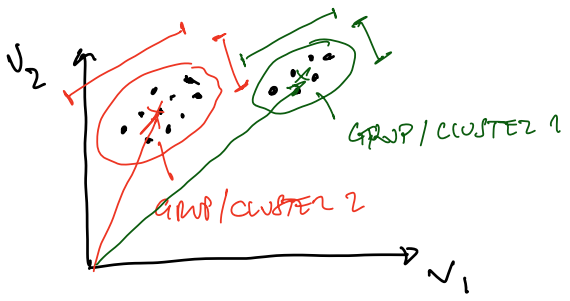
Tema 2: TÈCNIQUES D'AGRUPAMENT DE DADES (CLUSTERING ALGORITHMS)

└ NO-SUPERVISAT

$X_{NOBS \times NVAR}$ (dades numèriques)

Ex: NOBS = 16
NVAR = 2



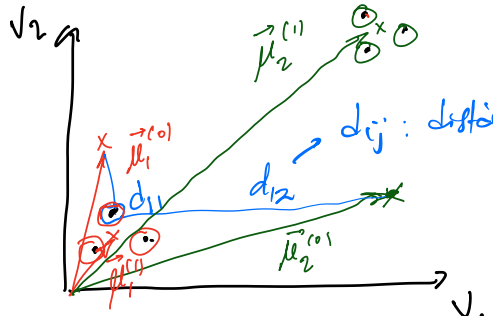
NOBS = 10^4
NVAR = 18
} → ?

Algorismes →
- K-means (K-centroides) → distàncies entre observacions
- Agrupament Aglomeratiu → distàncies entre observacions
- Models de mescla Gaussiana (GMM) → distribució probabilística pròpia

↓
HOW MANY CLUSTERS?

(1) K-MEANS : Algorisme reqüencial (iteratiu)

Ex:



[1] Establir # clusters que volem identificar
[2] Definir centroides aleatoris (k) per cadascun dels K grups

$\{\vec{\mu}_1^{(0)}, \vec{\mu}_2^{(0)}, \dots \vec{\mu}_K^{(0)}\}$

↓
centroide cluster 1 iteració (0)

[3] Assigna cada observació al cluster que tingui el centroide més proper.

[4] Recalcula centroides $\{\vec{\mu}_1^{(1)}, \vec{\mu}_2^{(1)}, \ldots, \vec{\mu}_K^{(1)}\}$

goto [3] until $|\Delta \vec{\mu}_i| < \varepsilon$

↓

clustering final

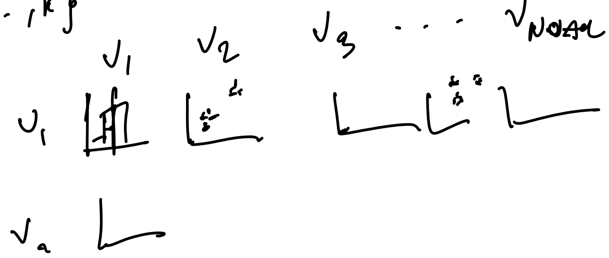
$Y = \{y_1, y_2, \ldots, y_{NOBS}\}$

↓

A quin cluster pertany l'observació 1

$y_1 \in \{1, \ldots, k\}$

K? → pairplot:

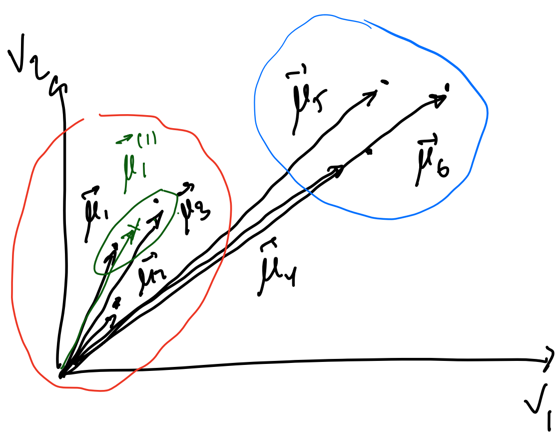


② AGRUPAMENT AGLOMERATIU

↳ Hierarchical clustering

↳ aglomeratiu

↳ divisiu



inicial clusters = # observacions

↓

agrupar clusters fins que només quedin (k)

[1] Indica # clusters K

[2] Inicialitza #clusters → # observacions

[3] Agrupa els 2 clusters més propers

$(\vec{\mu}_1, \vec{\mu}_3) \rightarrow \vec{\mu}_1^{(1)}$

$\vec{\mu}_2 \rightarrow \vec{\mu}_2^{(1)}$

$\vec{\mu}_3 \rightarrow X$

$\vec{\mu}_4 \rightarrow \vec{\mu}_3^{(1)}$

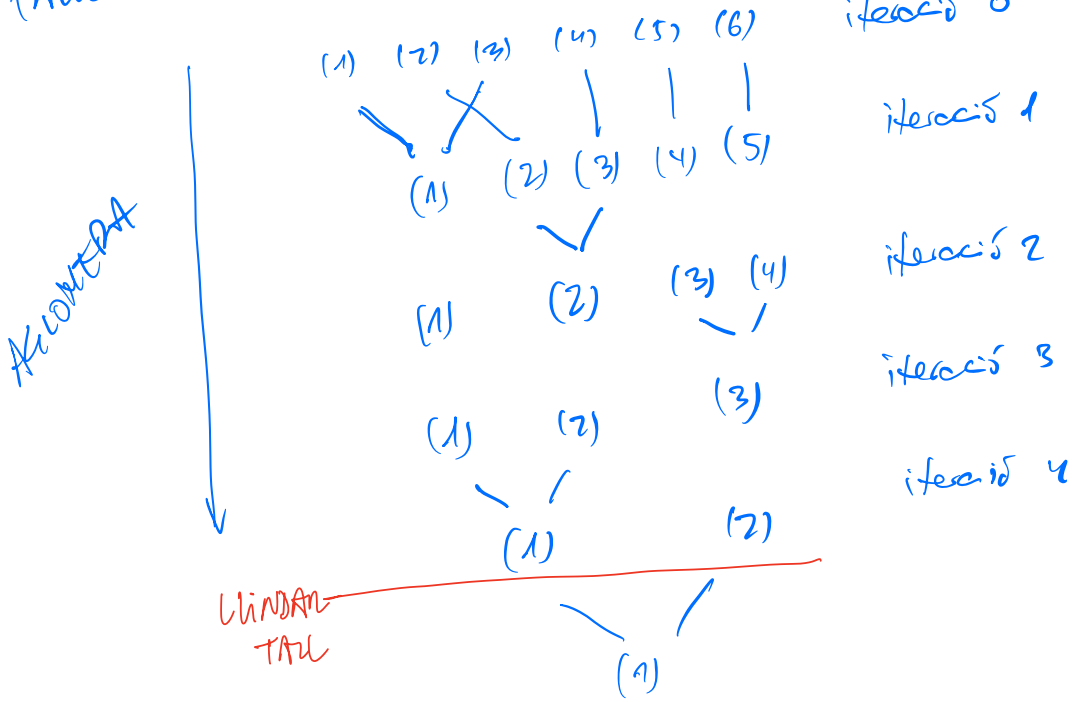
$\vec{\mu}_5 \rightarrow \vec{\mu}_4^{(1)}$

$\vec{\mu}_6 \rightarrow \vec{\mu}_5$

5 clusters (#obs − 1)

[4] Repeat until #clusters = K

La jerarquia d'agrupaments es pot representar gràficament en forma de DENDROGRAMA



③ MODELS DE MESCLA GAUSSIANA (GAUSSIAN MIXTURE MODELS)

$X_{NOBSS \times NVAR}$ (numèrics)

$= \longrightarrow \vec{x} \in \mathbb{R}^{NVAR}$

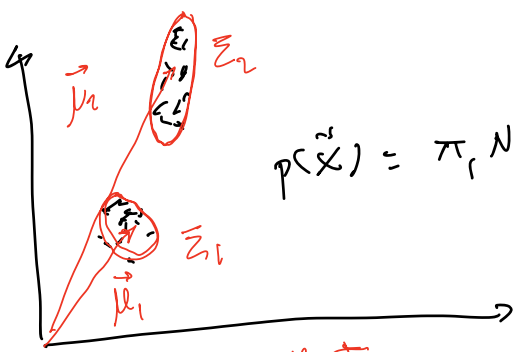
$$p(\vec{x}) = \sum_{i=1}^{NG} \pi_i \cdot N(\vec{\mu}_i, \Sigma_i)$$

- $p(\vec{x})$: pdf observacions
- π_i: pesos
- $\vec{\mu}_i$: centroide, $NVAR \times 1$
- Σ_i: matriu covariança, $NVAR \times NVAR$

descriu la distribució de les observacions a l'espai NVAR dimensional com una suma ponderada de gaussianes multivariades

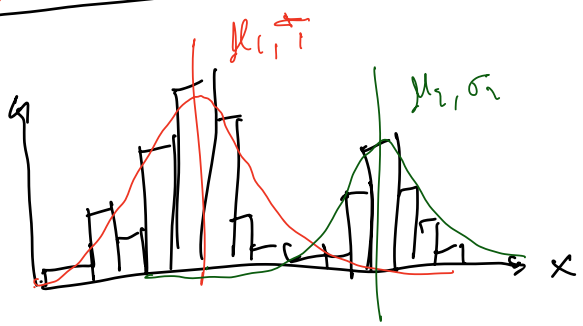
Ex:

2D



$$p(\vec{x}) = \pi_1 N(\vec{\mu}_1, \Sigma_1) + \pi_2 N(\vec{\mu}_2, \Sigma_2)$$

1D



$$p(x) = \pi_1 N(\mu_1, \sigma_1) + \pi_2 N(\mu_2, \sigma_2)$$

$$p(\vec{x}) = \sum_{i=1}^{NG} \pi_i N(\vec{\mu}_i, \Sigma_i) \rightarrow \text{paràmetres a estimar.}$$

VARIABLE LATENT
↓
assignació de observacions → Covariances
↓
Etiqueta de clustering

ESTIMACIÓ PARÀMETRES:
Algorisme d'Expectació-Maximització
(EM-algorithm)

Dades → GMM NG → clustering variable

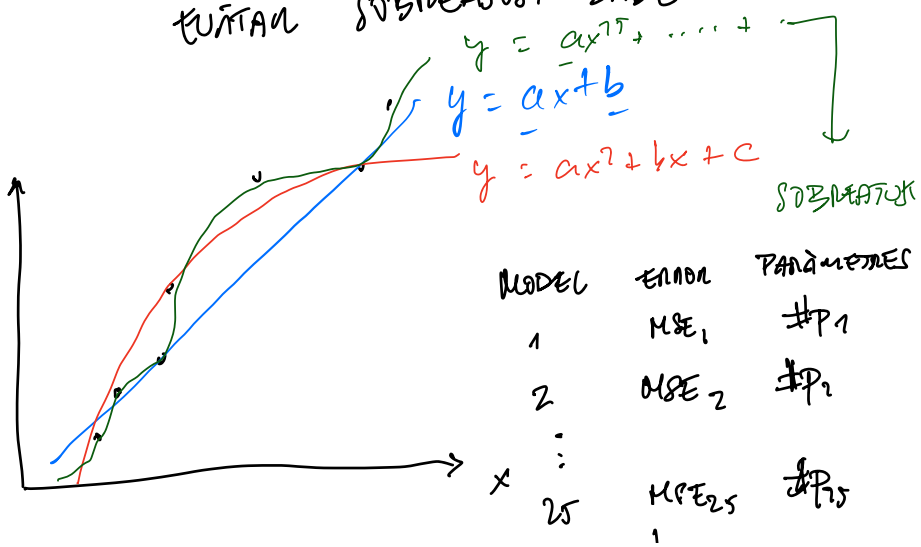
How many clusters? → MODEL SELECTION
Índex de puntuació (Bayesian Information Criterion / Akaike I.C.)

I ~ error model (MSE) − complexitat model (#paràmetres)

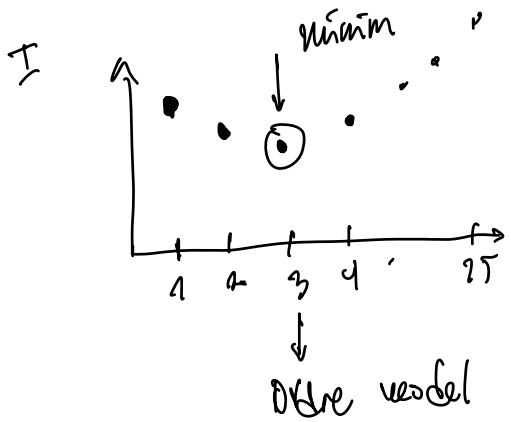
↓

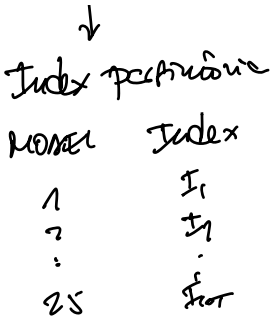
EVITAR SOBREAJUST DADES

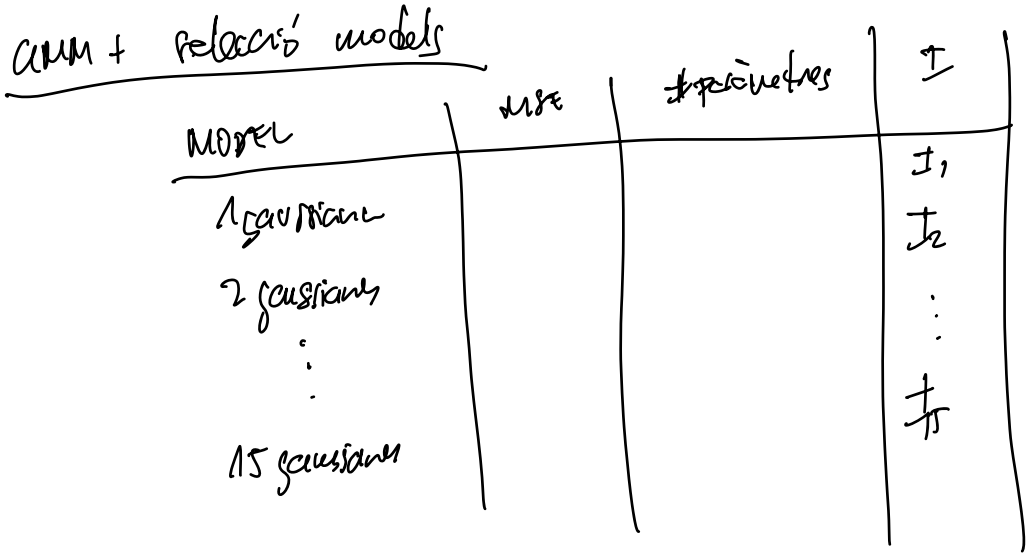
Ex: regressió polinòmica

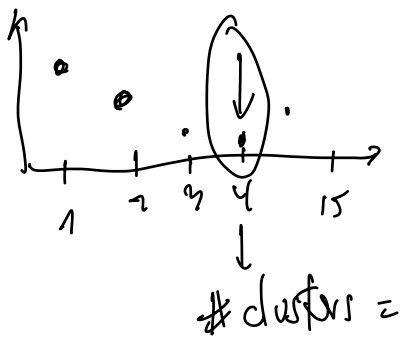


MODEL	ERROR	PARÀMETRES
1	MSE_1	$\#p_1$
2	MSE_2	$\#p_2$
⋮		
25	MSE_{25}	$\#p_{25}$









X → models GMM → model guanyador (mínim BIC) → clustering + # clusters