T1.1 → Intro data analysis & ML

T1.2 → CLUSTERING ALGORITHMS

↳ K-means

↳ Hierarchical clustering

↳ Gaussian Mixture Models (GMM)

- Model selection
- Parameter estimation in probabilistic models.

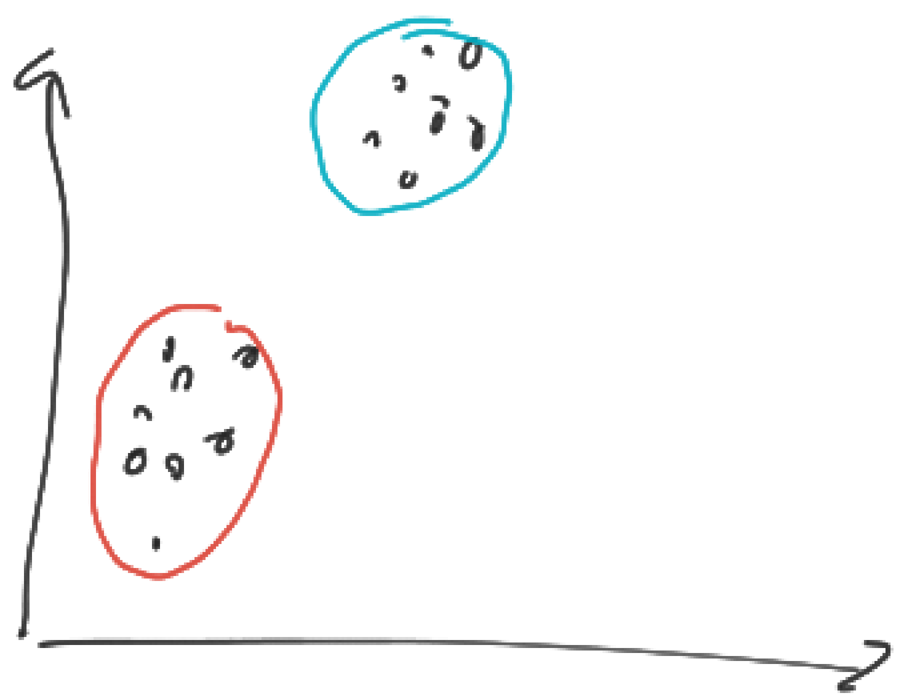
CLUSTERING

d-dimensional feature space $\vec{x} \in \mathbb{R}^d$

From a set of observations

$$\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots, \vec{x}_N\}$$

$d = 2$



UNSUPERVISED METHODS !

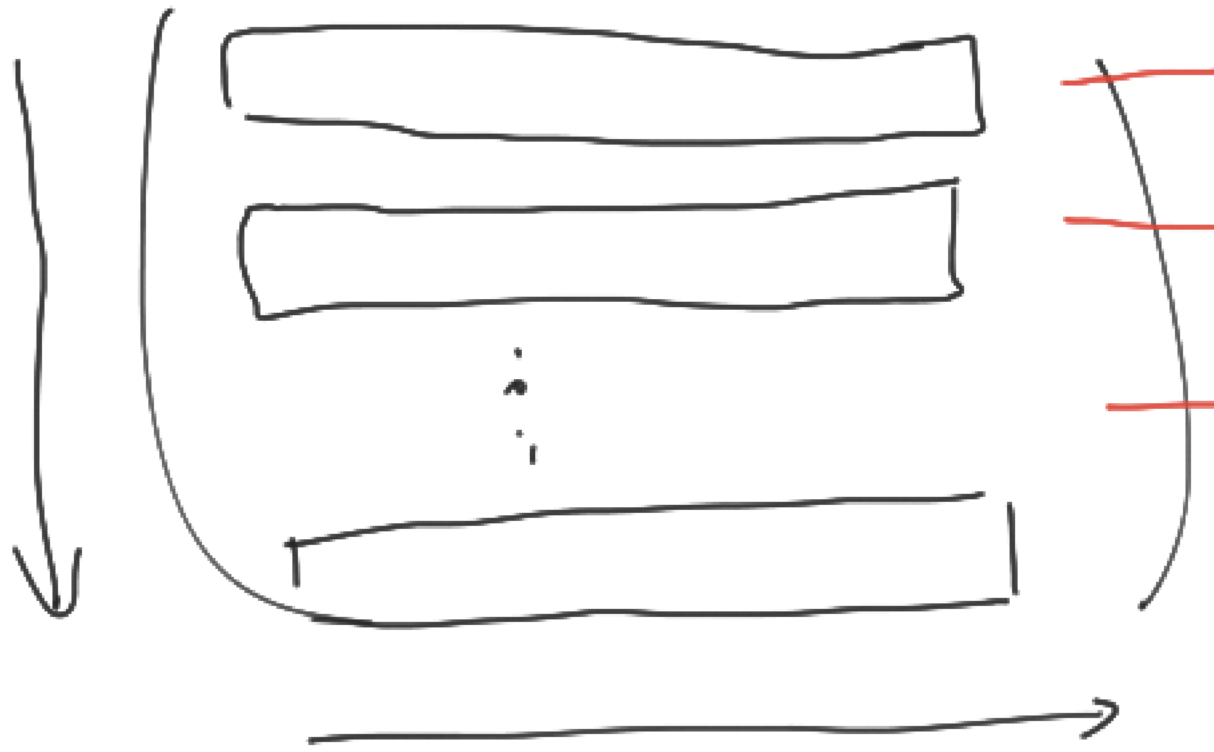
DATA ML $\longrightarrow$ CLASS MEMBERSHIP OF EACH OBSERVATION

data matrix

observations $\times$ features

$\omega_j \quad j = 1 \dots$ NOBS

obs



features | variable | attributes

$\omega_1 =$ 'FCB supporter'

$\omega_2 =$ 'other team'

$\omega_3 =$ 'other team'

$\vdots$

$\omega = \{$'FCB supporter', 'other team'$\}$

data matrix $\longrightarrow$ unsupervised techniques

data matrix

class-label vector

$\{\omega_1, \omega_2, \dots \omega_{NOBS}\}$

$\longrightarrow$ supervised techniques

$\downarrow$

TRAIN / FITTING / LEARNING

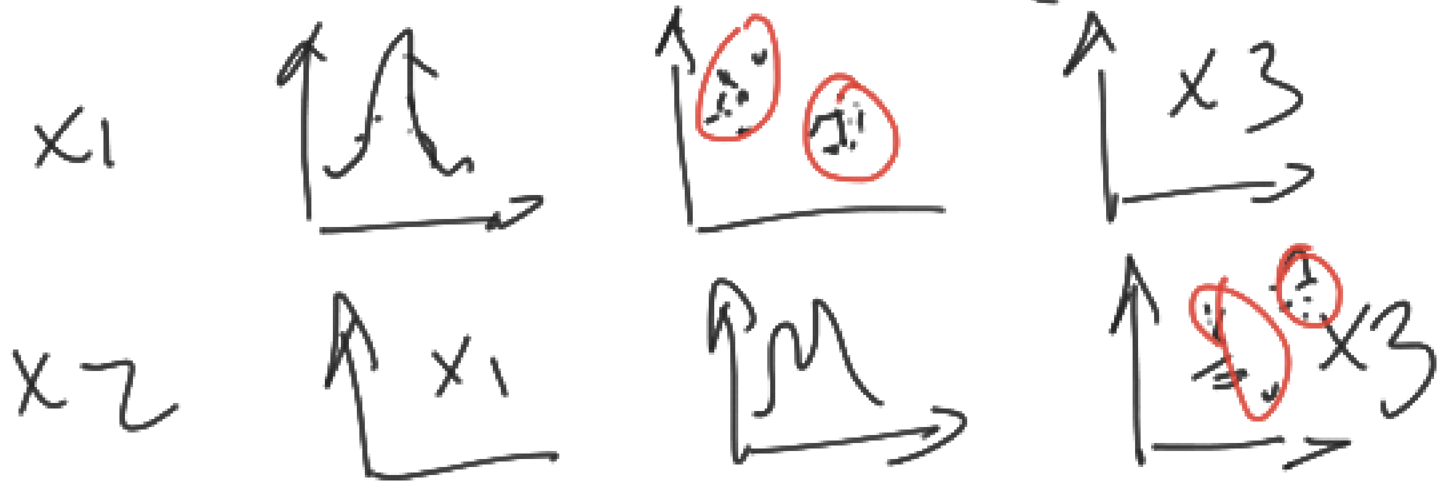
(ML)

k-means → require pre-define #clusters
Hierarchical to be found in data

2d → easy

>2d → plot data using a pairplot

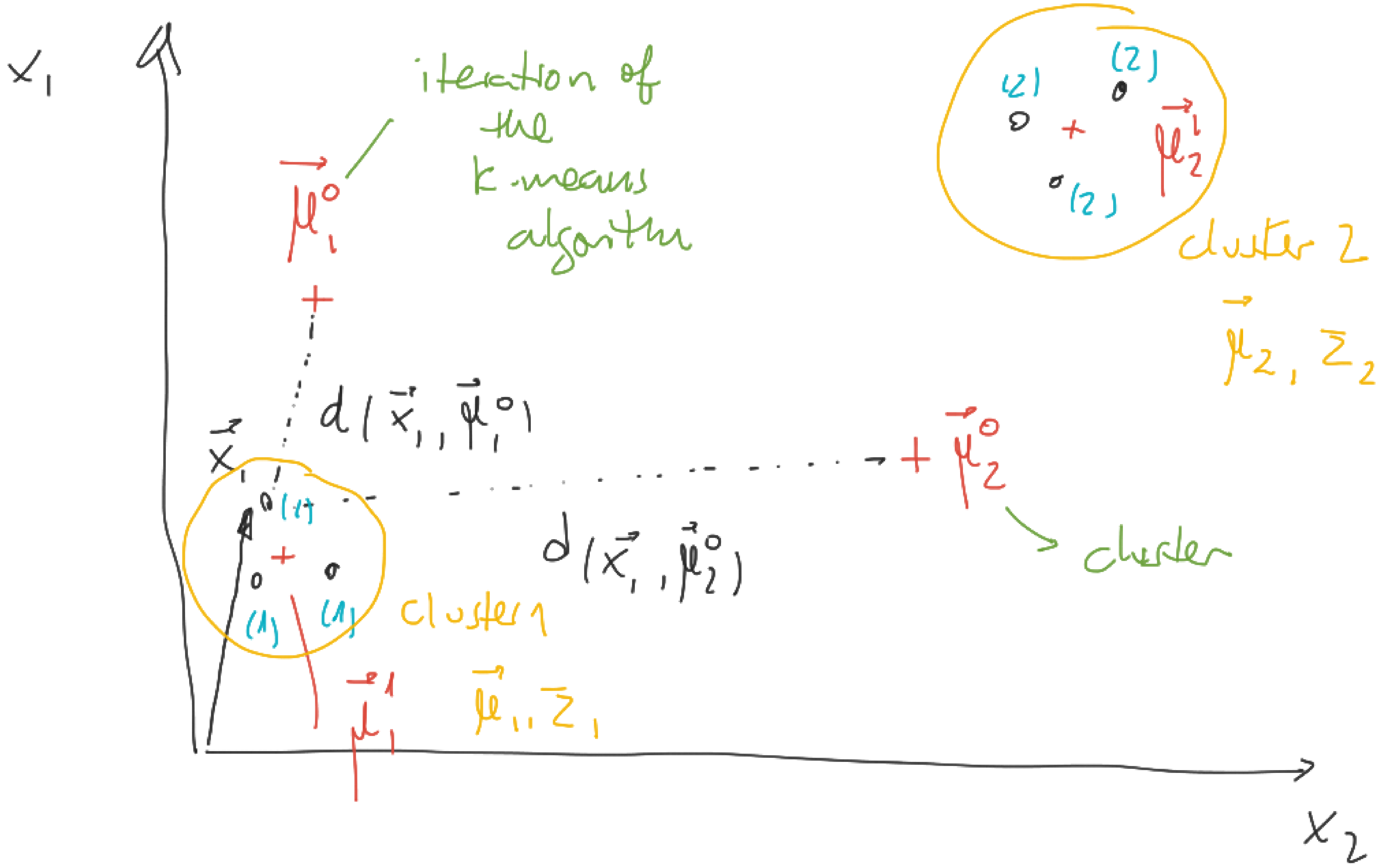
x1 x2 (scatter plot matrix)



k-means

$d = 2$

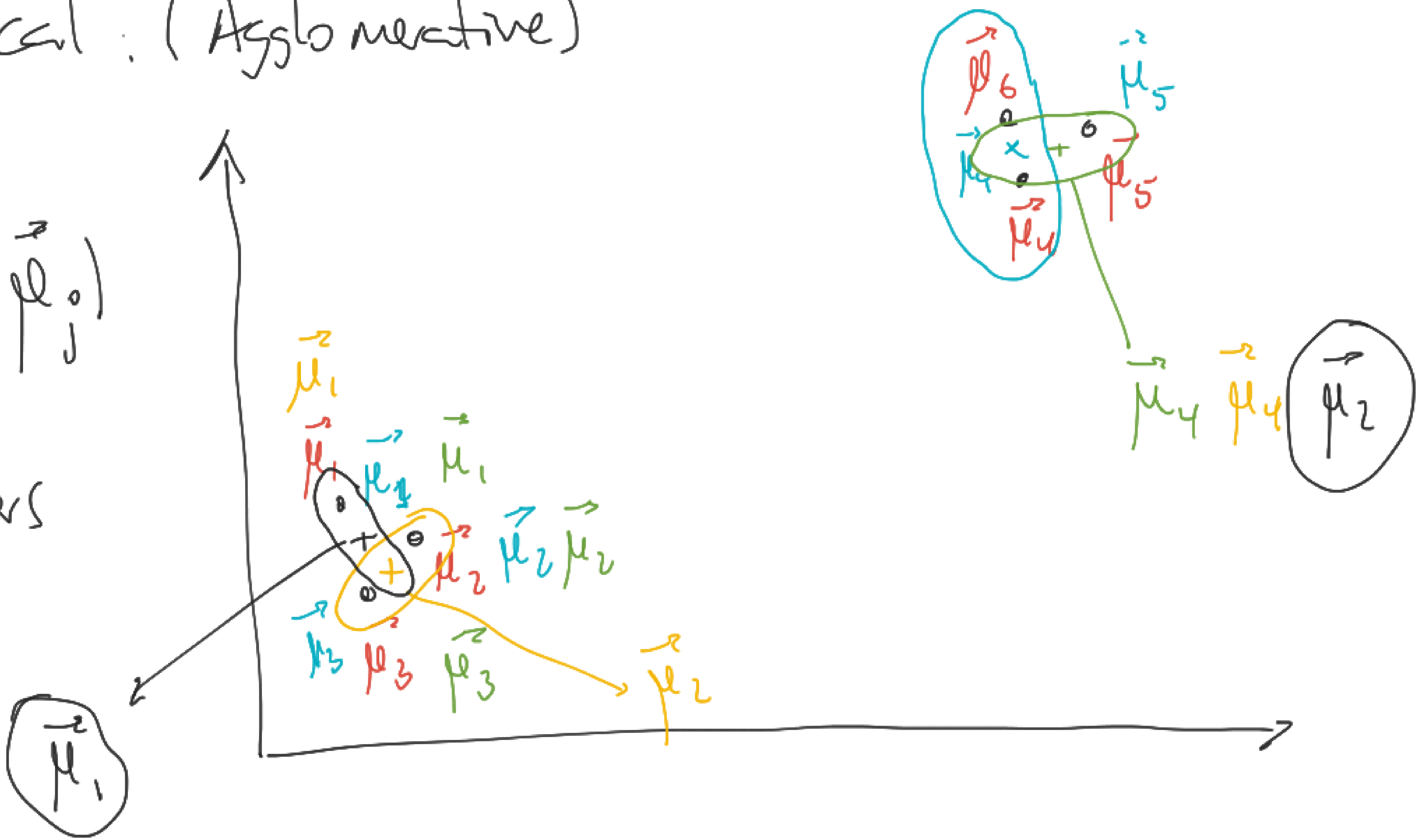
$\vec{\mu}_1 , \vec{\mu}_2$?

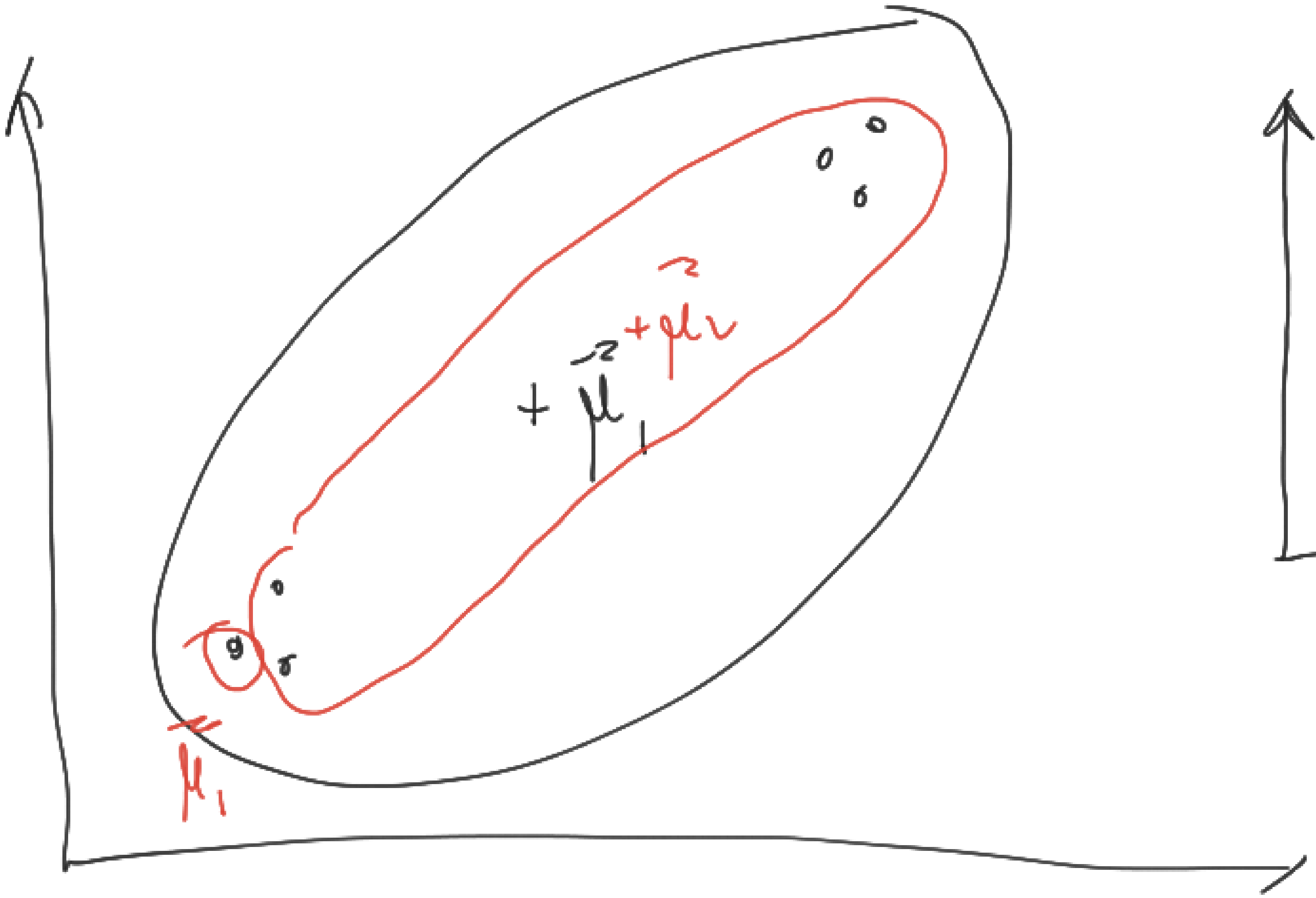


Hierarchical : (Agglomerative)

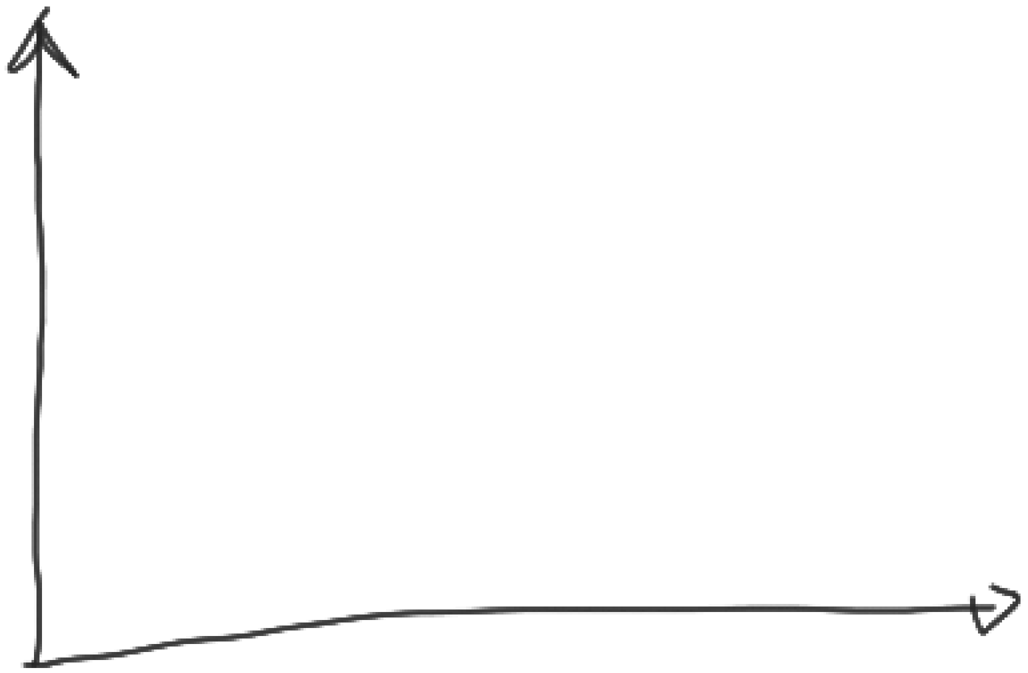
$\mathrm{pdist}(\vec{\mu}_i, \vec{\mu}_j)$

$K = 2$ clusters





$\vec{\mu}_1$
$\vec{\mu}_2$
$\vec{\mu}_1$



GMM $\quad \vec{x} \in \mathbb{R}^d$

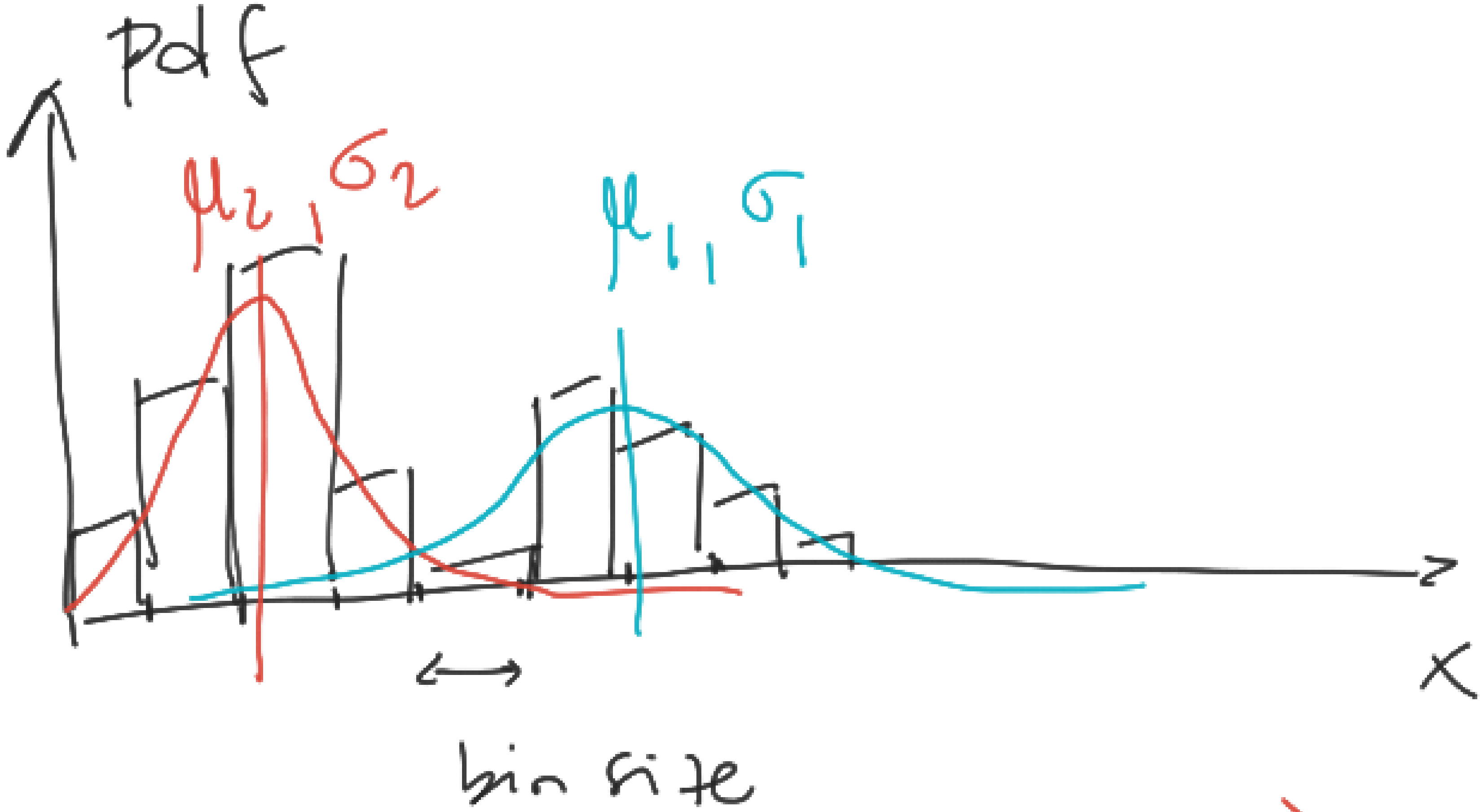
$$P(\vec{x}) = \sum_{j=1}^{G} \underbrace{\pi_j}_{\text{weights}} \cdot N(\vec{\mu}_j, \Sigma_j)$$

$\vec{\mu}_j$ → mean of gaussian j $(1 \times d)$

Σ_j → covariance matrix of gaussian j $(d \times d)$

Ex: $d = 1$ (univariate observ.) #

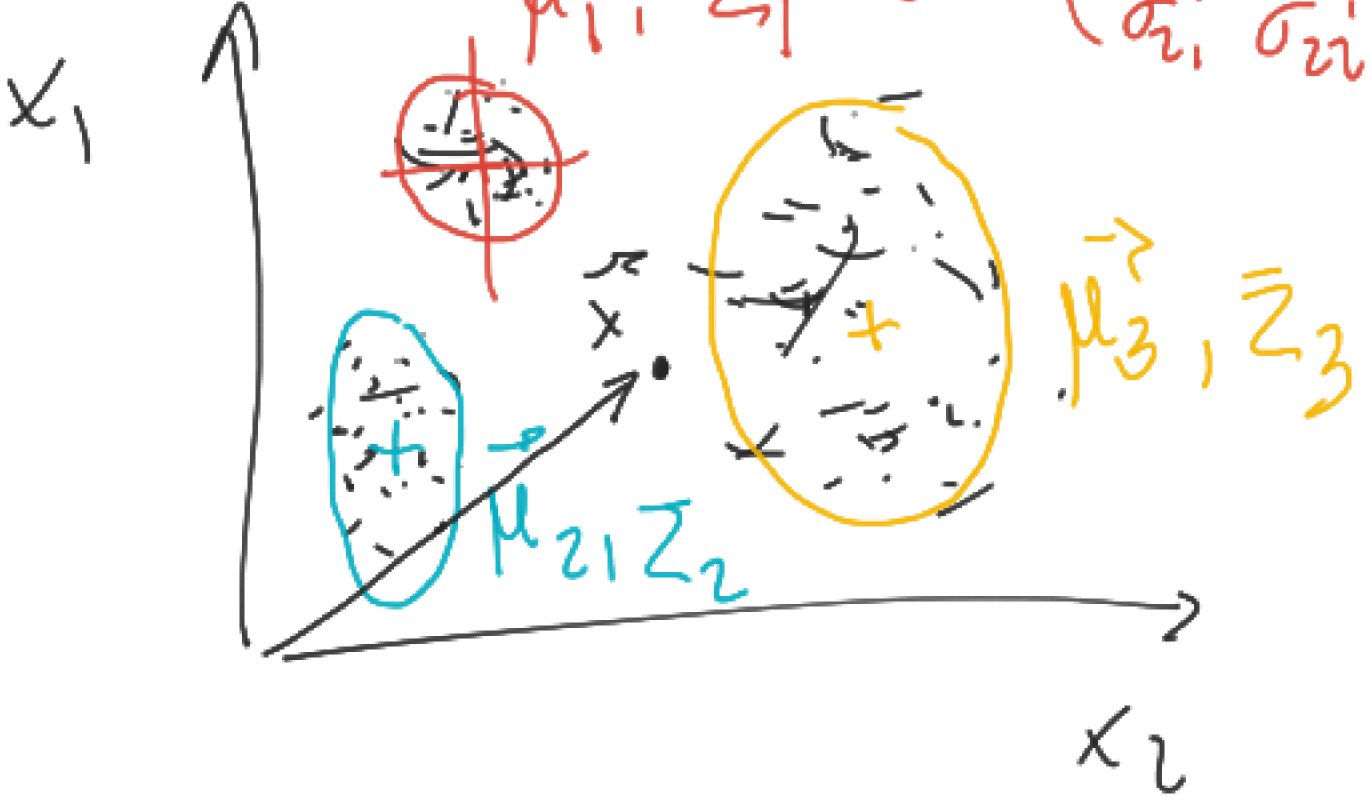
$G = 2$ gaussians



$\vec{\mu}_1, \Sigma_1 = \begin{pmatrix} \sigma_{11}^1 & \sigma_{12}^1 \\ \sigma_{21}^1 & \sigma_{22}^1 \end{pmatrix}$

Ex. $d = 2$

$G = 3$



$$p(\vec{x}) = \sum_{j=1}^{3} \pi_j \, N(\vec{\mu}_j, \Sigma_j)$$

GMM → pre-define # clusters → # gaussians

parameters (GMM)

MODEL COMPLEXITY

$K=1$ $\quad \vec{\mu}_1, \Sigma_1 \quad \underset{\vec{\mu}_1}{d} + \underset{\Sigma_1}{d \times d}$

$K=2$ $\quad \vec{\mu}_1, \Sigma_1, \vec{\mu}_2, \Sigma_2, \pi_1 \quad (\pi_1 + \pi_2 = 1) \quad \underbrace{2d}_{\vec{\mu}_1, \vec{\mu}_2} + \underbrace{2d \cdot d}_{\Sigma_1, \Sigma_2} + \underbrace{1}_{\pi_1}$

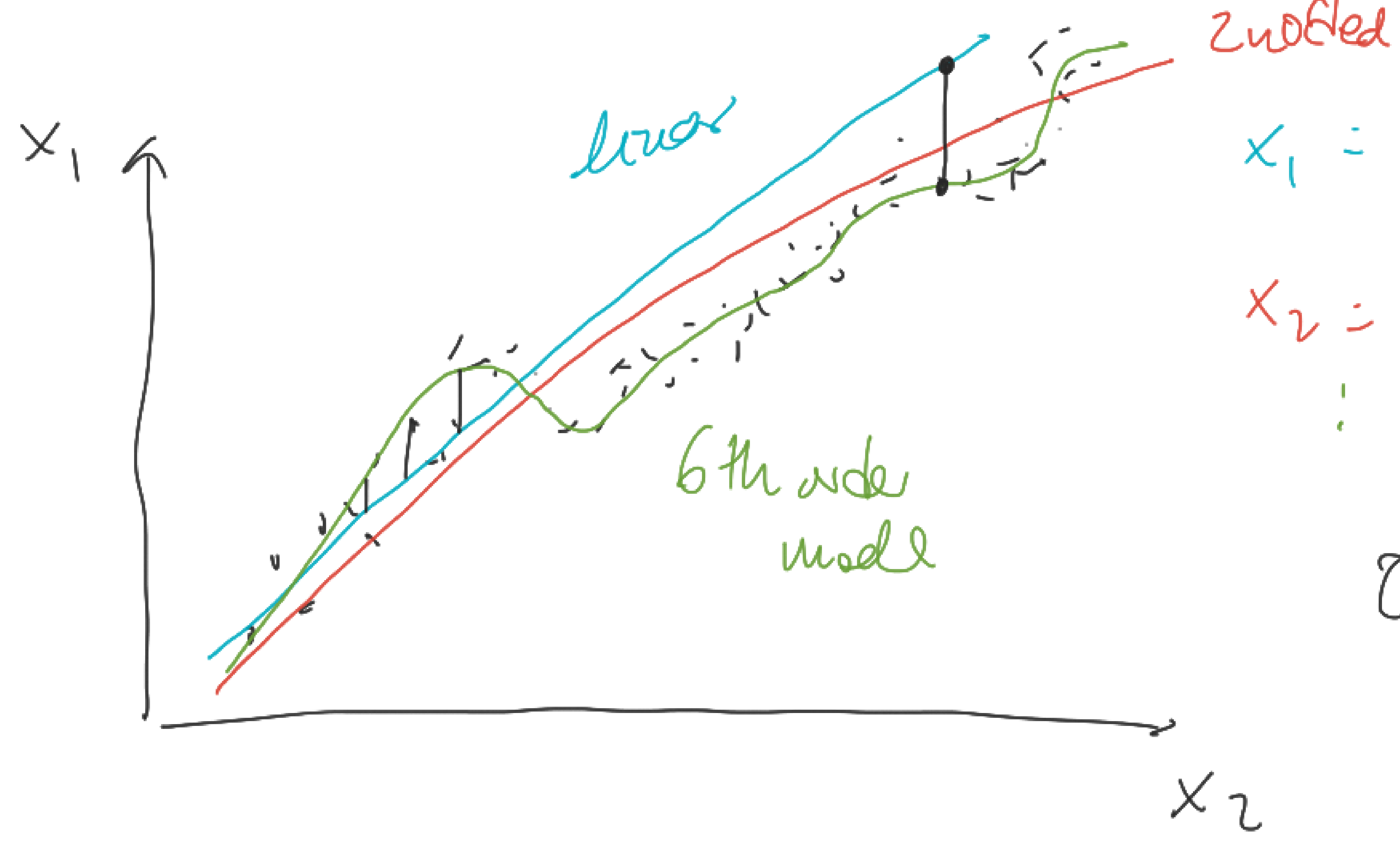
$\vdots$

PARSIMONY INDICES

You need several observations to fit a GMM model

PARSIMONY

= MODEL PERFORMANCE (RESIDUALS) (ERROR) MSE vs MODEL COMPLEXITY # parameter

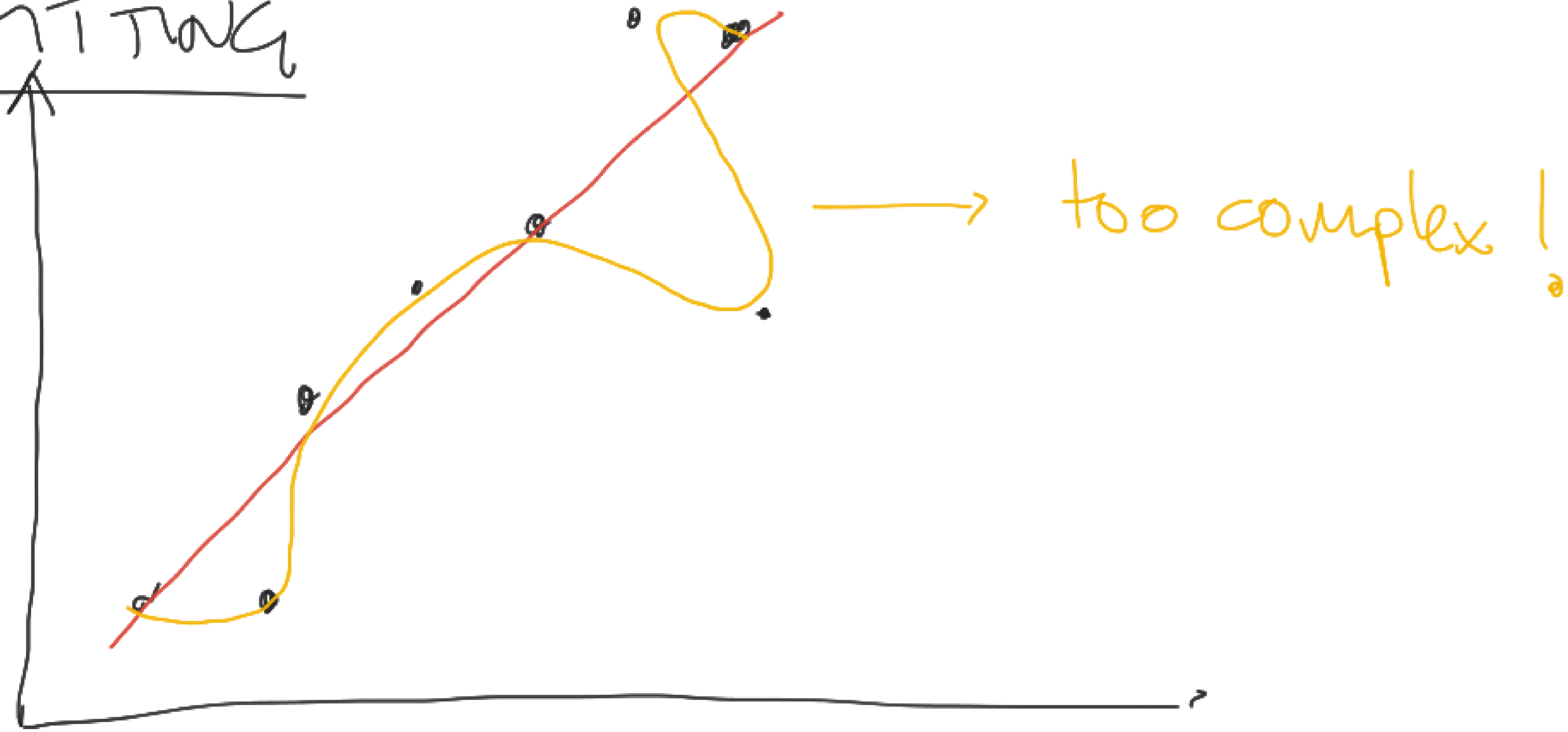


$x_1 = a x_2 + b$ 2

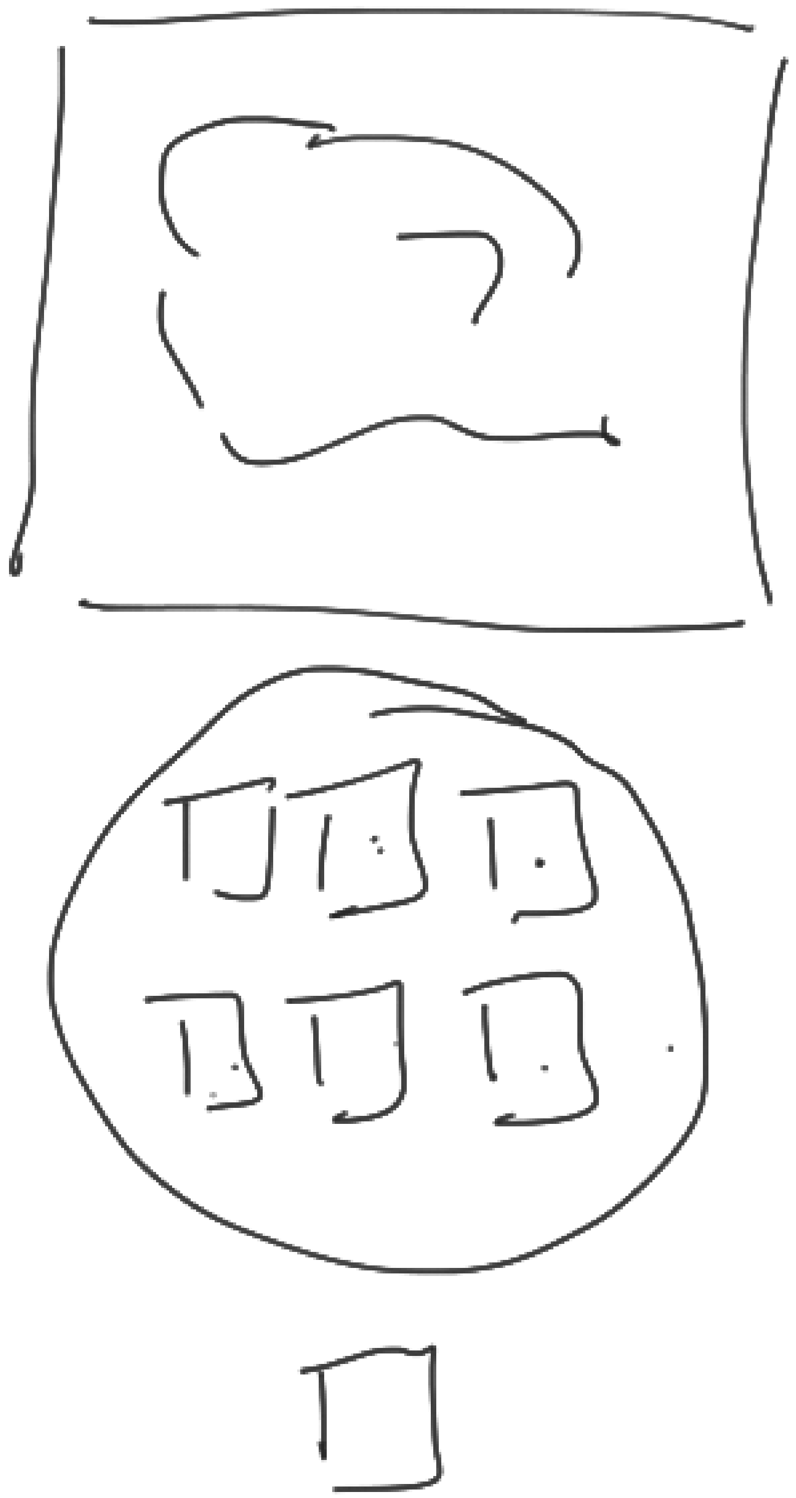
$x_2 = a x_2^2 + b x_2 + \dots$ 3

avoid data overfitting

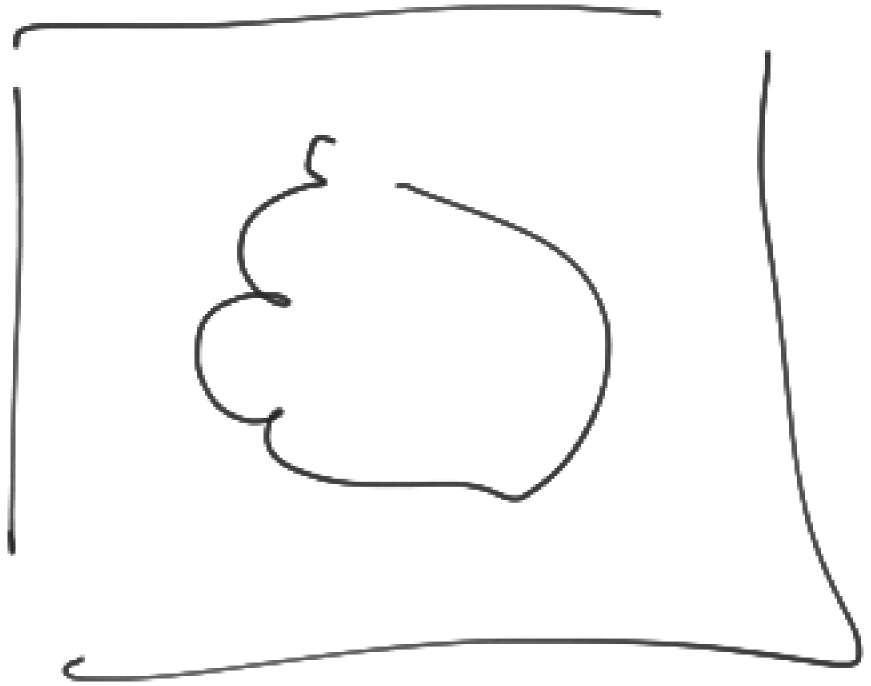
OVERFITTING



R



J



OVERFITTING $\Longleftrightarrow$ GENERALIZATION

GMM + MODEL SELECTION

BIC
PARSIMONY INDEX

DATA

fit GMM (G=1) — BIC$_1$ -748

fit GMM (G=2) — $\boxed{BIC_2 \quad -800}$

fit GMM (G=3) — BIC$_3$ -790

fit GMM (G=4) — BIC$_4$ -796

BIC = - error + complexity

$-2\ln(\hat{\mathcal{L}}) + k\ln(n)$

ML

parameters

How many clusters?

$\arg\max_{j=1..4} \{BIC_j\} \rightarrow$ 2 clusters

PARAMETER ESTIMATION IN PROBABILISTIC MODELS

Probabilistic model : $\vec{x}$: data observation (d-dimensional)

θ : model parameters

$\downarrow$

$p(\vec{x}, \theta)$ → joint pdf

Ex: $p(\vec{x} \mid \theta) = N(\vec{x} \mid \underbrace{\vec{\mu}, \bar{\Sigma}}_{\theta})$

θ : model parameters

$$p(\vec{x} \mid \theta) = \sum_{j=1}^{G} \pi_j \cdot N(\vec{\mu}_j, \bar{\Sigma}_j)$$

PARAMETER ESTIMATION :

sample of observations $\{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\} \longrightarrow \theta^*$ SAMPLE ESTIMATE of model parameters

→ MAXIMUM LIKELIHOOD ESTIMATE (ML)

Gaussian multivariate (unknown $\vec{\mu}$) → $\theta = \vec{\mu}$ (Σ is known)

likelihood function $\quad P(\vec{x} \mid \vec{\mu}) = N(\vec{x} \mid \vec{\mu}, \Sigma)$

define log-likelihood $\mathcal{L} = \ln P(\vec{x}|\vec{\mu})$ $\quad \mathcal{N}(\vec{x}|\vec{\mu}, \bar{\Sigma}) = \frac{1}{(2\pi)^d |\bar{\Sigma}|^{+\frac{1}{2}}}$

$$\mathcal{L} = -\frac{1}{2}\ln\left[(2\pi)^d \cdot |\bar{\Sigma}|\right] - \underbrace{\frac{1}{2}(\vec{x}-\vec{\mu})^T \cdot \bar{\Sigma}^{-1} \cdot (\vec{x}-\vec{\mu})}_{\text{mahalanobis distance}}$$

(quadratic weighted with variance in each direction)

Maximum likelihood:

$$\partial_{\vec{\mu}} \ln P(\vec{x}|\vec{\mu}) = \partial_{\mu}\mathcal{L} =$$

$$= \bar{\Sigma}^{-1} \cdot (\vec{x}-\vec{\mu}) \longrightarrow \vec{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

LATENT VARIABLE MODELS $\rightarrow$ hidden variable

GMM: $p(\vec{x}) = \sum_{k=1}^{G} \pi_k \cdot N(\vec{x} \mid \vec{\mu}_k, \Sigma_k)$

LATENT VARIABLE (clustering variable): $\vec{z} = (z_1, z_2, \dots z_G)$ $\quad z_j \in \{0,1\}$

one-hot encoding

$\vec{x}$ belongs to cluster 3 $\rightarrow$ $\vec{z} = (0, 0, 1, 0, 0 \dots, 0)$

$p(z_k = 1) = \pi_k$

joint distribution $p(\vec{x}, \vec{z}) = p(\vec{x}|\vec{z})\, p(\vec{z})$

latent

conditional probability $p(\vec{z}|\vec{x}) \rightarrow$ posterior

posterior likelihood prior

$$p(\vec{z}|\vec{x}) = \frac{p(\vec{x}|\vec{z})\, p(\vec{z})}{p(\vec{x})}$$

$p(\vec{x}) \rightarrow$ normaliz. constant

EXPECTATION - MAXIMIZATION ALGORITHM (EM-algorithm)

Iterative procedure to get the ML estimate of parameters in probabilistic models with latent variables.

$\boxed{\text{GMM + EM}}$ → start from initial value for the parameters

$$\underline{\underline{\theta}}^{0} = \{ \pi_j^0 , \vec{\mu}_j^0 , \Sigma_j^0 \}$$

① E-STEP: Evaluate posterior $P(\vec{z} \mid \vec{x}, \underline{\underline{\theta}}^0)$

② M-STEP: $\theta_{new} = \underset{\theta}{\arg\max}\ Q(\theta, \theta_0)$

→ θ^*

ML estimate of parameters

expect. $Q(\theta, \theta_0) = \sum_{\vec{z}} \underline{P(\vec{z} \mid \vec{x}, \theta^0)} \cdot \ln P(\vec{x}, \vec{z} \mid \theta)$

fitting a GMM :

set of observations $\{\vec{x}_1, \vec{x}_2, \cdots \vec{x}_n\}$

$\downarrow$

EM-algorithm

$\downarrow$

$\pi_j^*, \vec{\mu}_j^*, \Sigma_j^* \quad j = 1 \cdots G \qquad \vec{z}$

$\downarrow$

$$p(\vec{x}) = \sum_{j=1}^{G} \pi_j^* \, N(\vec{x} \mid \vec{\mu}_j, \Sigma_j) \longrightarrow$$

assign each obs. to a cluster.

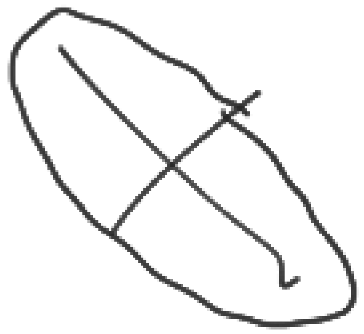
GMM - options

→ Randomized initial values for θ

→ Choice for the structure of the covariance matrix:

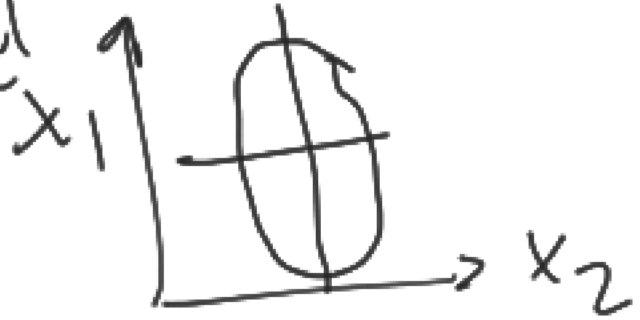
$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

full-covariance



$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

diagonal



tied

$\Sigma_1 \quad \Sigma_2$

Σ

$$\Sigma = \begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}$$