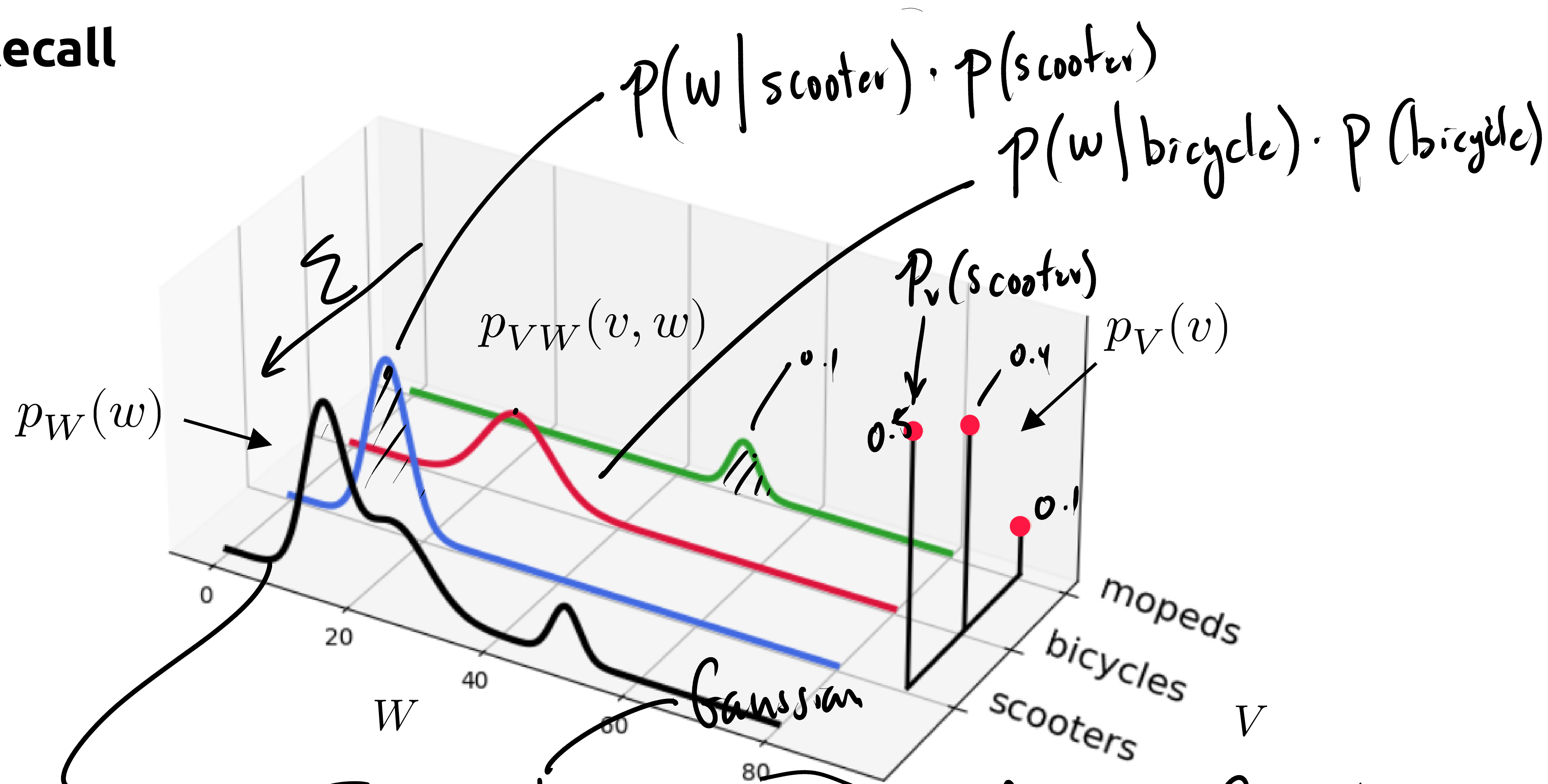
Statistics and Data Science for Engineers
E178 / ME276DS

Gaussian mixtures and K-means

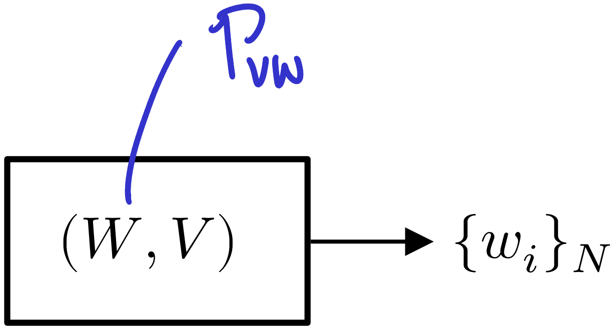
Recall



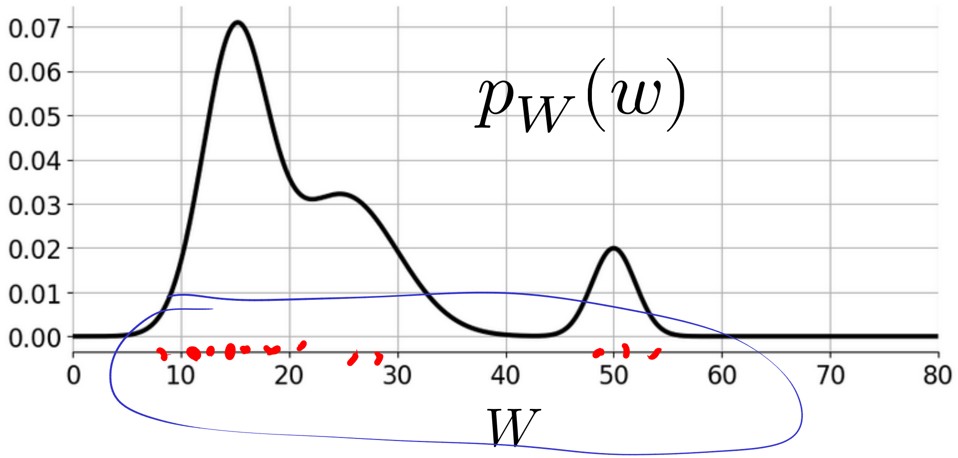
$$p_W(w) = \sum_v p(w|V=v) \cdot p_v(v)$$

fraction

... Gaussian mixture



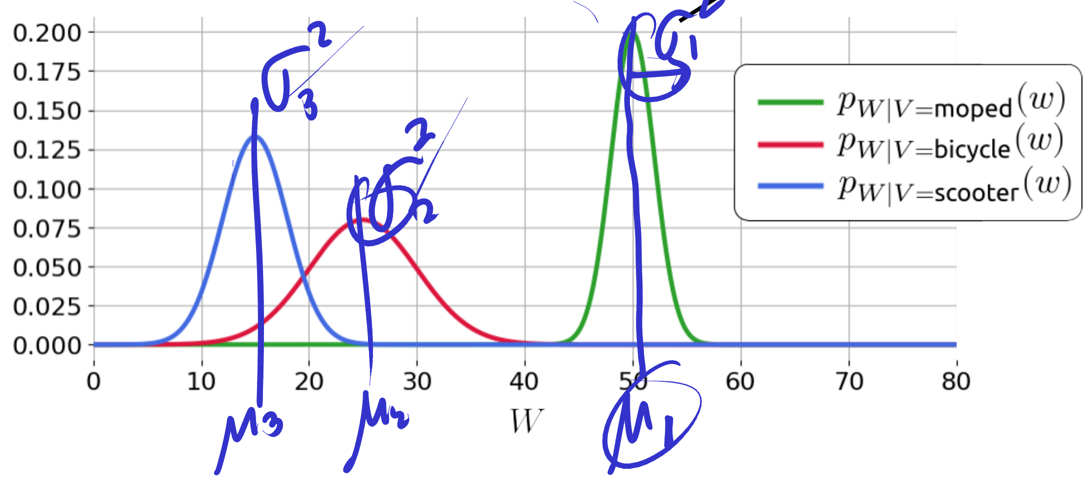
Problem: Can we estimate the P_{VW} from measurements of weight only.

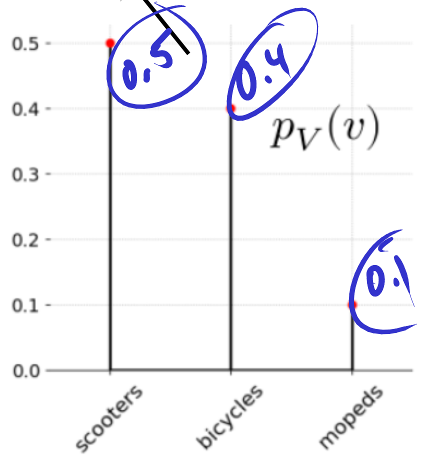


... Gaussian mixture.

Gaussian fraction

$$
\begin{aligned}
p_{VW}(\text{scooter}, w) &= p(w \mid \text{scooter})\, p_V(\text{scooter}) \\
p_{VW}(\text{bicycle}, w) &= p(w \mid \text{bicycle})\, p_V(\text{bicycle}) \\
p_{VW}(\text{moped}, w) &= p(w \mid \text{moped})\, p_V(\text{moped})
\end{aligned}
$$





Assumption: The class-conditioned weights are Gaussian.

Normal distrib

$$W \mid V = v \sim \mathcal{N}(\mu_v, \sigma_v^2) \qquad \forall v \in \{\text{scooter}, \text{bicycle}, \text{moped}\}$$

mean variance

More generally:

$$\boxed{(Y,Z)} \longrightarrow \{y_i\}_N$$

- w • Observations: Y $\quad \Omega_Y = \mathbb{R}$
- v • Hidden variable: Z $\quad \Omega_Z = \{1 \ldots K\}$
- Marginal distribution of Z: $\quad \pi_k = p_Z(k)$

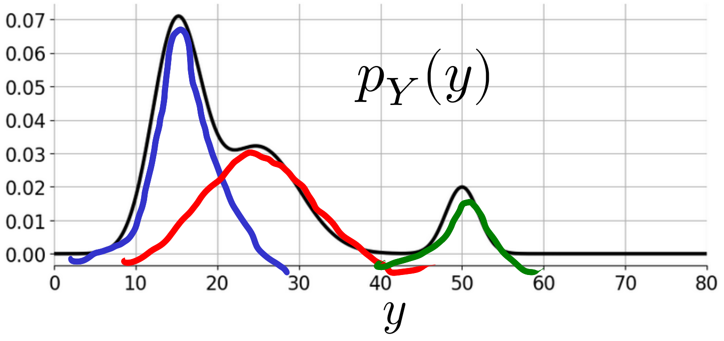
$\{0.5, 0.4, 0.1\}$

$$\sum_{k=1}^{K} \pi_k = 1 \quad , \quad \pi_k \geq 0$$

- Class conditioned observations are Gaussian: $\quad w \mid V = v = \mathcal{N}(\mu, \sigma^2)$.

$$\mathcal{N}_k(y_i) = \quad p(y \mid Z = k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu_k)^2}{\sigma_k^2}\right) = \mathcal{N}_k(y)$$

- Observations are a Gaussian mixture:



$$p_Y(y) = \int_{\Omega_Z} p_{YZ}(y, z)dz$$

$$= \sum_{k=1}^{K} p_{YZ}(y, k)$$

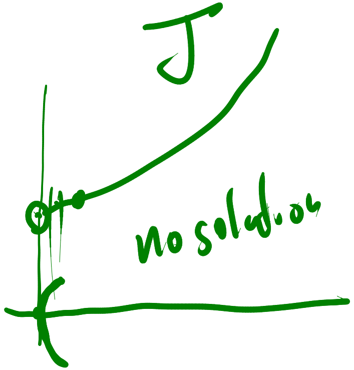
$$= \sum_{k=1}^{K} \underbrace{p(y|k)}_{\mathcal{N}_k} \underbrace{p(k)}_{\pi_k} \quad \ldots \text{ Gaussian mixture.}$$

MLE for Gaussian mixtures

$3K-1$

- $\mathcal{D} = \{y_i\}_N$
- $\underline{\theta} = \{(\pi_k, \mu_k, \sigma_k^2)\}_K$ ... $3K$ parameters to estimate.
- Log-likelihood: $\ln \mathcal{L}(\underline{\theta}\,; \mathcal{D}) = \sum_{i=1}^{N} \ln p_Y(y_i\,;\, \underline{\theta})$

- Optimization problem:

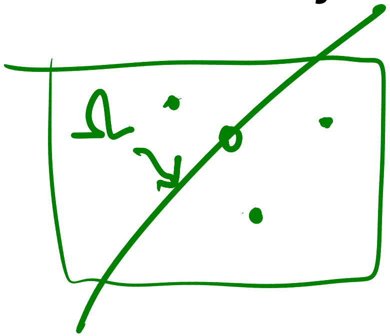


$$\underset{\{(\pi_k,\mu_k,\sigma_k^2)\}_K}{\text{maximize}} \quad \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \overbrace{\mathcal{N}_k(y_i)}^{\text{Gaussian pdf with parameters } \mu_k,\ \sigma_k^2}\right)$$

$$\text{subject to} \quad \sum_{k=1}^{K} \pi_k - 1 = 0$$

$$\left.\begin{aligned} &\pi_k \geq 0 \qquad k \in \{1 \ldots K\} \\ &\sigma_k^2 > 0 \qquad k \in \{1 \ldots K\} \end{aligned}\right\} \text{boundary points.}$$

could be no solution.



Note:

1. Fixing K
2. Objective function is non-convex ∴ may end up with local (not global) solution.
3. Equality constraint

- Append the equality constraint to the objective function:

$\pi_k \ldots$ class fractions.

$$
\begin{aligned}
\underset{\lambda,\underline{\theta}}{\text{maximize}} \quad & \overbrace{\sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k\, \mathcal{N}_k(y_i)\right) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)}^{J} \\
\text{subject to} \quad & \pi_k \geq 0 \qquad k \in \{1 \ldots K\} \\
& \sigma_k^2 > 0 \qquad k \in \{1 \ldots K\}
\end{aligned}
$$

Ω.

- Equality constraint $\longrightarrow$ Lagrange multiplier. $\lambda \ldots$ variable

Derivative with respect to μ_r

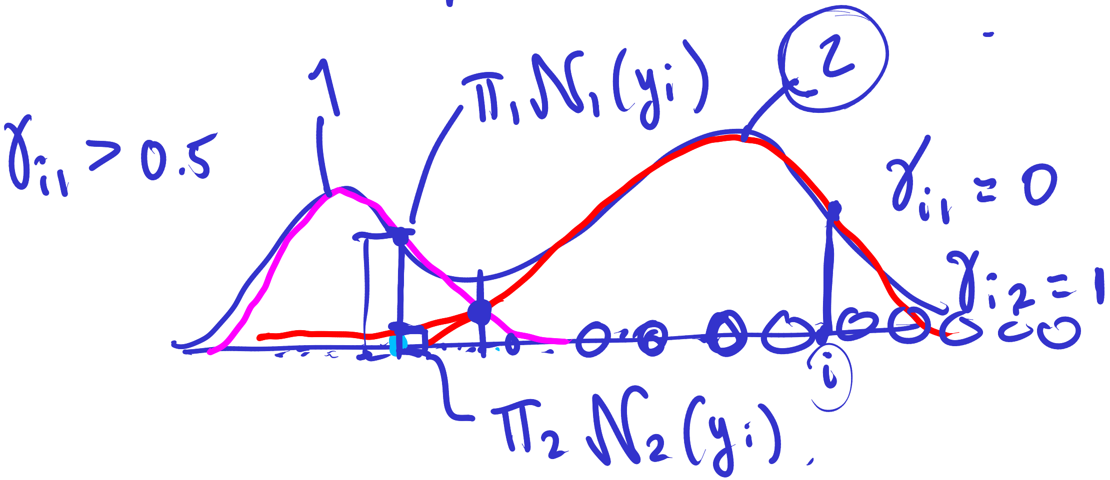
$r \in 0 \ldots K$

$$\frac{\partial J}{\partial \mu_r} = \quad \ldots \quad = \sum_{i=1}^{N} \underbrace{\frac{\pi_r \, \mathcal{N}_r(y_i)}{\sum_{k=1}^{K} \pi_k \, \mathcal{N}_k(y_i)}}_{\gamma_{ir}} \frac{(y_i - \mu_r)}{\sigma_r^2} = 0$$

$$\gamma_{ir} = \frac{\pi_r \, \mathcal{N}_r(y_i)}{\sum_{k=1}^{K} \pi_k \, \mathcal{N}_k(y_i)}$$

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

$$\gamma_{ik} \geq 0.$$



γ_{ik} ... responsibility of class k for data point i

$$\sum_{i=1}^{N} \gamma_{ik} \frac{y_i - \mu_k}{\cancel{\sigma_k^2}} = 0$$ ... stationarity condition.

$$\therefore \quad \sum_i \gamma_{ik} y_i - \mu_k \overbrace{\sum_i \gamma_{ik}}^{N_k} = 0.$$

N_k ... total responsibility for class k

$$\boxed{\mu_k = \frac{1}{N_k} \sum_i \gamma_{ik} y_i}$$ ... responsibility weighted mean.

Derivative with respect to σ_r^2

$$\frac{\partial J}{\partial \sigma_r^2} = \quad \cdots \quad = \sum_{i=1}^{N} \gamma_{ir} \left(\frac{(y_i - \mu_r)^2}{\sigma_r^2} - 1 \right) = 0$$

$$\sum_{i=1}^{N} \gamma_{ir} \frac{(y_i - \mu_r)^2}{\sigma_r^2} - \underbrace{\sum_i \gamma_{ir}}_{N_r} = 0$$

$$\boxed{\sigma_k^2 = \frac{1}{N_k} \sum \gamma_{ik} (y_i - \mu_k)^2}$$

responsibility-weighted sample variance

Derivative with respect to π_r

$$\frac{\partial J}{\partial \pi_r} = \quad \dots \quad = \sum_{i=1}^{N} \frac{\mathcal{N}_r(y_i)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}_j(y_i)} + \lambda \; = \; 0$$

$\vdots$

$$\boxed{\pi_k = \frac{N_k}{N}}$$

$$\sum_{k=1}^{K} N_k = N$$

$$\sum \pi_k = 1 \quad \checkmark$$

Expectation-Maximization (EM) algorithm

Random initialization of $\{(\mu_k, \sigma_k^2, \pi_k)\}_K$

E

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}_k(y_i)}{\sum_{j=1}^{K} \pi_j \mathcal{N}_j(y_i)}$$

$$N_k = \sum_{i=1}^{N} \gamma_{ik}$$

M

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} y_i$$

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik}(y_i - \mu_k)^2$$

$$\pi_k = \frac{N_k}{N}$$

In 2D:

$\mu_k = \begin{bmatrix} \cdot \\ \cdot \end{bmatrix}$

$\sigma_k^2 = \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix}$

From GMM to K-means clustering

Assumptions:

1. Gaussians are "circles" : $\sigma_k^2 = \begin{bmatrix} \varepsilon & & 0 \\ & \varepsilon & \\ 0 & & \varepsilon & \\ & & & \ddots \end{bmatrix} = \varepsilon I$

2. $\varepsilon \longrightarrow 0$. $\Longrightarrow$ $\gamma_{ik} = \begin{cases} 1 & \text{if } \mu_k \text{ is closest to } y_i \\ 0 & \text{otherwise.} \end{cases}$

K-means clustering

Random initialization of $\{\mu_k\}_K$

E

$\gamma_{ik} =$ Assign each point to the nearest centroid

μ_k

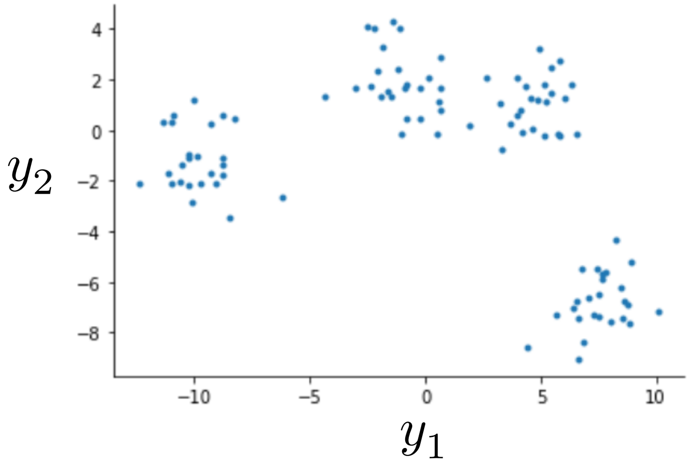
$N_k =$ Number of points assigned to this cluster

M

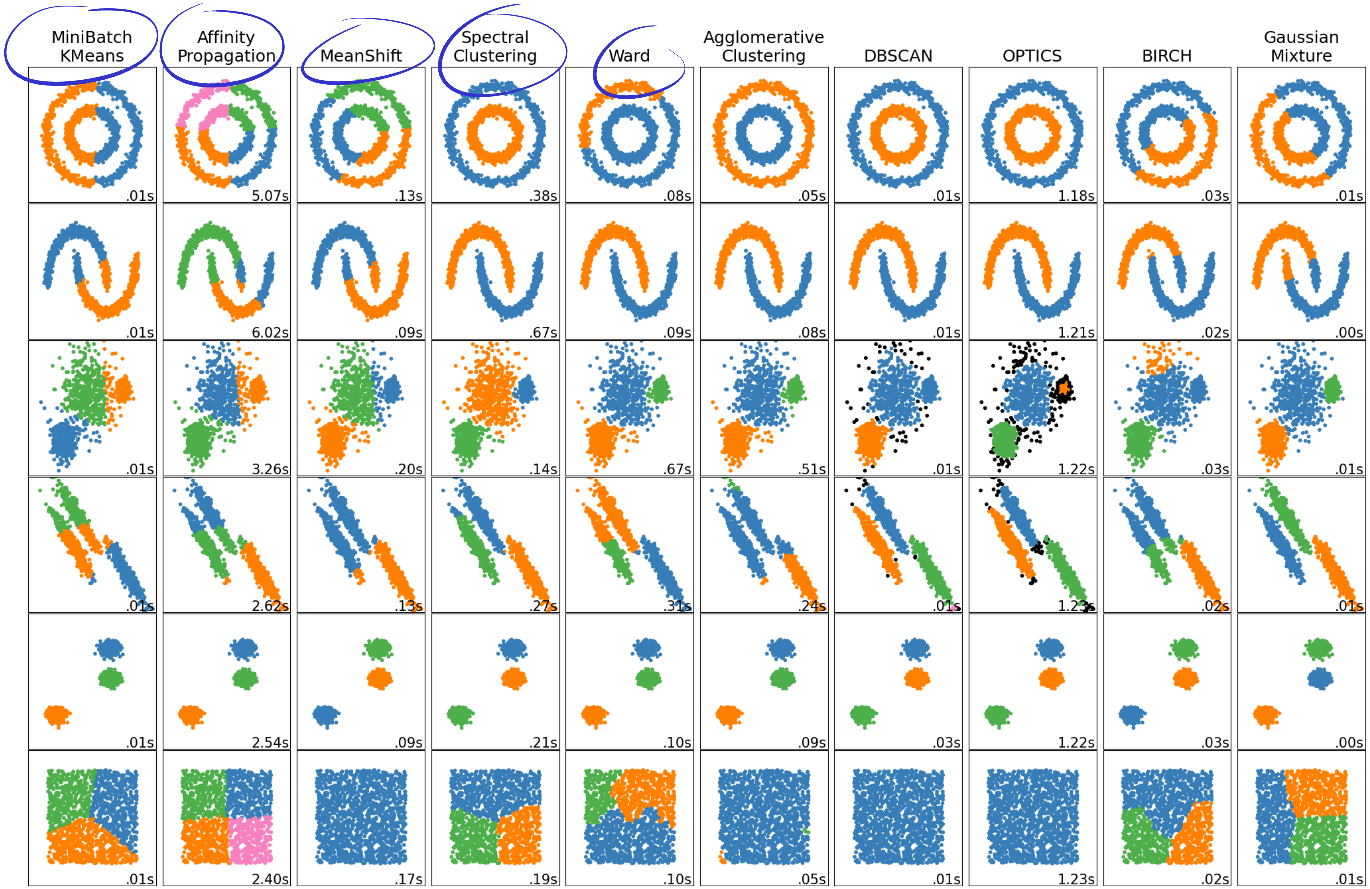
$\mu_k =$ Place centroid at the COM of the cluster

General clustering

- Given $\{y_i\}_N$ with $y_i \in \mathbb{R}^D$
- Goal: Group the data into *clusters*
- What is a cluster?

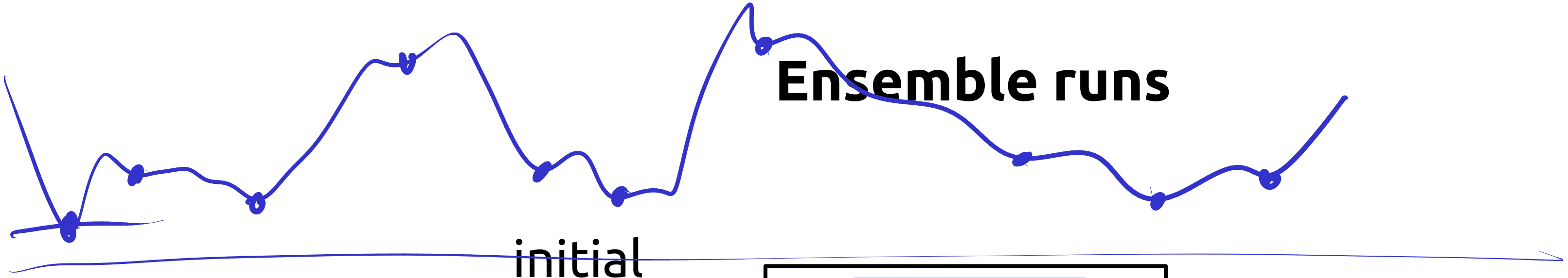


A few clustering algorithms (from scikit-learn)

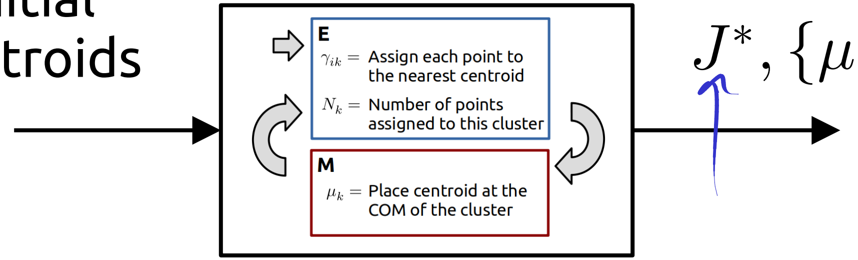


https://scikit-learn.org/stable/modules/clustering.html#clustering

Ensemble runs



initial centroids



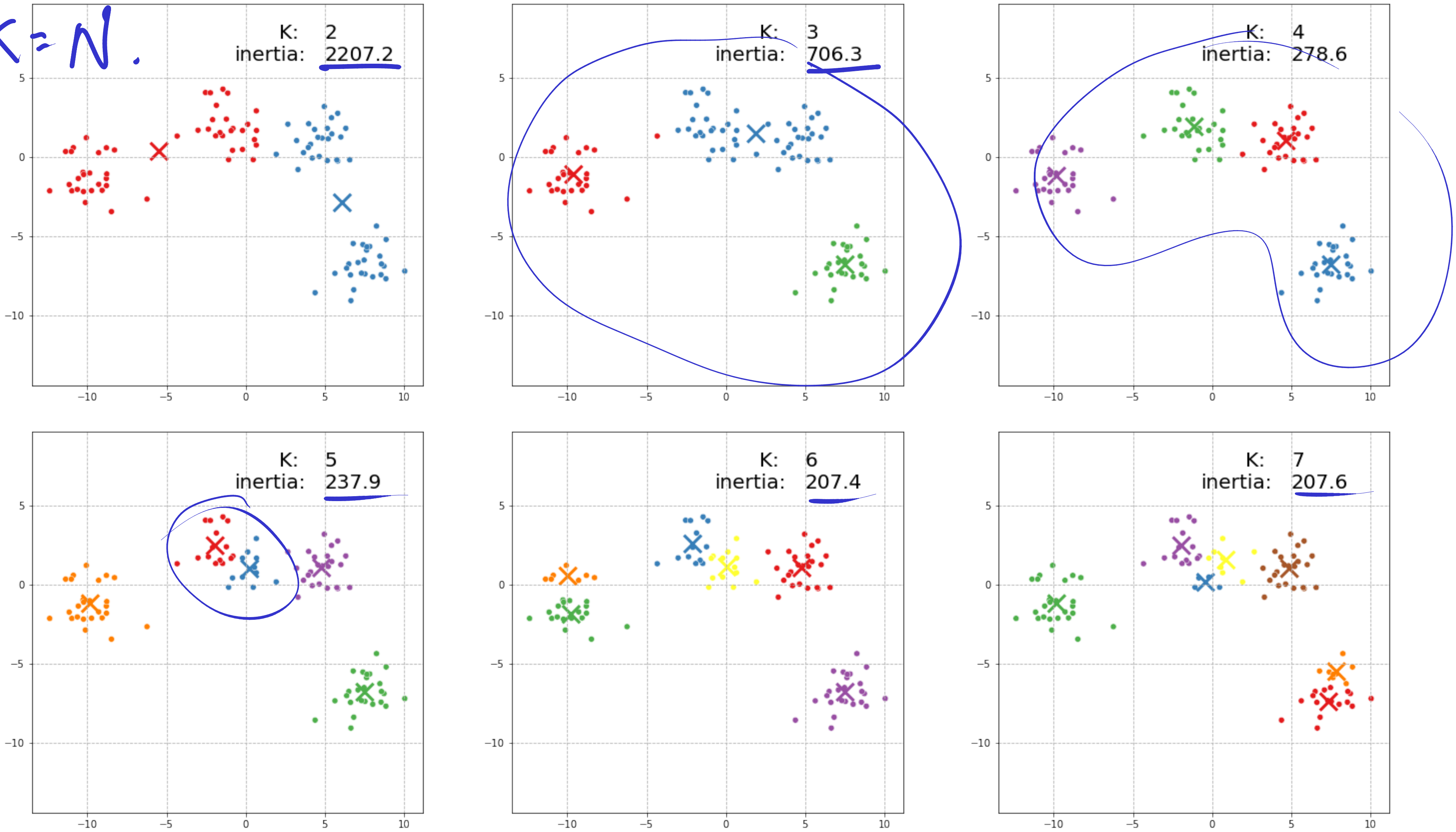
$J^*, \{\mu_k^*\}_K, \{\gamma_{ik}^*\}_K$

- Minimize J^* with respect to initial conditions (K fixed).
- Repeat for many values of K

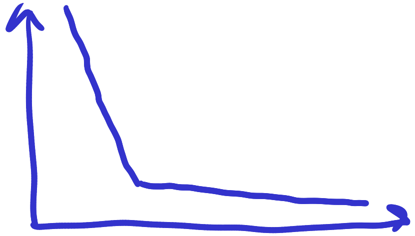
$$J = \sum_{k} \sum_{\substack{\text{data } i \\ \text{my cluster}}} (y_i - \mu_k)^2$$

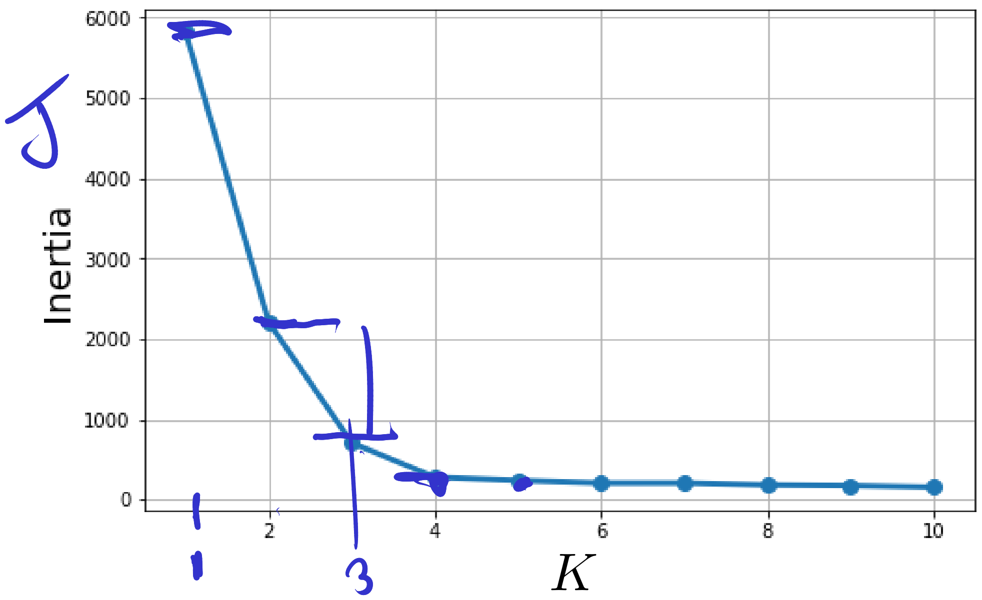
"inertia"

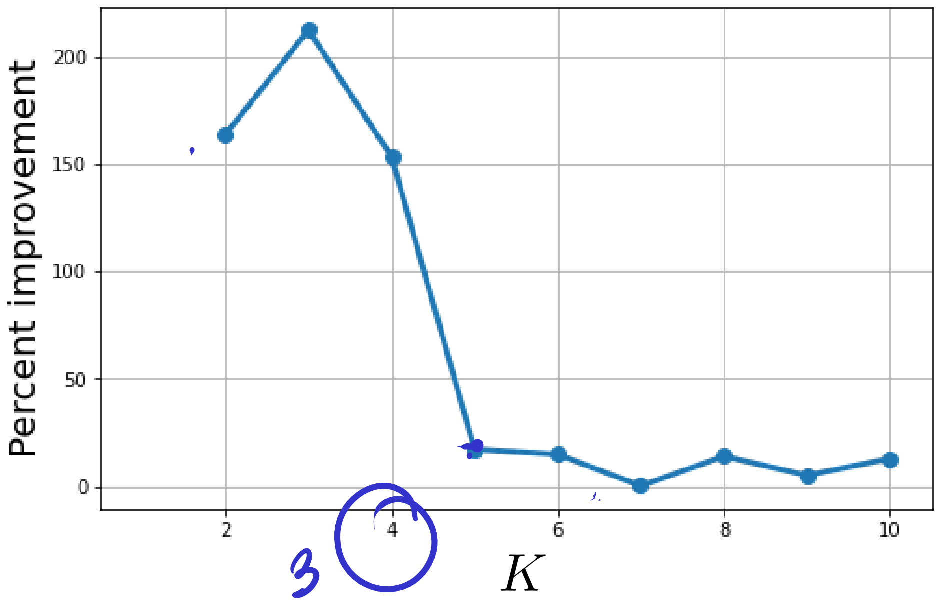
K=N.


K: 2
inertia: 2207.2
K: 3
inertia: 706.3
K: 4
inertia: 278.6
K: 5
inertia: 237.9
K: 6
inertia: 207.4
K: 7
inertia: 207.6

Selecting K: The elbow method







$$\frac{4000}{6000}$$