OVERVIEW

- semi-unsupervised learning - graphs.
  ↳ spectral clustering
- Clustering:
  - k-means
  - GMM
  - spectral clustering
  - evaluation
- Matrix Factorisation
  - SVD
  - low-rank approximation
  - PCA
  - Robust PCA

ROBUST PCA

Q:

$X \in \mathbb{R}^{m \times n}$, $D_\tau(X)$ - singular value threshold

Prove: (a) $\|D_\tau(X)\|_{*,F} \leq \|X\|_{*,F}$

(b) $\text{rank}(D_\tau(X)) \leq \text{rank}(X)$

A:

$$X = U \cdot \Sigma \cdot V^T$$

$$D_\tau(X) = U \cdot S_\tau(\Sigma) \cdot V^T$$

$$S_\tau(x) = \begin{cases} x - \tau & x > \tau \\ 0 & |x| \leq \tau \\ x + \tau & x < -\tau \end{cases}$$

$$\|X\|_* = \sum_{i=1}^{r} \sigma_i$$ ← singular values

(a) $\sigma_1, \_\_\_, \sigma_r$ – are singular values of X. > 0

$\Rightarrow S_\tau(\sigma_1), \_\_\_, S_\tau(\sigma_r)$ – " " " of $D_\tau(X)$

$S_\tau(\sigma_i) \leq \sigma_i \quad (\sigma_i \geq 0)$

$\hookrightarrow \begin{pmatrix} \sigma_i - \tau & \sigma_i > \tau \\ 0 & \sigma_i < \tau \end{pmatrix} \leq \sigma_i$

$\Rightarrow \|D_\tau(X)\|_* = \sum_{i=1}^{r} S_\tau(\sigma_i) \leq \sum_{i=1}^{r} \sigma_i = \|X\|_*$

(b) If $\sigma_i > 0 \Rightarrow S_\tau(\sigma_i) > 0$
$= 0$

If $\sigma_i = 0 \Rightarrow S_\tau(\sigma_i) = 0$.

$\Rightarrow$ total number of non-zero singular values can only go down

$\Rightarrow \underbrace{\text{rank}(D_\tau(X))}_{\#\text{ non-zero sing. values}} \leq \text{rank}(X)$

Q: When is: $\|D_\tau(X)\|_* = \|X\|_*$?

$\text{rank}(D_\tau(X)) = \text{rank}(X)$?

$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$

$S_\tau(\sigma_1) \geq S_\tau(\sigma_2) \geq \cdots \geq S_\tau(\sigma_r) \overset{?}{>} 0$

$\text{rank}(D_\tau(X)) = \text{rank}(X)$ iff $S_\tau(\sigma_r) > 0$

$\Updownarrow$

$\sigma_r > \tau$.

$\Updownarrow$

$\sigma_1, \_\_\_, \sigma_r > \tau$

$$S_\tau(\sigma_i) \overset{?}{=} \sigma_i \qquad (\tau > 0)$$

$\checkmark\ 0$ $\qquad$ $\sigma_i - \tau$ X

$$0 = \sigma_i \ \rightarrow \text{ for all } i = 1, \dots, r.$$

$$\Downarrow$$

$$X = 0$$

$$E = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 71 & 0 \\ 87 & 0 & 0 & 0 \\ 0 & 0 & 78 & 65 \end{pmatrix} \qquad \|E\|_0 = 4 \ \checkmark \longleftarrow$$

2:

$$X = \begin{pmatrix} 1 & -2 & 0 & 1 \\ 2 & 3 & 1 & 0 \\ 3 & 1 & 72 & 1 \\ 85 & 4 & 0 & -2 \\ -1 & -5 & 77 & 66 \end{pmatrix}$$

Find $X = L + E$ s.t. $\quad \text{rank}(L) \leq 2.$

$\|E\|_0 \leq 5.$ (25%)

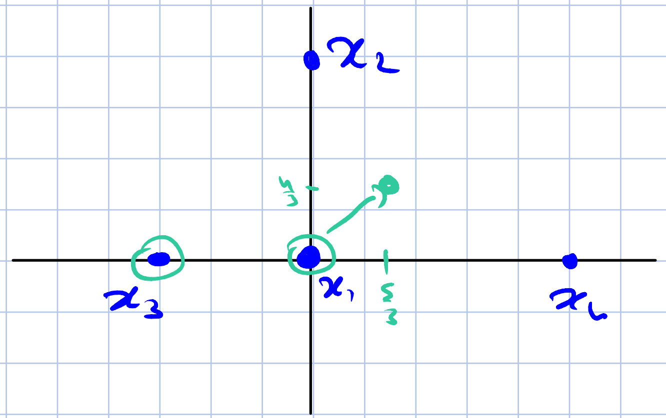
Take: $L = \begin{pmatrix} 1 & -2 & 0 & 1 \\ 2 & 3 & 1 & 0 \\ 3 & 1 & 1 & 1 \\ -2 & 4 & 0 & -2 \\ -1 & -5 & -1 & 1 \end{pmatrix}$ $\begin{matrix} \\ \\ \rightarrow R_3 = R_1 + R_2 \\ \rightarrow R_4 = 2 \cdot R_1 \\ \rightarrow R_5 = R_1 - R_2 \end{matrix}$

rank(L) = 2 ✓.

k-means

$x_1 = (0,0)$ $\quad x_2 = (0,4)$, $\quad x_3 = (-3,0)$

$x_4 = (5,0)$



run k-means with $\underline{k=2}$

$\mu_1^{(0)} = (0,0)$, $\mu_2^{(0)} = (-3,0)$

Solution: Step 1: $C_1 = \{x_1, x_2, x_4\}$

$C_2 = \{x_3\}$

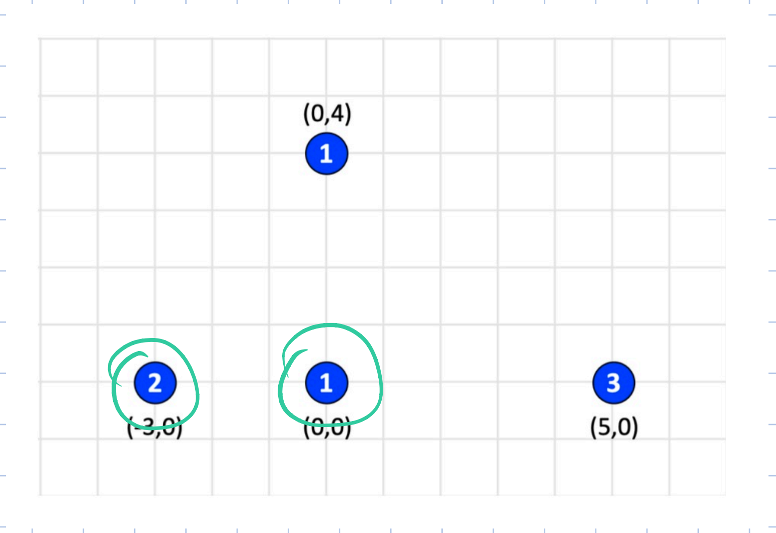
$\mu_1^{(1)} = \left(\frac{5}{3}, \frac{4}{3}\right)$, $\mu_2^{(1)} = (-3,0)$

Step 2: $C_1 = \{x_1, x_2, x_4\}$

$C_2 = \{x_3\}$

no chanes

Stop



$x_1 = (0,0)$, $x_2 = (0,4)$ $\quad x_3 = x_4 = (-3,0)$

$x_5 = x_6 = x_7 = (5,0)$

$\mu_1^{(0)} = (0,0)$, $\mu_2^{(0)} = (-3,0)$

Solution:

Step 1:

$$C_1 = \{x_1, x_2, x_5, x_6, x_7\}$$

$$C_2 = \{x_3, x_4\}.$$

$$\mu_1^{(1)} = \left(3, \tfrac{4}{5}\right), \quad \mu_2^{(1)} = (-3, 0)$$

Step 2:

$$C_1 = \{x_2, x_5, x_6, x_7\}$$

$$C_2 = \{x_1, x_3, x_4\}.$$

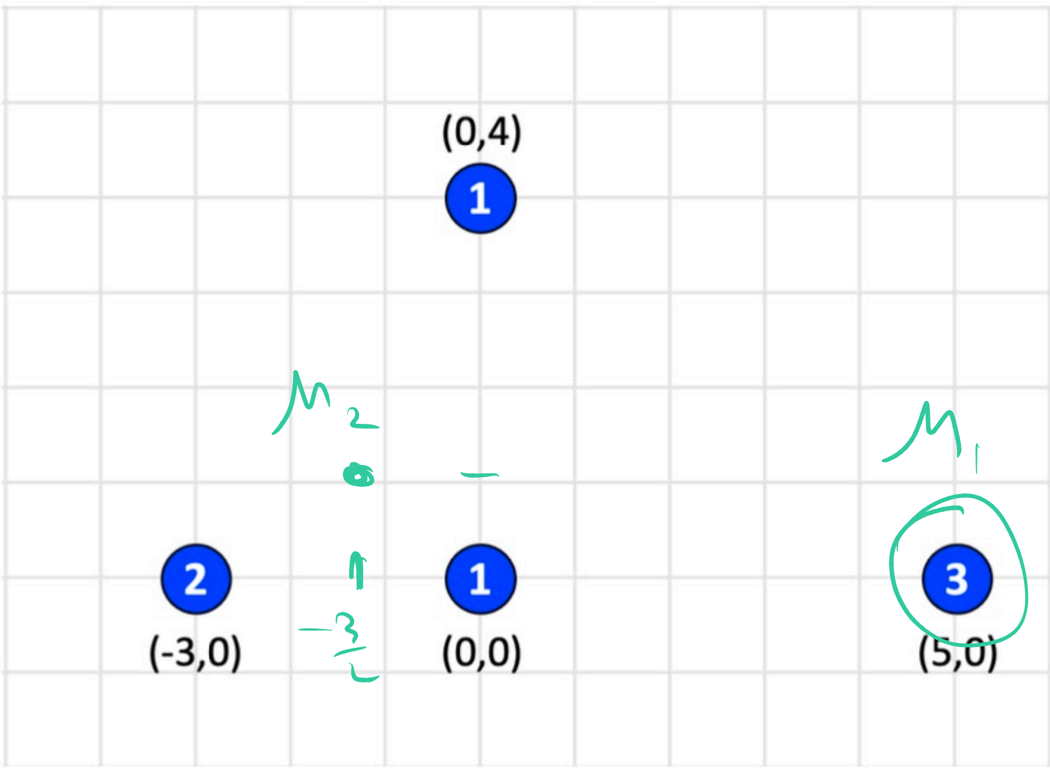
$$\mu_1^{(2)} = \left(\tfrac{15}{4}, 1\right) \quad \mu_2^{(2)} = (-2, 0)$$

Step 3:

$$C_1 = \{x_5, x_6, x_7\}$$

$$C_2 = \{x_1, x_2, x_3, x_4\}$$

$$\mu_1^{(3)} = (5, 0), \quad \mu_2^{(3)} = \left(-\tfrac{3}{2}, 1\right)$$



Step 4:

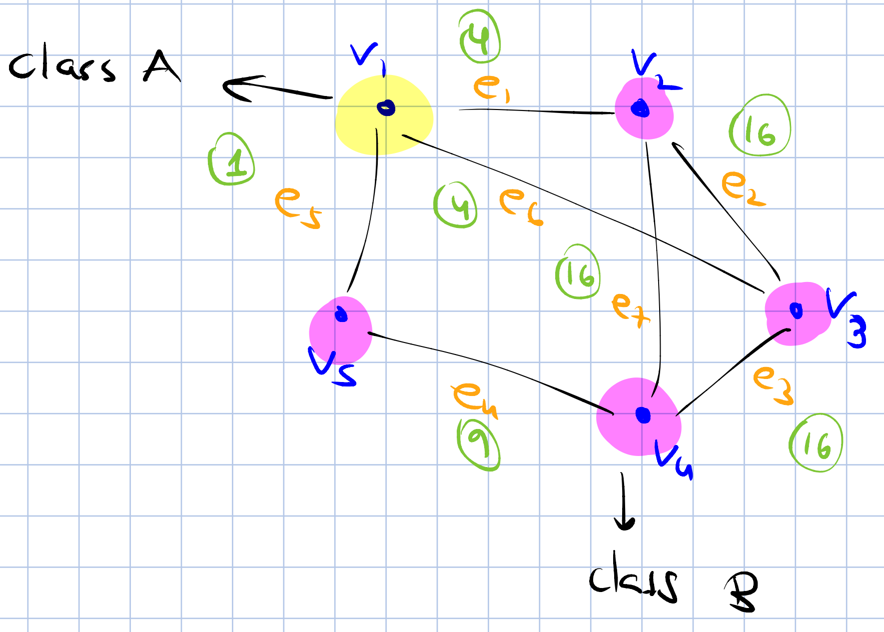
$$C_1 = \{x_5, x_6, x_7\}$$

$$C_2 = \{x_1, x_2, x_3, x_4\}$$

$\downarrow$

no change.

Semi-supervised Learning



Labelled: $V_L = \{1, 4\}$

$V_U = \{2, 3, 5\}$

$$L = \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 9 & -4 & -4 & 0 & -1 \\ -4 & 36 & -16 & -16 & 0 \\ -4 & -16 & 36 & -16 & 0 \\ 0 & -16 & -16 & 41 & -9 \\ -1 & 0 & 0 & -9 & 10 \end{pmatrix} \end{array} \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}$$

$$P_L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad P_U = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Known labels: $y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

Solve: $Ag = b$

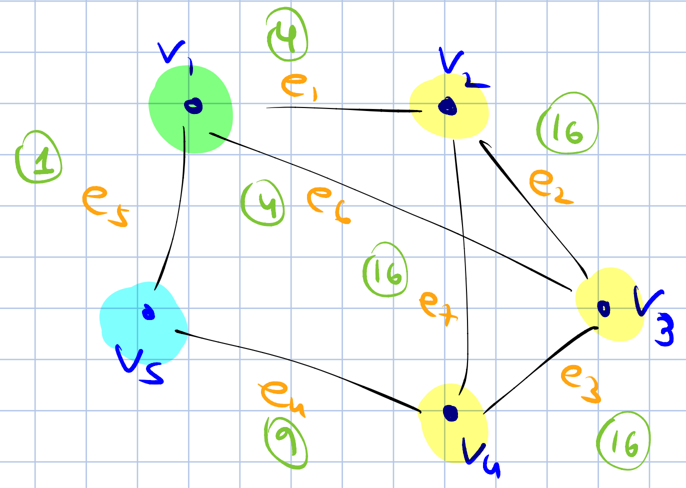
$$A = P_U L P_U^T = \begin{pmatrix} 36 & -16 & 0 \\ -16 & 36 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

$$b = -(P_U L P_L^T) \cdot y = \begin{pmatrix} -4 & -16 \\ -4 & -16 \\ -1 & -9 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -16 \\ -16 \\ -9 \end{pmatrix}$$

$$\Downarrow$$

$$g^* = \begin{pmatrix} 0.8 \\ 0.8 \\ 0.9 \end{pmatrix} \rightarrow f^* = \begin{pmatrix} 0 \\ 0.8 \\ 0.8 \\ 1 \\ 0.9 \end{pmatrix} \Rightarrow \text{Labels: } \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Suppose - no labels → Spectral clustering:



$$L = \begin{pmatrix} 9 & -4 & -4 & 0 & -1 \\ -4 & 36 & -16 & -16 & 0 \\ -4 & -16 & 36 & -16 & 0 \\ 0 & -16 & -16 & 41 & -9 \\ -1 & 0 & 0 & -9 & 10 \end{pmatrix}$$

$\lambda_1 = 0$

$\lambda_2 \geq \lambda_1$

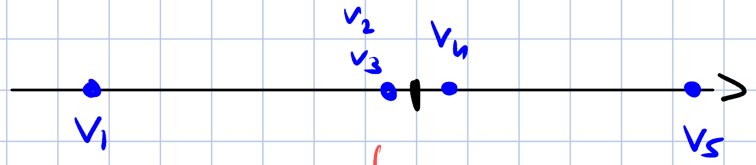
$\Downarrow$

$\lambda_2 = 9.276$

$$u_2 = \begin{array}{c} \begin{pmatrix} -0.7 \\ -0.06 \\ -0.06 \\ 0.44 \\ 0.69 \end{pmatrix} \end{array} \begin{array}{c} - \\ - \\ - \\ + \\ + \end{array}$$

$\downarrow$

$$\text{labels: } \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$



↓

3 clusters!

Spectral clustering for k=3:

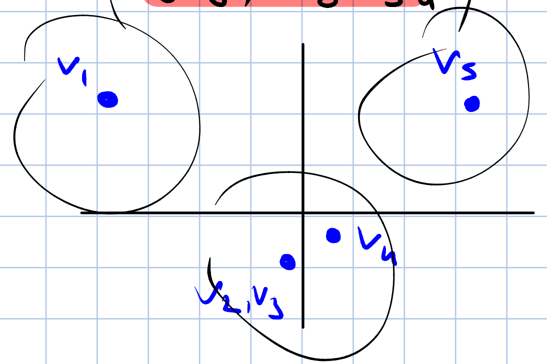
$\lambda_3 = 13.83$

$$u_3 = \begin{pmatrix} 0.55 \\ -0.4 \\ -0.4 \\ -0.29 \\ 0.54 \end{pmatrix} \rightarrow U = \begin{array}{cc} u_2 & u_3 \\ \hline -0.7 & 0.55 \\ -0.06 & -0.4 \\ -0.06 & -0.4 \\ 0.44 & -0.29 \\ 0.69 & 0.54 \end{array}$$

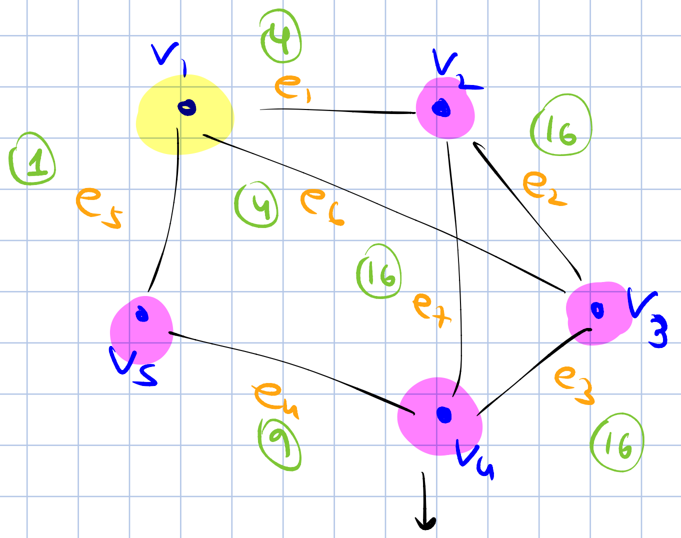
$P(v_1) = (-0.7, 0.55)$

$P(v_2) = (-0.06, -0.4)$

$\vdots$



Evaluation of clustering



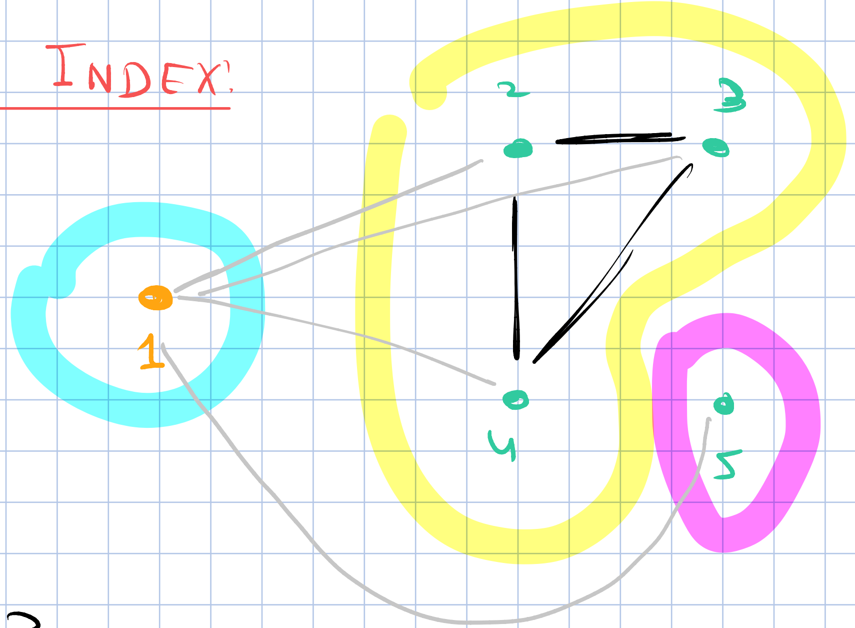
External evaluation:

Our clusters: $C_1 = \{1\}$, $C_2 = \{2, 3, 4, 5\}$

Ground truth: $C_1^* = \{1\}$, $C_2^* = \{2, 3, 4\}$

$C_3^* = \{5\}$

RAND INDEX:



$TP = 3$

$TN = 4$

$FP = 3$

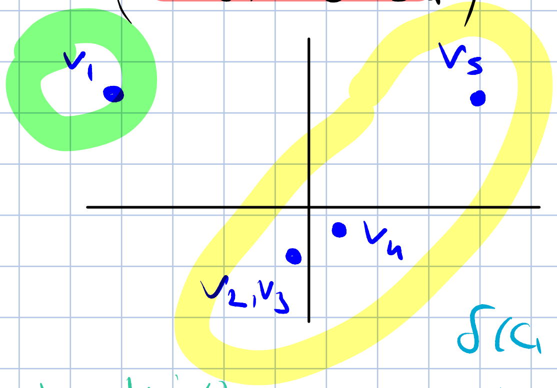
$FN = 0$

$$\rightarrow RI = \frac{TP + TN}{\underbrace{TP + TN + FP + FN}_{\binom{n}{2} \rightarrow \binom{5}{2} = 10}} = 0.7$$

Internal evaluation:

⊗ need distances → use spectral representation

$$
\begin{array}{cc} & \begin{array}{cc} u_2 & u_3 \end{array} \\ U = & \begin{pmatrix} -0.7 & 0.55 \\ -0.06 & -0.4 \\ -0.06 & -0.4 \\ 0.44 & -0.29 \\ 0.69 & 0.54 \end{pmatrix} \end{array}
$$



$\delta(c_1, c_2) = \min(\cdots)$

Distance matrix:

0	1.1452	1.1452	1.1863	1.3959
1.1452	0	0.0000	0.2241	1.2070
1.1452	0.0000	0	0.2241	1.2070
1.1863	0.2241	0.2241	0	1.0034
1.3959	1.2070	1.2070	1.0034	0

→ $\Delta(c_2) = \max(\cdots)$

DUNN INDEX:

Inter-cluster distance: $\delta(c_1, c_2) = 1.1452$

Intra-cluster distance: $\Delta(c_1) = 0$

$\Delta(c_2) = 1.207$

$$DI = \frac{\min_{i,j} \delta(c_i, c_j)}{\max_i \Delta(c_i)} = \frac{1.1452}{1.207} = 0.948$$

single-linkage

diameter