

Lecture 05b

Distance Measures

Similarity and Dissimilarity

- Similarity
 - Numerical measure of how alike two data objects are.
 - Is higher when objects are more alike.
 - Often falls in the range $[0,1]$
- Dissimilarity
 - Numerical measure of how different are two data objects
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Euclidean Distance

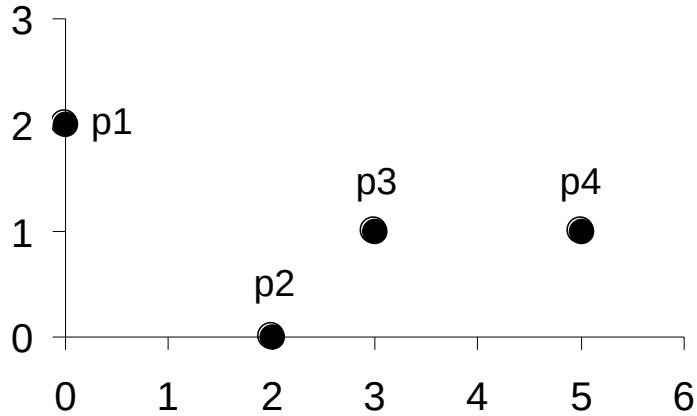
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q .

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .

Minkowski Distance: Examples

- $r = 1$. Cityblock (Manhattan, taxicab, L_1 norm) distance.
 - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean (L_2 norm) distance.
- $r \rightarrow \infty$. Chebyshev (L_{\max} norm, L_{∞} norm, maximum, supremum) distance.
 - This is the maximum difference between any component of the vectors
 - Example: L_{∞} of $(1, 0, 2)$ and $(6, 0, 3) = ??$
 - Do not confuse r with n , i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

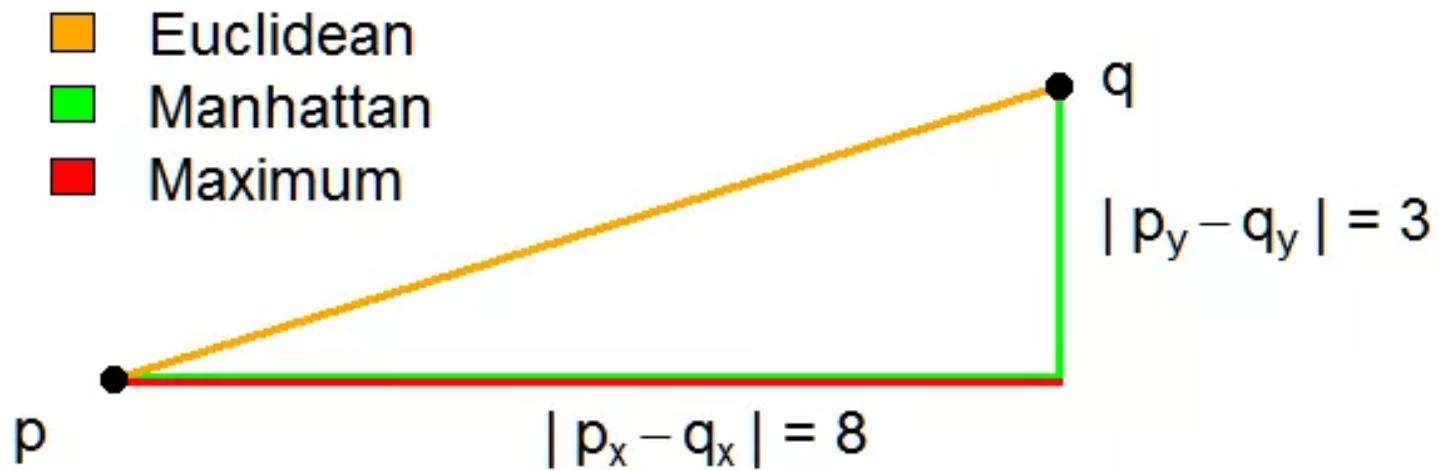
L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

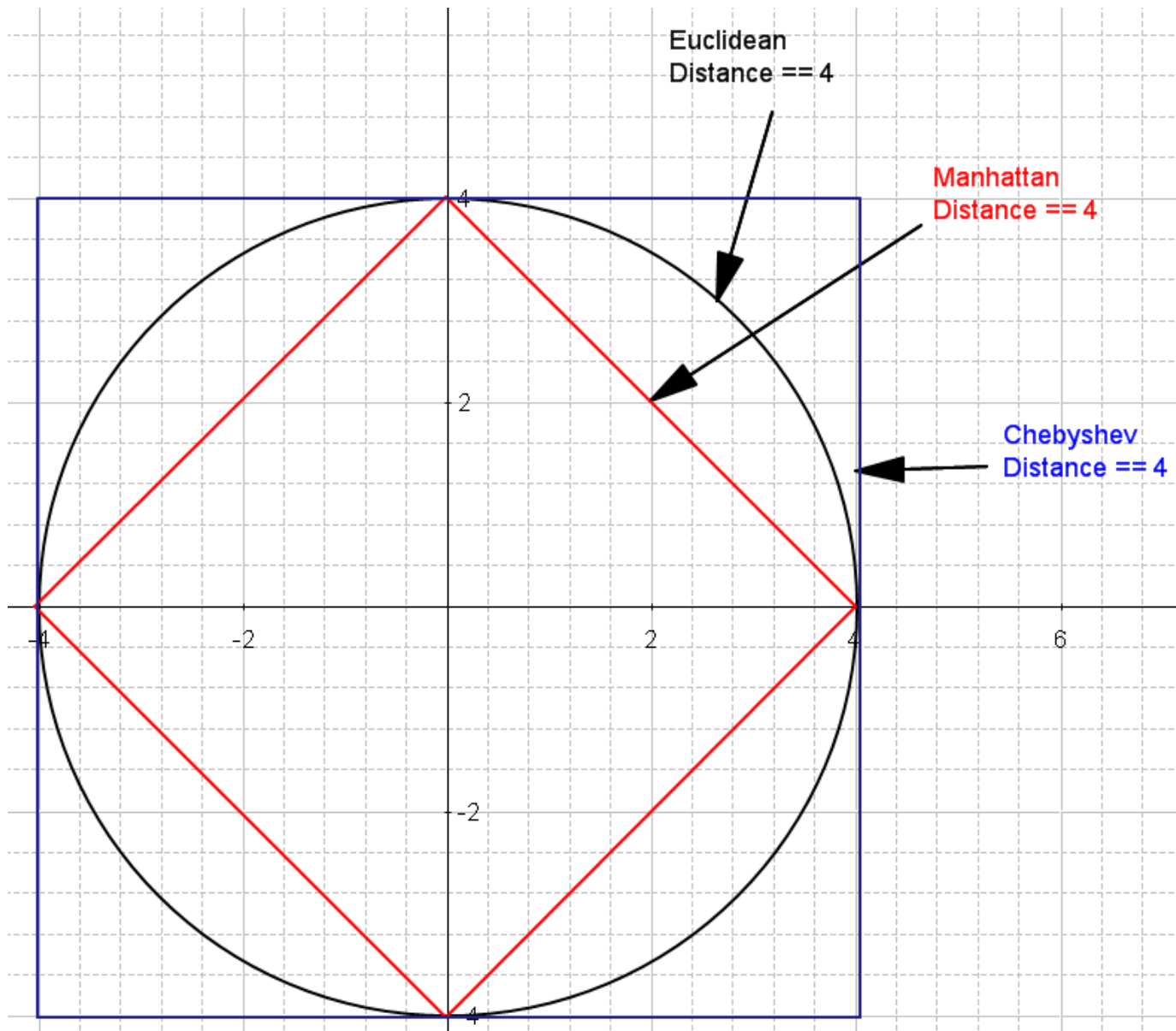
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Minkowski Distance

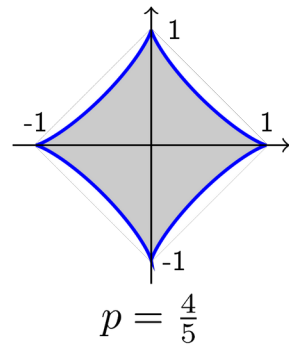
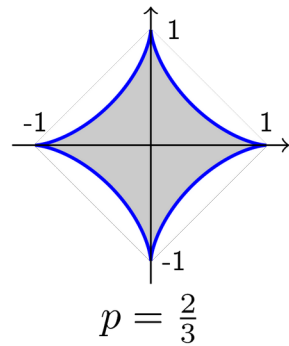
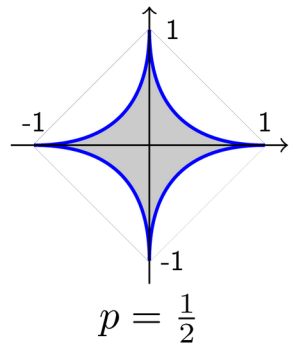
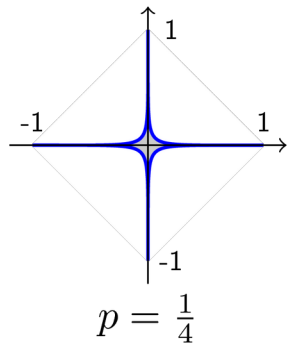
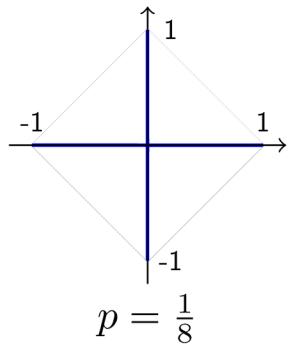


Minkowski Distance

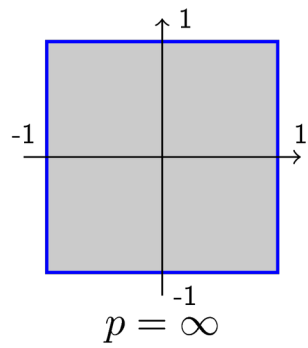
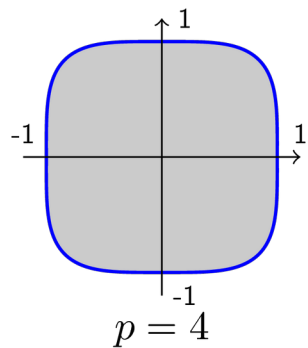
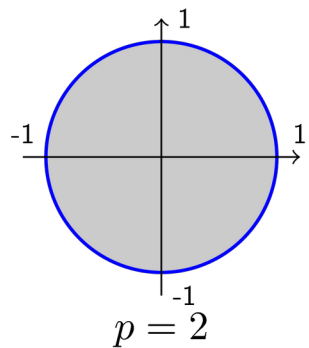
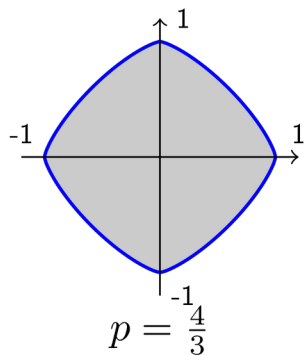
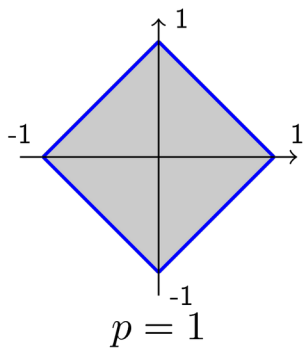


Minkowski Distance

$$\mathcal{C}_p = \{(x, y) \mid (|x|^p + |y|^p)^{1/p} \leq 1\}$$

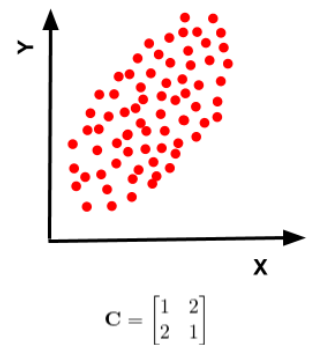
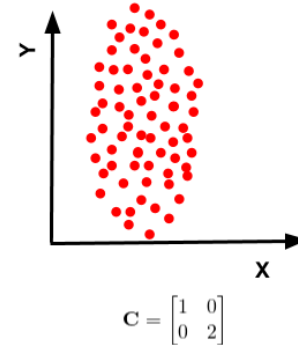
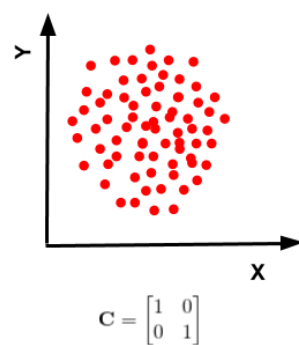
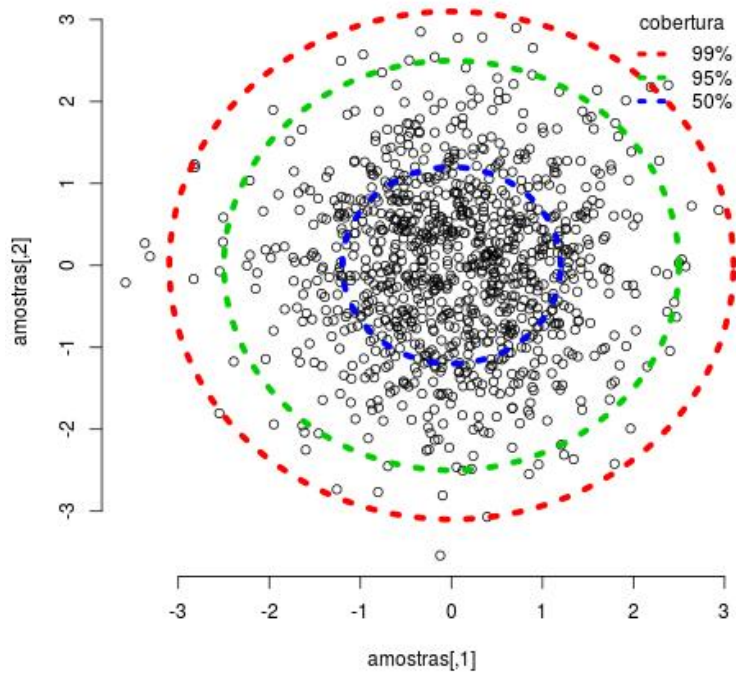


$p < 1$: nonconvex sets



$p \geq 1$: convex sets

Normalized Euclidean



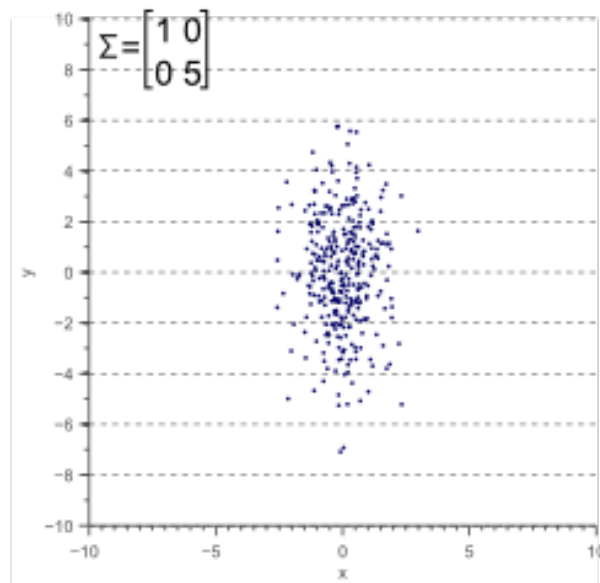
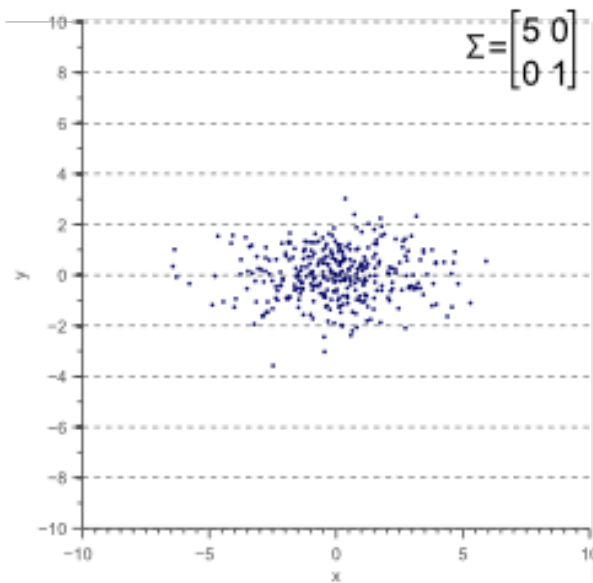
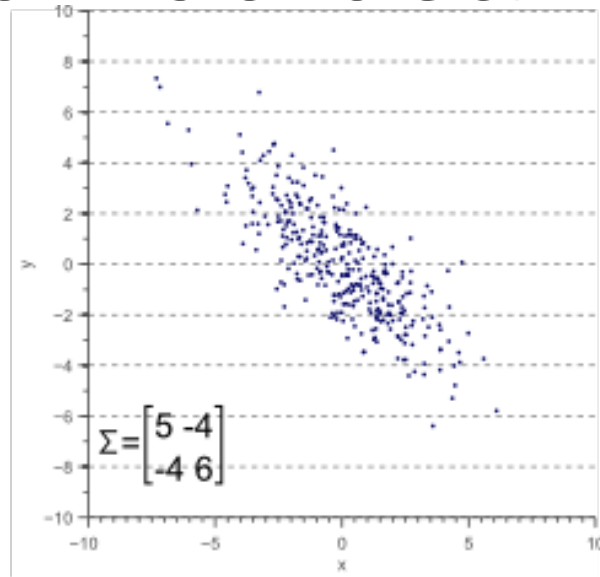
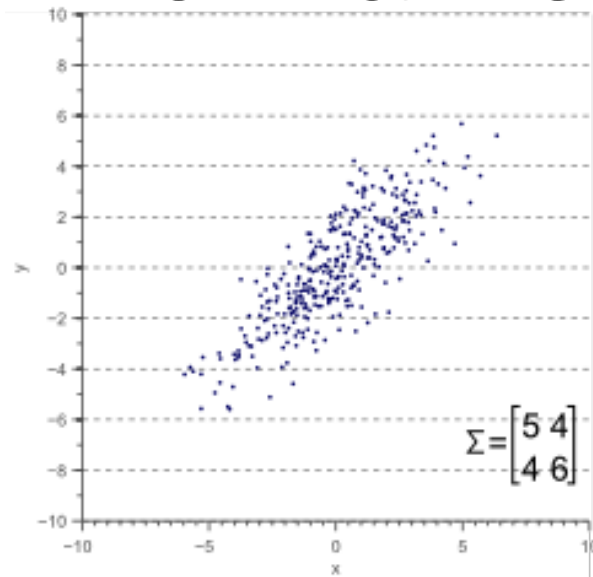
Normalized Euclidean

- Um pescador quer medir a dissimilaridade entre os salmões, para assim classificá-los em tipos e vender por um melhor preço os mais grandes
- Para cada salmão ele mede a comprimento e a largura
- O comprimento dos salmões é uma variável aleatória entre 50 e 100cm
- A largura dos salmões é uma variável aleatória entre 10 e 20cm
- Observa-se que a largura tem valores menores, portanto menos importância terá na distância euclidiana



Salmo salar ♂

Normalized Euclidean



Normalized Euclidean

- Por essa razão o pescador resolve incorporar a estatística dos dados, segundo sua variância
 - As variáveis com menos variância terão mais importância que as de maior variância
- Dessa forma pretende-se igualar a importância do comprimento e da largura na métrica de distância

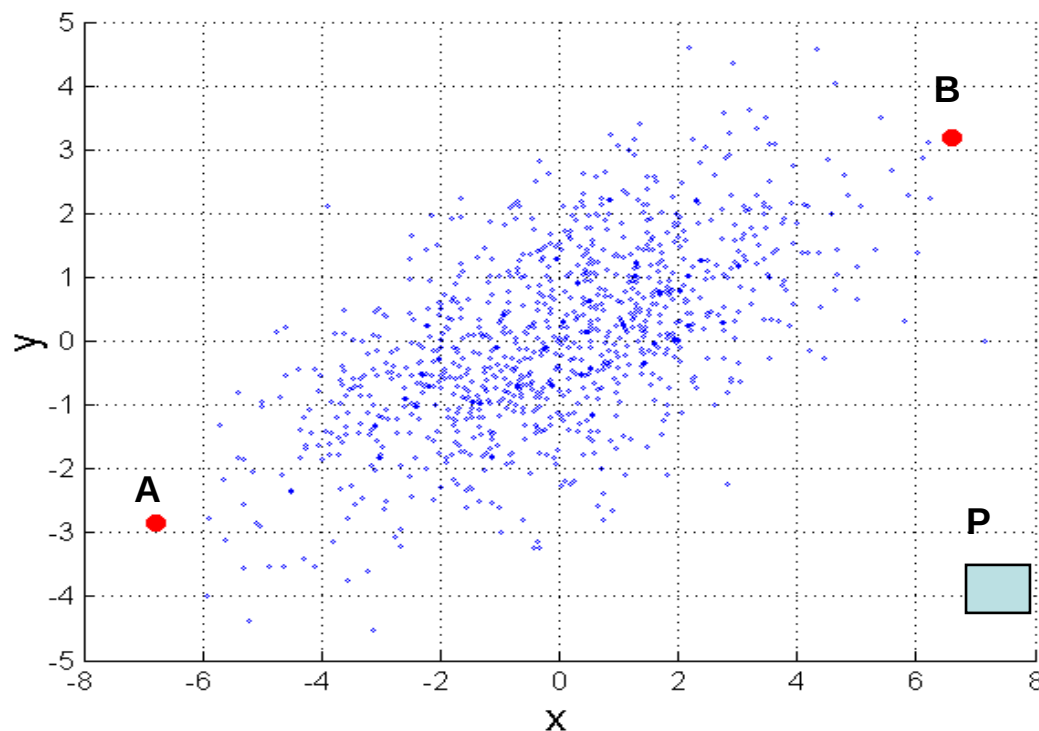
$$d_e(\vec{x}_1, \vec{x}_2) = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$$

$$d_2(\vec{x}_1, \vec{x}_2) = \sqrt{\left(\frac{(x_{11} - x_{12})}{\sigma_1}\right)^2 + \left(\frac{(x_{21} - x_{22})}{\sigma_2}\right)^2}$$

$$d_e(\vec{x}_1, \vec{x}_2) = \sqrt{(\vec{x}_1 - \vec{x}_2)^T S^{-1} (\vec{x}_1 - \vec{x}_2)}$$

Mahalanobis Distance

$$\text{mahalanobis}(p, q) = (p - q)^T \Sigma^{-1} (p - q)$$



Σ is the covariance matrix of the input data X

$$\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

When the covariance matrix is identity Matrix, the mahalanobis distance is the same as the Euclidean distance.

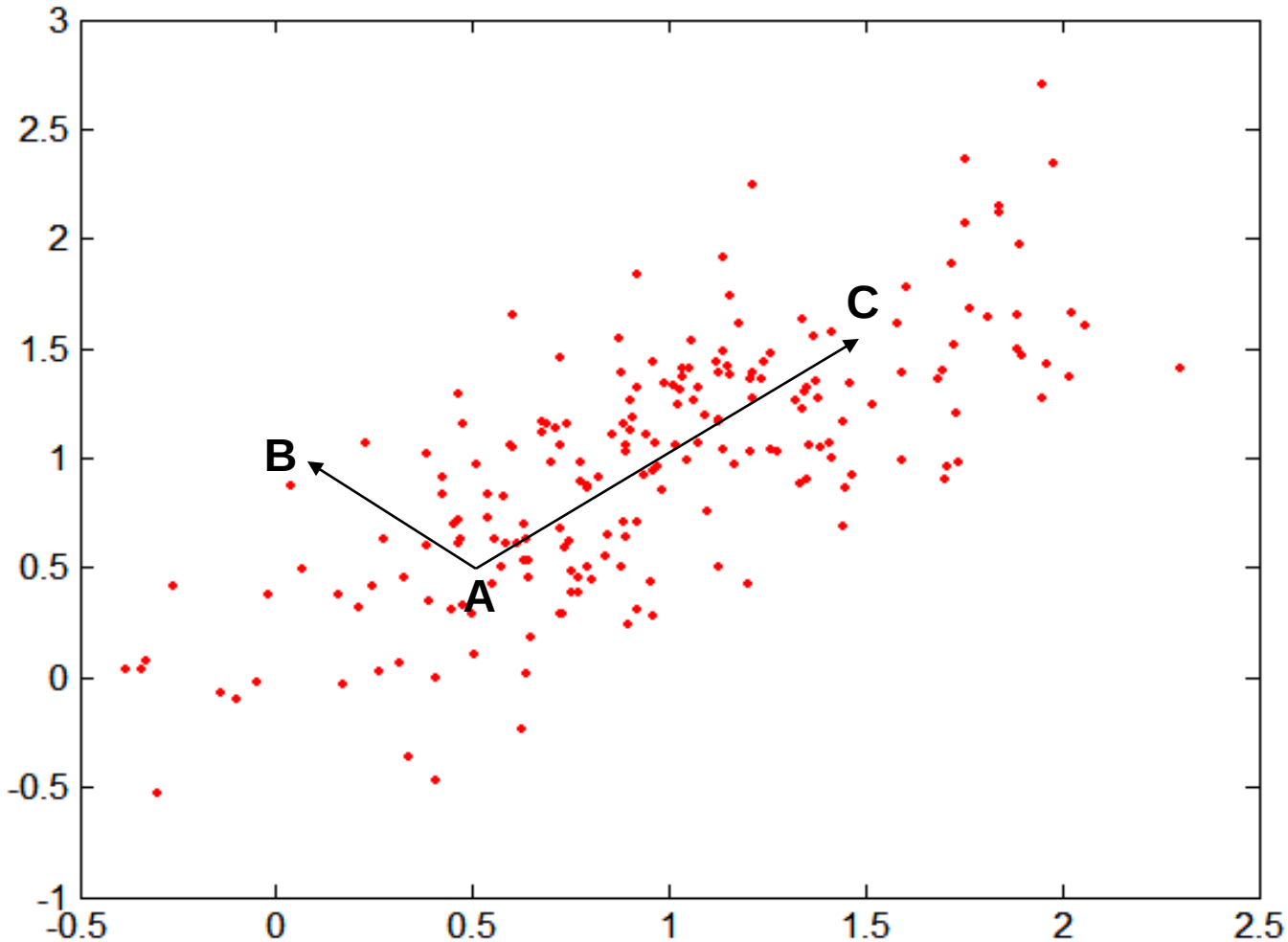
Useful for detecting outliers.

Q: what is the shape of data when covariance matrix is identity?

Q: A is closer to P or B?

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q , and r . (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q .

- A distance that satisfies these properties is a **metric**, and a space is called a **metric space**

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all p and q . (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q .

Similarity Between Binary Vectors

- Common situation is that objects, p and q , have only binary attributes

- Compute similarities using the following quantities

M_{01} = the number of attributes where p was 0 and q was 1

M_{10} = the number of attributes where p was 1 and q was 0

M_{00} = the number of attributes where p was 0 and q was 0

M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Distance/Coefficients

SMC = number of matches / number of attributes

$$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

J = number of value-1-to-value-1 matches / number of not-both-zero attributes values

$$= (M_{11}) / (M_{01} + M_{10} + M_{11})$$

SMC versus Jaccard: Example

$$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$$

$$M_{01} = 2 \quad (\text{the number of attributes where } p \text{ was } 0 \text{ and } q \text{ was } 1)$$

$$M_{10} = 1 \quad (\text{the number of attributes where } p \text{ was } 1 \text{ and } q \text{ was } 0)$$

$$M_{00} = 7 \quad (\text{the number of attributes where } p \text{ was } 0 \text{ and } q \text{ was } 0)$$

$$M_{11} = 0 \quad (\text{the number of attributes where } p \text{ was } 1 \text{ and } q \text{ was } 1)$$

$$\text{SMC} = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0 + 7) / (2 + 1 + 0 + 7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

where \bullet indicates vector dot product and $\| d \|$ is the length of vector d .

- Example:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150, \text{ distance}=1-\cos(d_1,d_2)$$