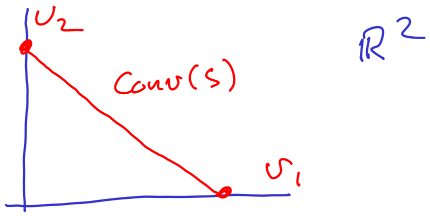
CS4103 - Geometric Foundations of Data Analysis II - Lecture 2

Defn Given a set $S = \{v_1, v_2, \dots, v_k\} \subseteq \mathbb{R}^n$ of vectors in $\mathbb{R}^n$, we define the convex hull

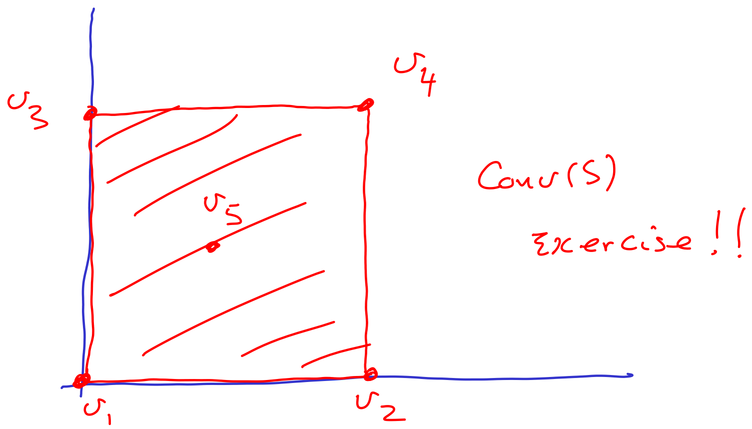
$$\mathrm{Conv}(S) = \left\{ w = \lambda_1 v_1 + \lambda_2 v_2 + \dots + \lambda_k v_k \; : \; \lambda_i \geq 0 \text{ all } i \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

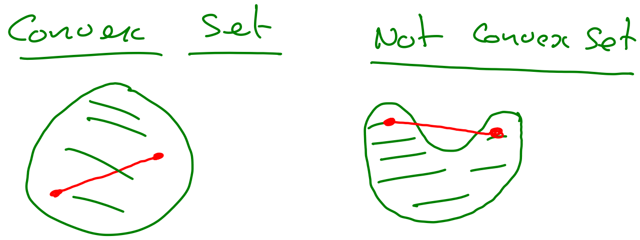
Example $S = \{v_1 = (1,0), v_2 = (0,1)\} \subseteq \mathbb{R}^2$

$$\mathrm{Conv}(S) = \{\lambda_1 v_1 + \lambda_2 v_2 : \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1\}$$



Example $S = \{v_1 = (0,0), v_2 = (1,0), v_3 = (0,1), v_4 = (1,1), v_5 = (\frac{1}{2}, \frac{1}{2})\}$.
$S \subseteq \mathbb{R}^2$





Fact Conv(S) is the "smallest" convex set containing all points of S.

Defn Let $v_0, v_1, \dots, v_k \in \mathbb{R}^d$ be $k+1$ vectors in $\mathbb{R}^d$ such that the k vectors

$$v_1 - v_0, \quad v_2 - v_0, \quad \dots, \quad v_k - v_0$$

are linearly independent. We say that

$$\mathrm{Conv}(\{v_0, v_1, \dots, v_k\})$$

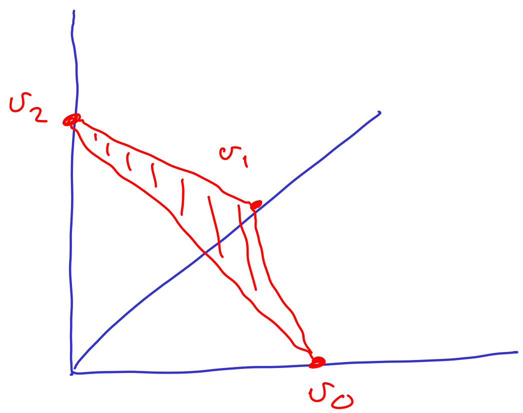
is a geometric k-simplex.

Example $S = \{ v_0 = (1,0,0),\ v_1 = (0,1,0),\ v_2 = (0,0,1) \} \subseteq \mathbb{R}^3$

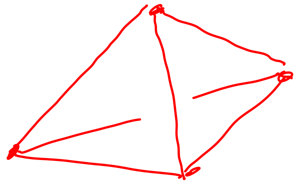
Note $v_1 - v_0 = (-1, 1, 0)$, $v_2 - v_0 = (-1, 0, 1)$ are linearly independent.

So $\mathrm{Conv}(\{v_0, v_1, v_2\})$ is a geometric 2-simplex.

$\mathbb{R}^3$



Example Geometric 3- simplex



Suppose (K, V) is a simplicial complex with finite vertex set $V = \{v_1, v_2, \dots, v_n\}$ say.

Let's identify v_i with the ith standard basis vector of $\mathbb{R}^n$

$$v_i = (0, \dots 0, \underset{\substack{\uparrow \\ \text{position } i}}{1}, 0, \dots 0) \in \mathbb{R}$$

For each k-simplex

$$\sigma = \{v_{i_0}, v_{i_1}, \dots, v_{i_k}\}$$

we let

$$|\sigma| = \mathrm{Conv}(\{v_{i_0}, v_{i_1}, \dots, v_{i_k}\})$$

denote the corresponding geometric k-simplex.

Defn The geometric realization $|K|$ of a simplicial complex K, is the subset of $\mathbb{R}^n$ arising as the union of the geometric simplices $|\sigma|$ with $\sigma \in K$.

Note that $|K|$ is a topological subspace of $\mathbb{R}^n$.

Back to data

Suppose given an $n \times n$ distance matrix $D = (d_{ij})$, recording dissimilarities between n items.

For any $\varepsilon > 0$ we can construct a simplicial complex (K_ε, V) with

$$V = \{1, 2, \cdots, n\}$$

$$K_\varepsilon = \{\sigma \subseteq V : \quad d_{ij} \leq \varepsilon \text{ for all } i, j \in \sigma\}.$$

We can now associate the topological space $|K_\varepsilon|$.

Example

200 people were asked to visit Galway harbour at their convenience once during a 2-week period. They were asked to record the height of the water on their arrival, and then again 2 hours later, and then again 4 hours after arrival. Each person returns the recordings $(h_0, h_2, h_4) \in \mathbb{R}^3$.

From the set

$$S = \{x_i = (h_{i_0}, h_{i_2}, h_{i_4}) : \quad 1 \leq i \leq 200\}$$

a data analyst can construct

$$D = (d_{ij})$$

with

$$d_{ij} = \|x_i - x_j\| \qquad \text{Euclidean metric.}$$

The analyst could view the (graph of the) simplicial complex K_ε for various values of ε.

A computer demonstration (see video) shows that for larger values of ε the simplicial complexes K_ε seem to contain a "dominant circular hole".

A possible interpretation is that the "circular hole" might represent some periodicity in the data. It seems reasonable to conjecture that this might be related to the periodicity of the moon.

This example shows how a good representation of our data set can lead to interesting/surprising conjectures about the data.