CS 4103 - Geometric Foundations of Data Analysis II

PCA & Johnson-Lindenstrauss

$\rho : \mathbb{R}^d \longrightarrow \mathbb{R}^k$

Represent data points $v \in \mathbb{R}^d$ as points $\rho(v) \in \mathbb{R}^k$, $k < d$.

Clustering

$\rho : \mathbb{R}^d \longrightarrow X$

Represents data points $v \in \mathbb{R}^d$ as points $\rho(v) \in X$, X a finite space of points.

In general, let's define dimension reduction as any useful representation

$\rho : \mathbb{R}^d \longrightarrow X$

with X a simpler (topological) space.

Example Nicolau, Levine, Carlsson (2011)

295 samples from breast cancer tumors.

15 samples from normal healthy breast tissue

A dissimilarity matrix (310 × 310) was created and (somehow) used the matrix to create a graph X.

In the graph X the nodes are "bins" containing samples. A sample may lie in two bins, in which case we join the bins/nodes are connected by an edge.

Bins are coloured:

Blue = Similar to normal healthy samples

Red = very different to normal samples

ER^+ = Estrogen receptor positive

ER^{-1} = " " negative

ER^+ branch of the graph had a good survival rate.

The c-MYB+ portion of the graph is defined as the region lying between the two sparse regions.

The graph was constructed using only the dissimilarity matrix.

However the c-MYB+ region had a 100% death rate.

Conclusion: c-MyB+ warrant being identified as a breast cancer group of genes.

Defn A simplicial complex consists of a set V and a collection K of certain subsets $\sigma \subseteq V$. The following axioms must hold:

1) $\{v\}$ is in the collection K for each $v \in V$.

2) If $\sigma \in K$ and if $\phi \neq \sigma' \subseteq \sigma$ then $\sigma' \in K$.

We call the elements $v \in V$ vertices, and the subsets $\sigma \in K$ simplices.

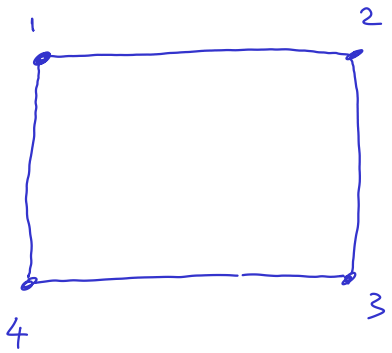
We call $\sigma \in K$ an n-simplex if $\sigma = \{v_0, v_1, \dots, v_n\}$ consists of $n+1$ vertices in V.

Example $V = \{1, 2, 3, 4\}$

$K = \{ \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{2,3\}, \{3,4\}, \{1,4\} \}$

In this example we have only 0-simplices and 1-simplices.

We can picture this example.



Example 2 $V = \{1, 2, 3, 4, 5, 6\}$

$K = \{ \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\},$
$\{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{4,5\}, \{4,6\}, \{5,6\},$
$\{1,2,3\}, \{4,5,6\} \}$.

We have six 0-simplices, seven 1-simplices, two 2-simplices. We will often call a 1-simplex and edge.

The picture: