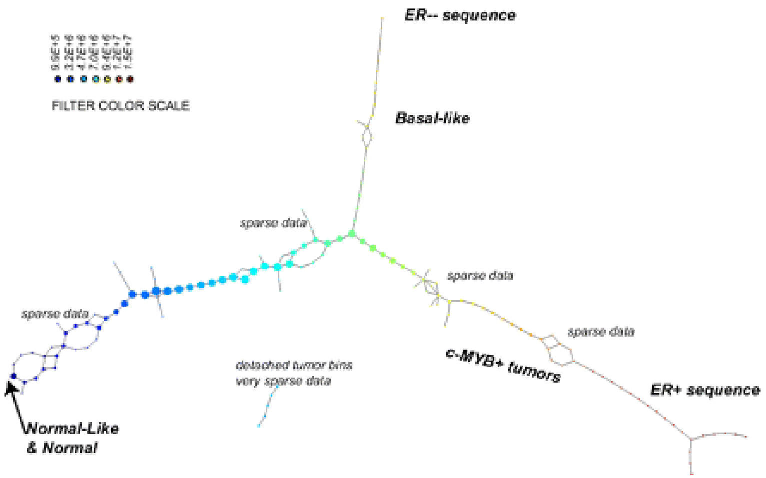
Breast cancer microarray gene expression data



Nicolau, Levine, Carlsson (PNAS, 2011): identified a subgroup of ER+ breast cancers. These patients exhibit 100% survival.

Part 2

PCA & Johnson-Lindenstrauss $\rho : \mathbb{R}^d \longrightarrow \mathbb{R}^k$

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

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

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

In general, let's define <u>dimension reduction</u> as <u>any</u> 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 x 310) was created and (somehow) used to create a graph X.

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

Bins are coloured:

Blue = similar to normal

Red = very different to normal

ER^+ = Estrogen receptor positive

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

The c-Myb$^+$ portion of the graph is defined as the region lying between two sparse regions

The graph was constructed using only the dissimilarity matrix.

However c-Myb$^+$ region had 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 all $v \in V$.
2) If $\sigma \in K$ and if $\emptyset \neq \sigma' \subset \sigma$ then $\sigma' \in K$.

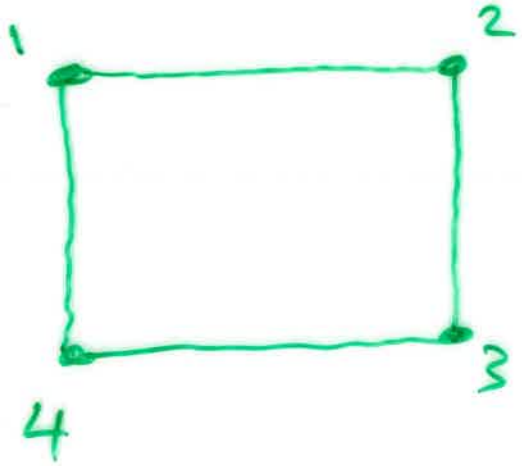
We call the elements $v \in V$ *vertices*, and the subsets $\sigma \in K$ *simplices*. We call $\sigma \in K$ an *n-simplex* if $\sigma = \{v_0, v_1, \dots, v_n\}$ consists of $n+1$ vertices.

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 call 0-simplices vertices, and 1-simplices edges. We can picture K as a graph



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

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

we have six 0-simplices, seven 1-simplices, and two 2-simplices

we picture K as