

### Theorem (Johnson - Lindenstrauss)

Let  $\varepsilon \in (0, \frac{1}{2})$ . Let  $S \subseteq \mathbb{R}^d$  be a set of  $n$  points. Let

$$k = \frac{20 \log(n)}{\varepsilon^2} \quad \text{integer ceiling}$$

Then there exists a linear homomorphism

$$f: \mathbb{R}^d \rightarrow \mathbb{R}^k$$

such that, for all  $u, v \in S$ ,

$$(1-\varepsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2$$

### Proposition (Norm preservation)

Let  $x \in \mathbb{R}^d$ . Let  $A$  be a  $k \times d$  matrix with entries independently sampled from  $N(0, 1)$ .

Consider the event

$$E_x: (1-\varepsilon)\|x\|^2 \leq \left\| \frac{1}{\sqrt{k}} Ax \right\|^2 \leq (1+\varepsilon)\|x\|^2$$

Then

$$\text{Prob}(E_x) \geq 1 - 2e^{-(\varepsilon^2 - \varepsilon^3)k/4}$$

Proof that Prop implies Thm

$\text{Prob}(\exists u, v \in S \text{ such that } E_{u,v} \text{ fails})$

$$\leq \sum_{u, v \in S} \text{Prob}(E_{u,v} \text{ fails})$$

$$\leq n^2 2 e^{-(\varepsilon^2 - \varepsilon^3)k/4}$$

$$= 2n^2 e^{-(\varepsilon^2 - \varepsilon^3) \frac{20 \log(n)}{\varepsilon^2 \cdot 4}}$$

$$= 2n^2 \left( e^{\log(n)} \right)^{-(\varepsilon^2 - \varepsilon^3) 5/\varepsilon^2}$$

$$= 2n^2 / n^{5(1-\varepsilon)}$$

$$\leq 2n^2 / n^{\frac{1}{2}} n^2$$

for  $n \geq 4$

$$< 1$$

Thus the probability of event  $E_{u,v}$  holding for all  $u, v \in S$  is  $> 0$ .

Let  $A$  be the matrix for  $\delta_{u,v}$  to hold for all  $u, v \in S$ . Set

$$f: \mathbb{R}^d \rightarrow \mathbb{R}^k, \quad x \mapsto \frac{1}{\sqrt{12}} A x.$$

See link on web page for a proof of the norm Proposition.

END OF PART 1

## Part 2

PCA

&

Johnson-Lindenstrauss

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

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

clustering  $\rho: \mathbb{R}^d \rightarrow X$

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

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

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

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 ( $310 \times 310$ ) 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

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

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

The graph was constructed using only the dissimilarity matrix. However, the  $c\text{-Myb}^+$  region had 100% survival rate.

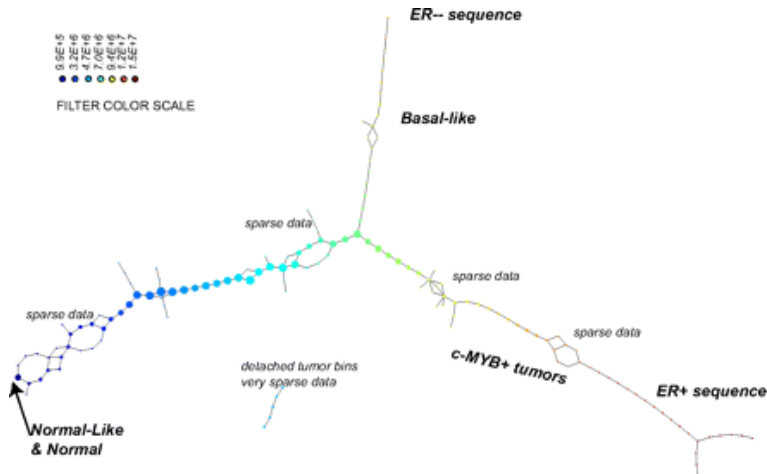
Conclusion:  $c\text{-Myb}^+$  warrants being identified as a breast cancer group of genes.

# Breast cancer microarray gene expression data





## Breast cancer microarray gene expression data



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