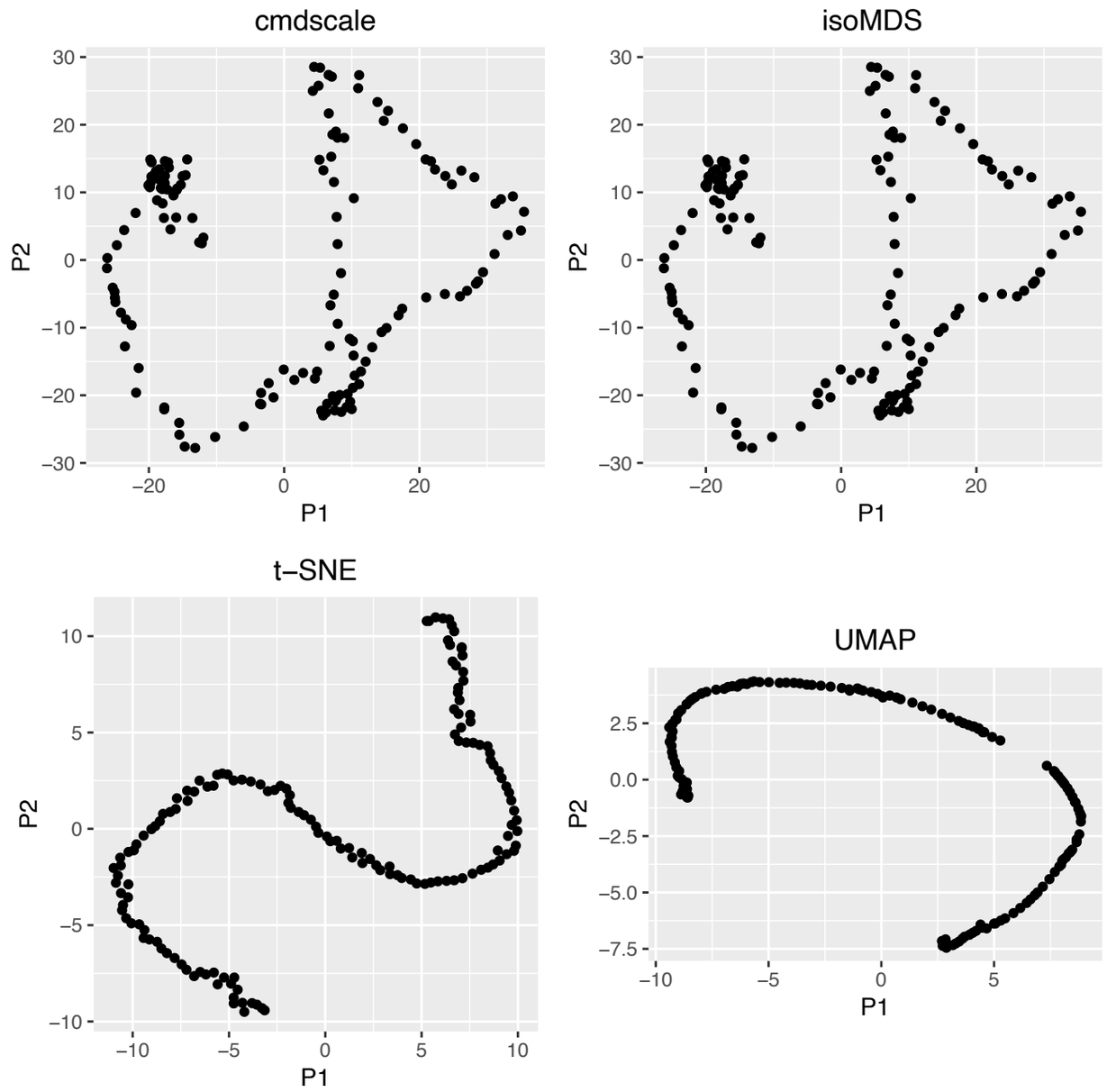
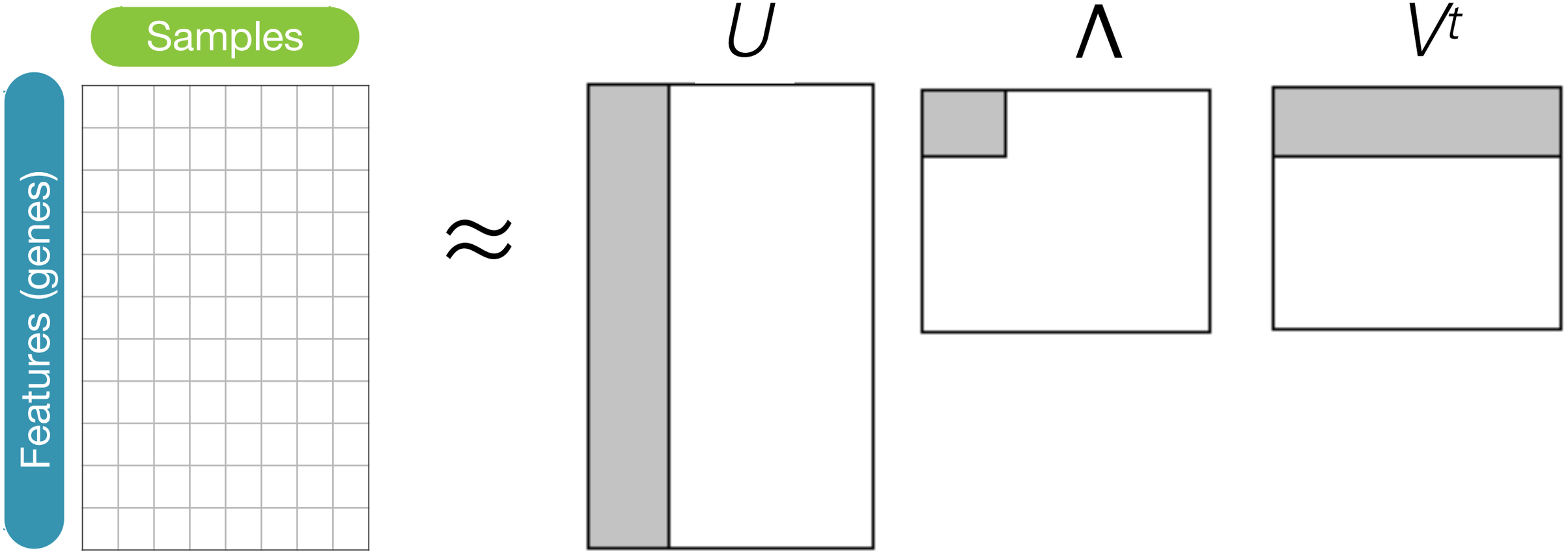
On low-dimensional embeddings of high-dimensional data



Wolfgang Huber and Susan Holmes

Dimension reduction / embedding



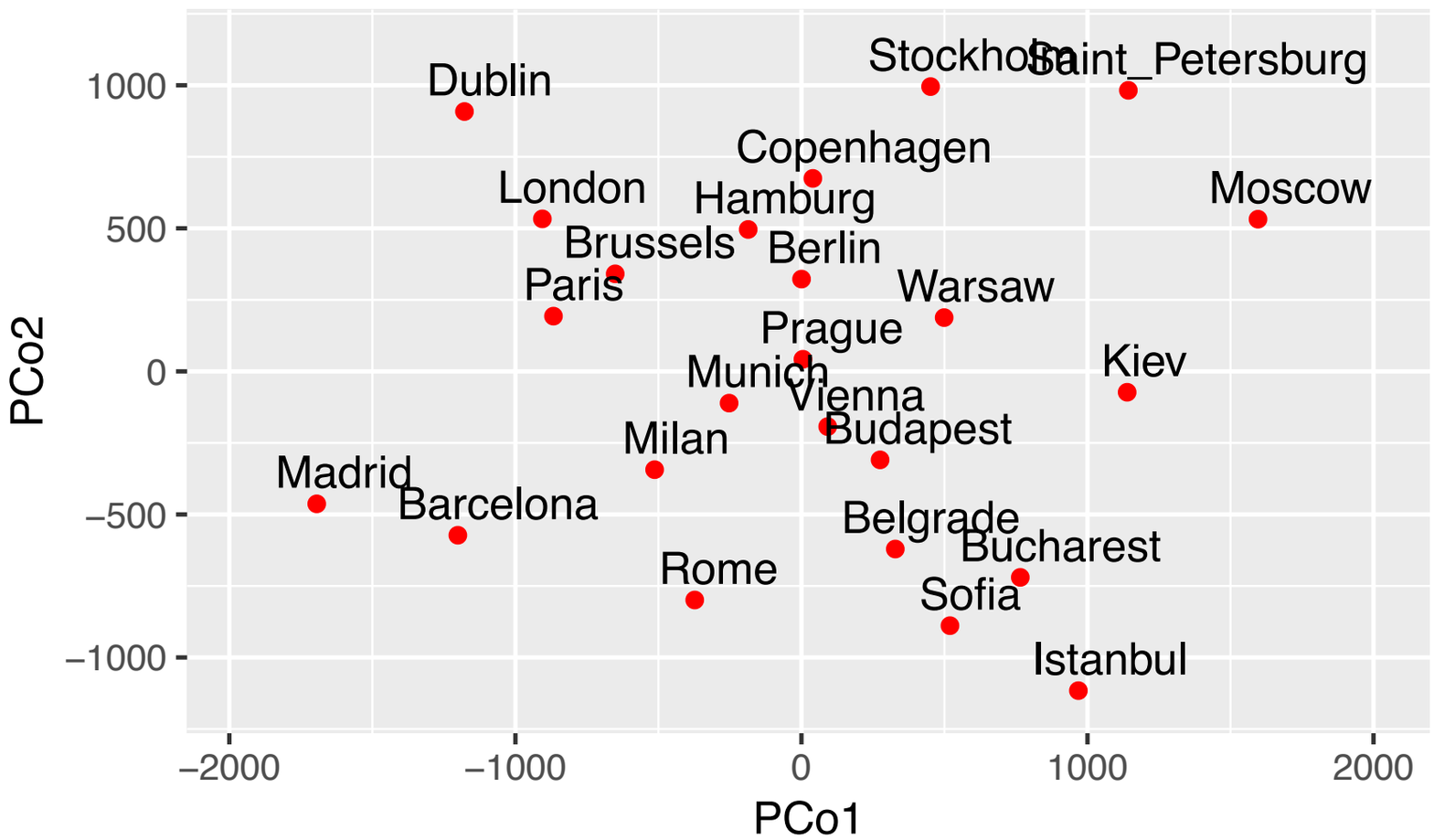
$n \times m$ $\quad\quad$ $n \times p$ $\quad\quad$ $p \times p$ $\quad\quad$ $p \times m$

$n = 20000$, $m=1000$:
$n \times m = 2 \times 10^8$

$p = 3$
$(n + 1 + m) \times p = 63003$

Applications: Principal component analysis (PCA), Non-negative matrix factorization (NMF), ...

Multi-dimensional scaling (MDS)



Two-dimensional layout of European cities based on matrix D of their pairwise distances

Classical MDS is achieved by singular value decomposition of (double centred) D^2. In R: `cmdscale`

Non-linear extensions: t-SNE, UMAP, ...

What does this have to do with RNA-seq?

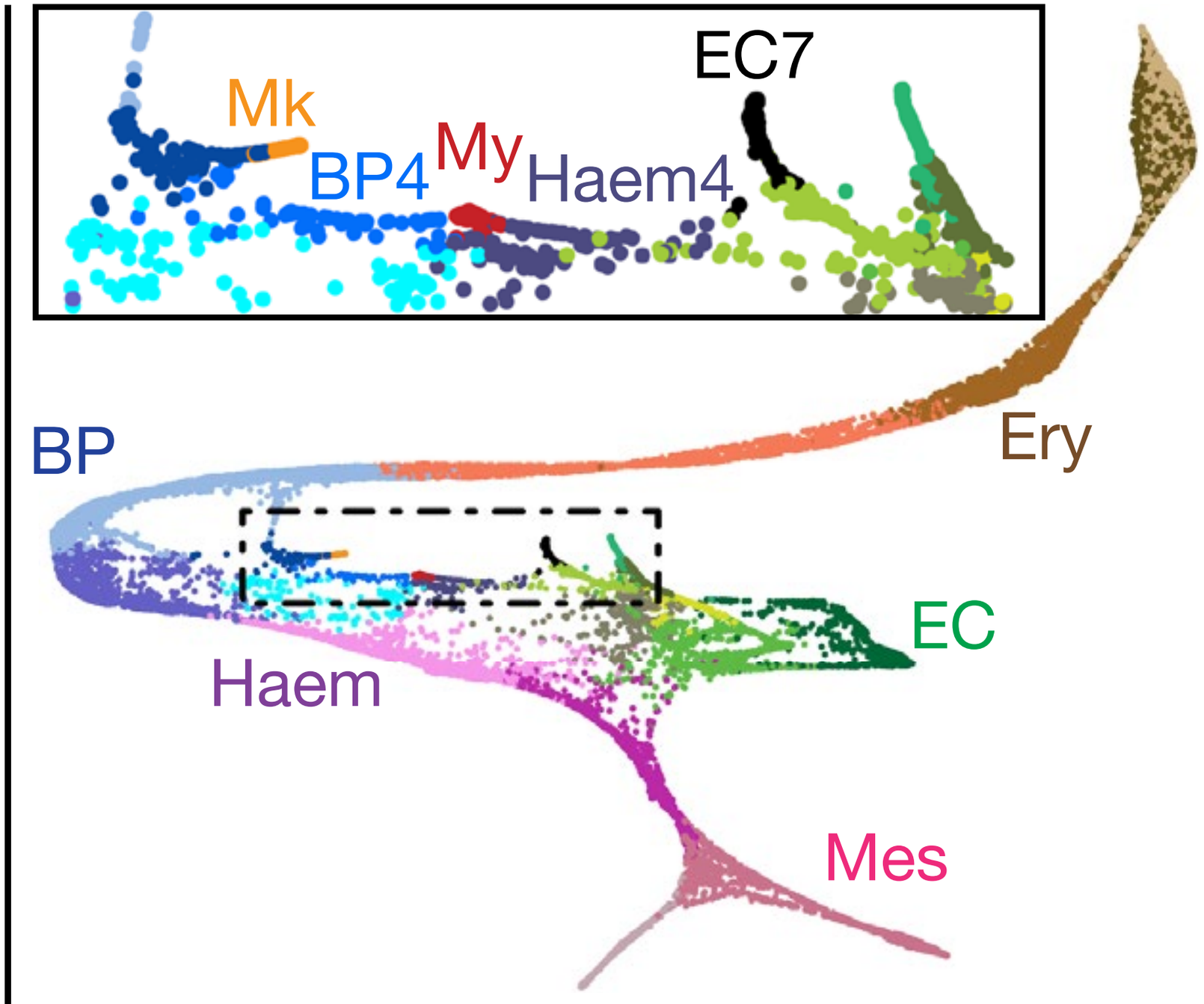


Fig. 3a from Pijuan-Sala et al. (Nature 2019)

15,875 cells from blood lineage (out of 116,312) from mouse embryos at 9 time points from 6.5-8.5 days post-fertilization

Clusters	Trajectories	Branch points
👉🏽 Cell types	👉🏽 Differentiation	👉🏽 Lineages

But the geometry of high-dimensional spaces is weird

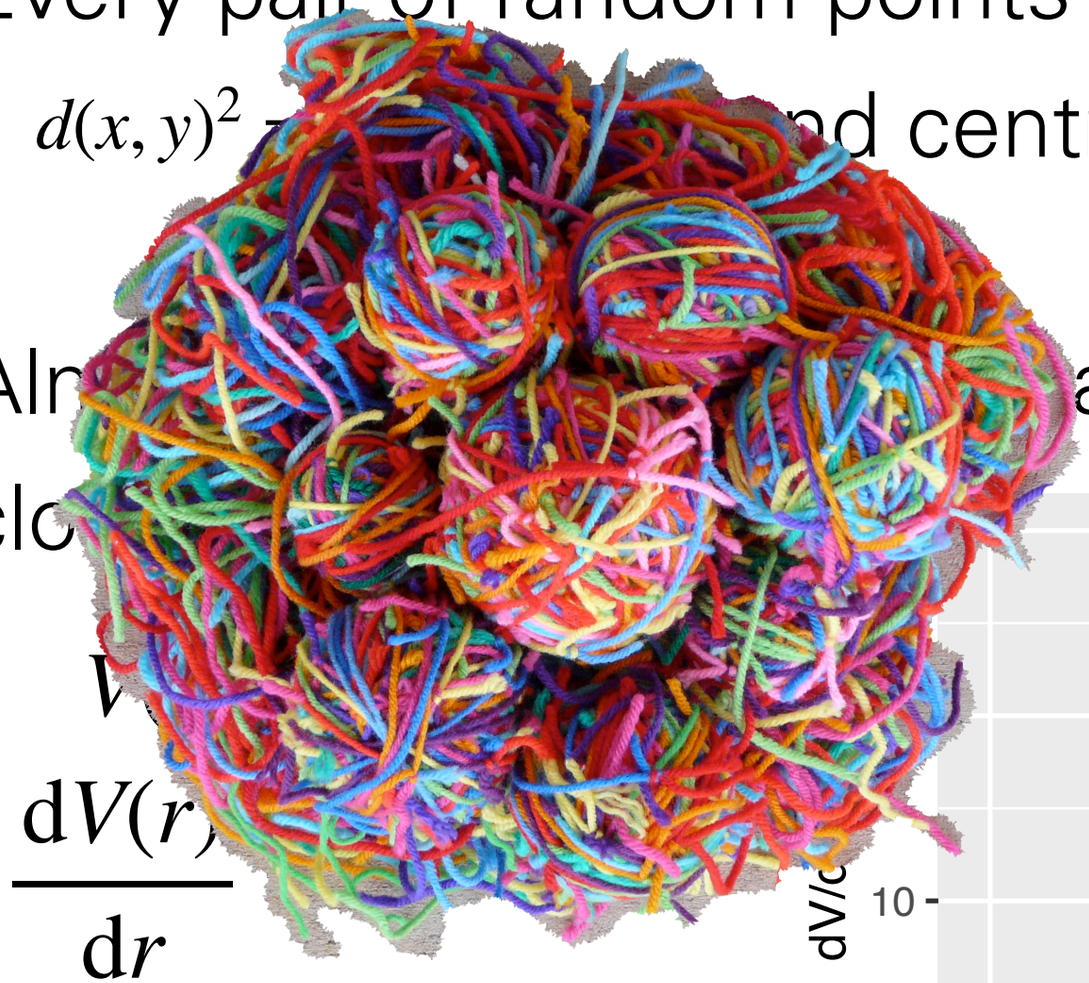
- Every pair of random points has nearly the same distance:

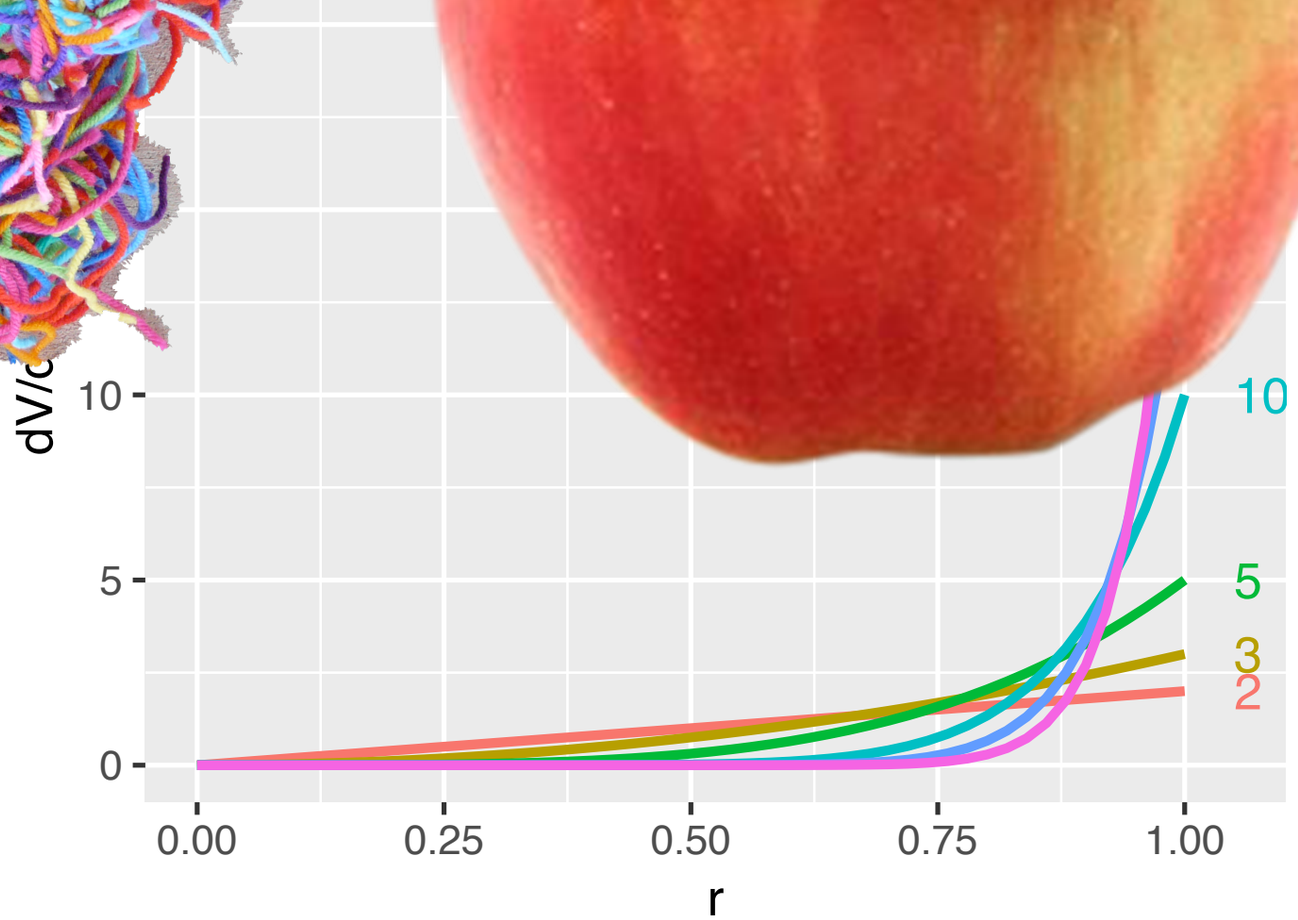
  $d(x, y)^2 \rightarrow$ nd central limit

- Aln ape e) is clo

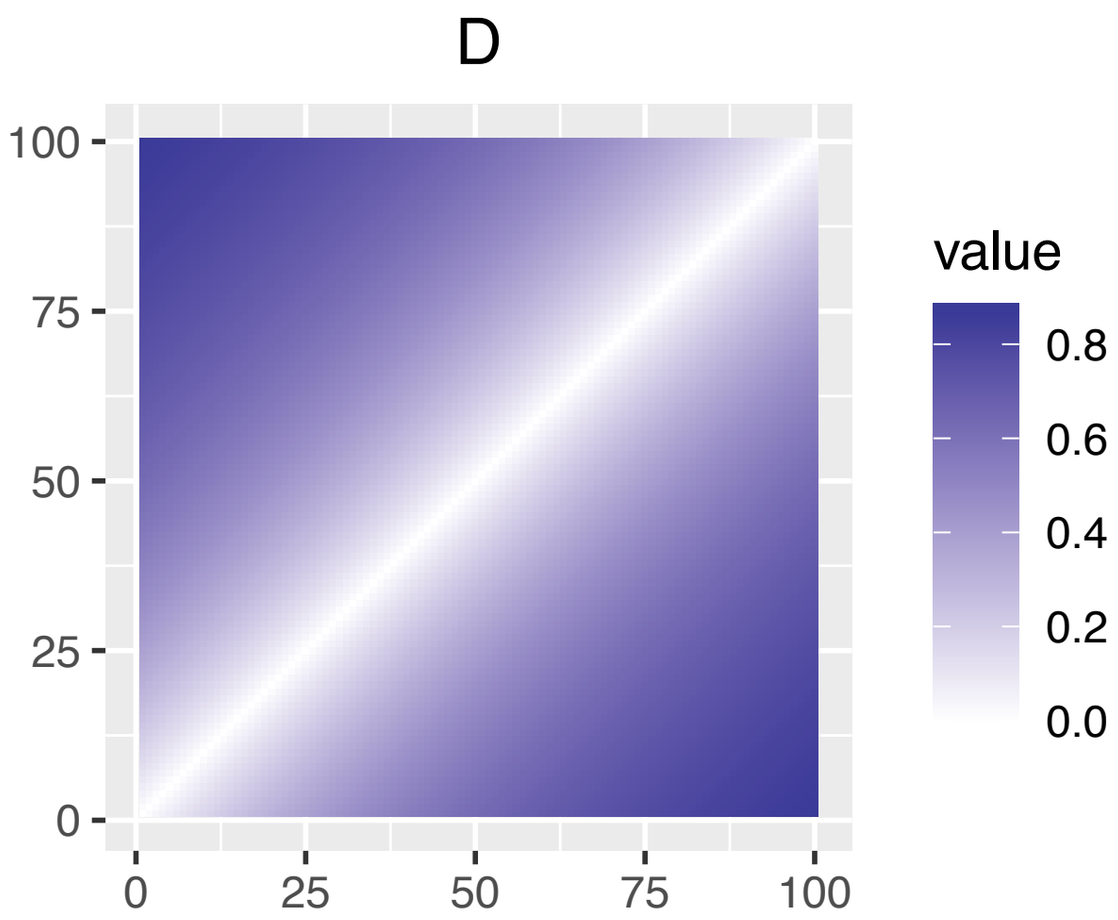
$$V$$

$$\frac{\mathrm{d}V(r)}{\mathrm{d}r}$$





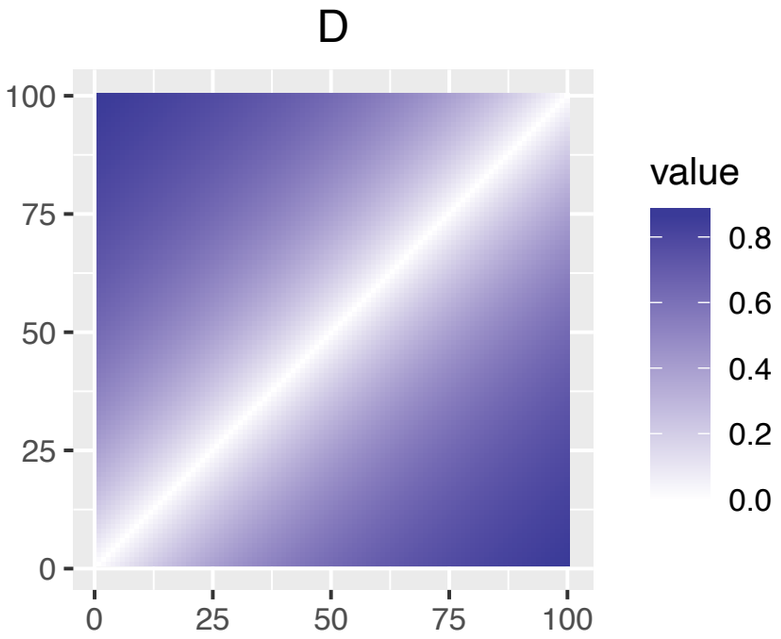
MDS of 100 objects

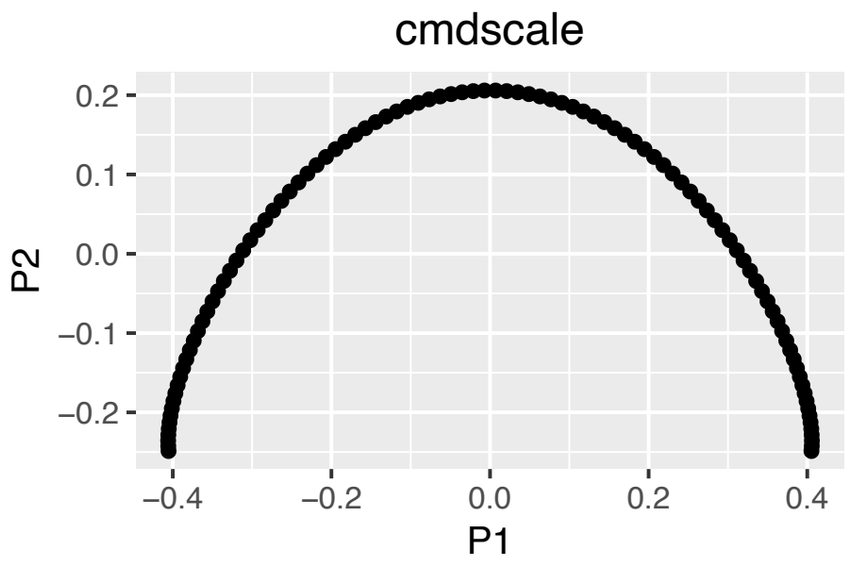


idealized model for such data:

$$D_{ij} = 1 - e^{-\lambda|i-j|}$

MDS of a 100-dimensional dataset

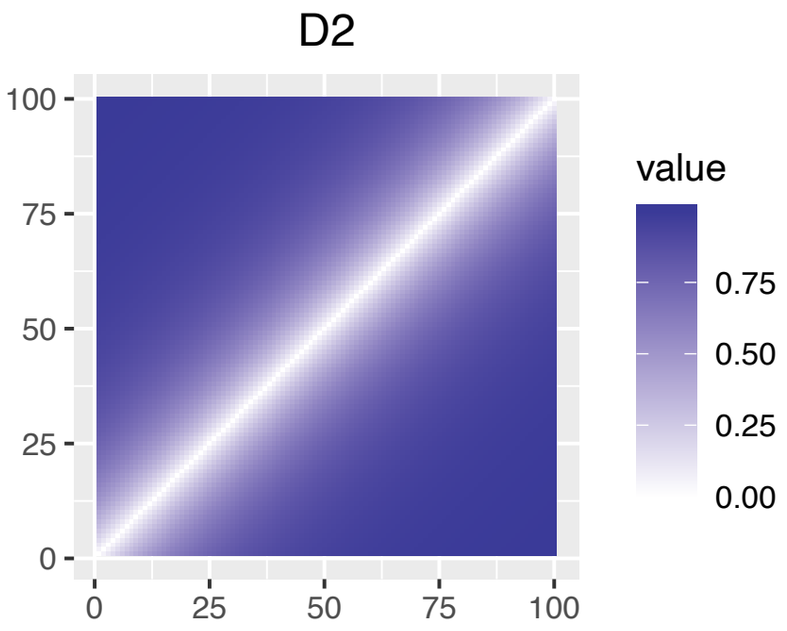




$$D_{ij} = 1 - e^{-\lambda |i-j|}$$

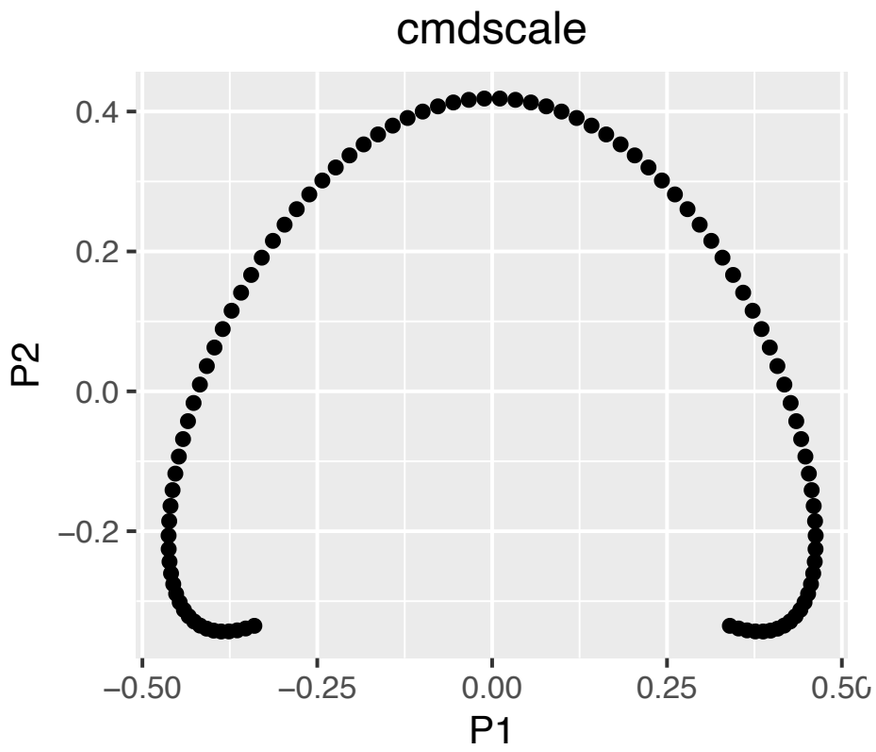
$$\lambda = 2$$

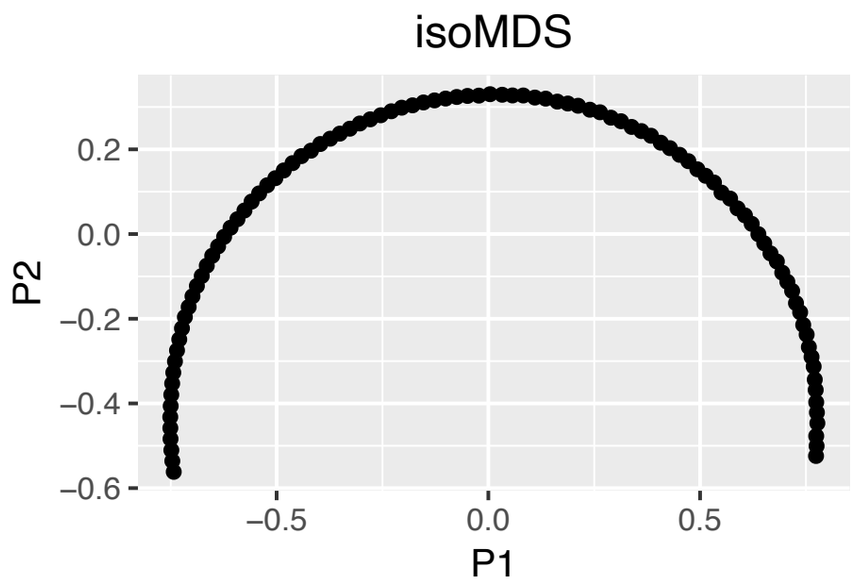
MDS of a 100-dimensional dataset

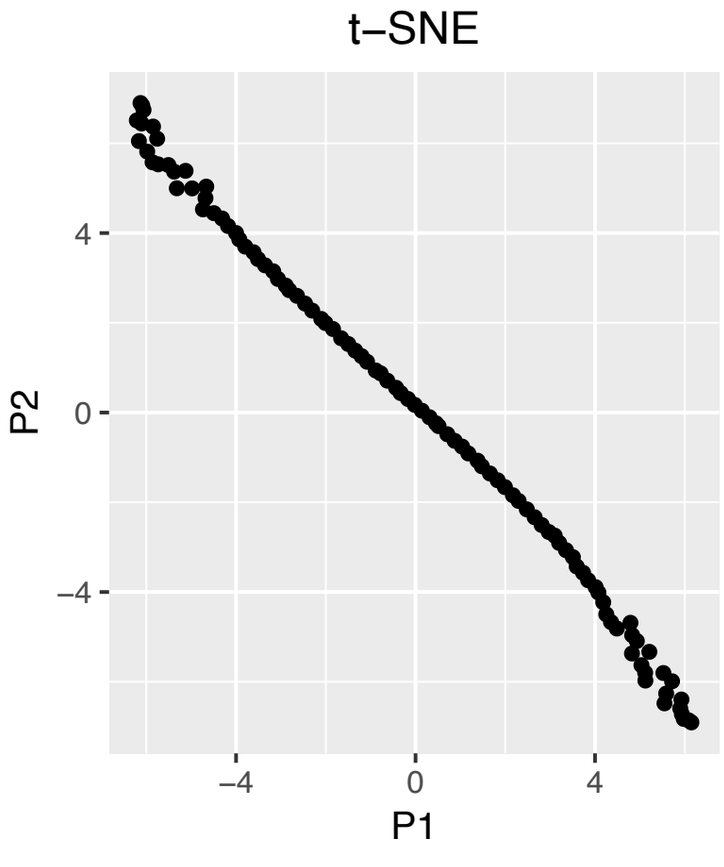


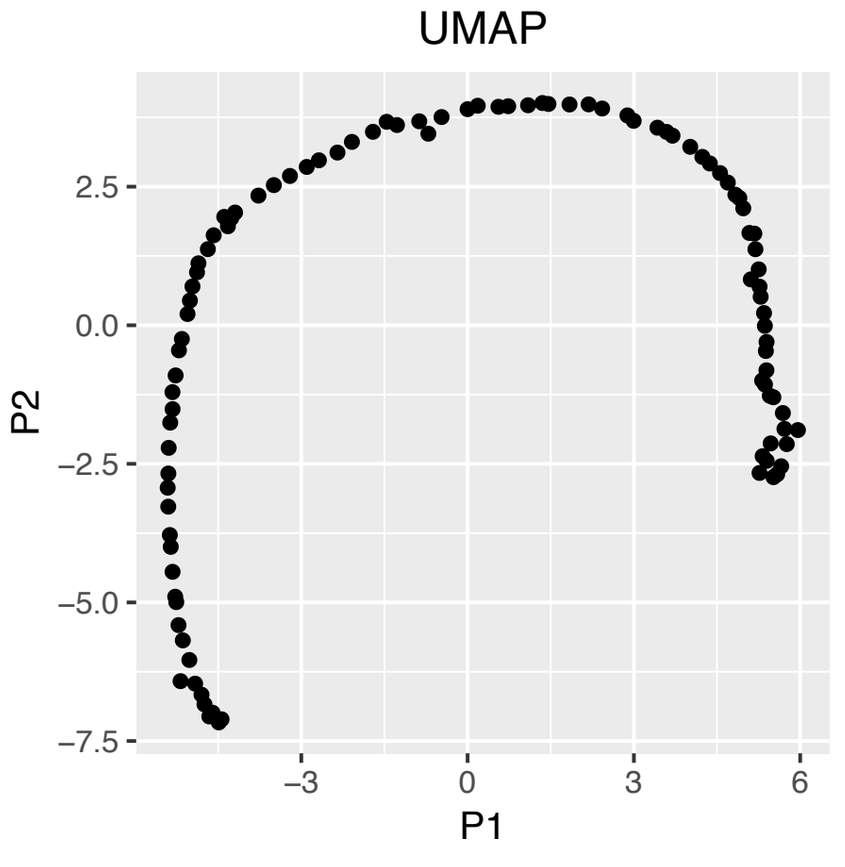
$$D_{ij} = 1 - e^{-\lambda|i-j|}$$

$$\lambda = 6$$

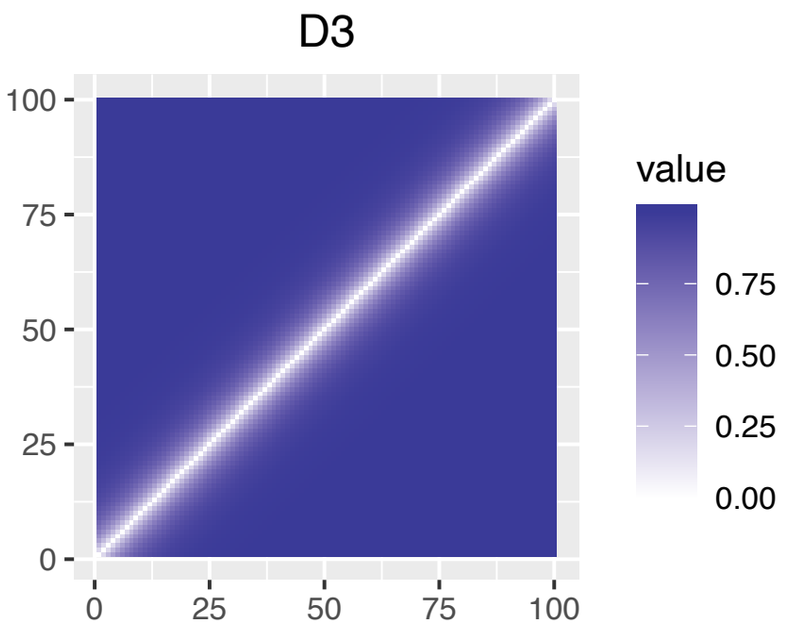






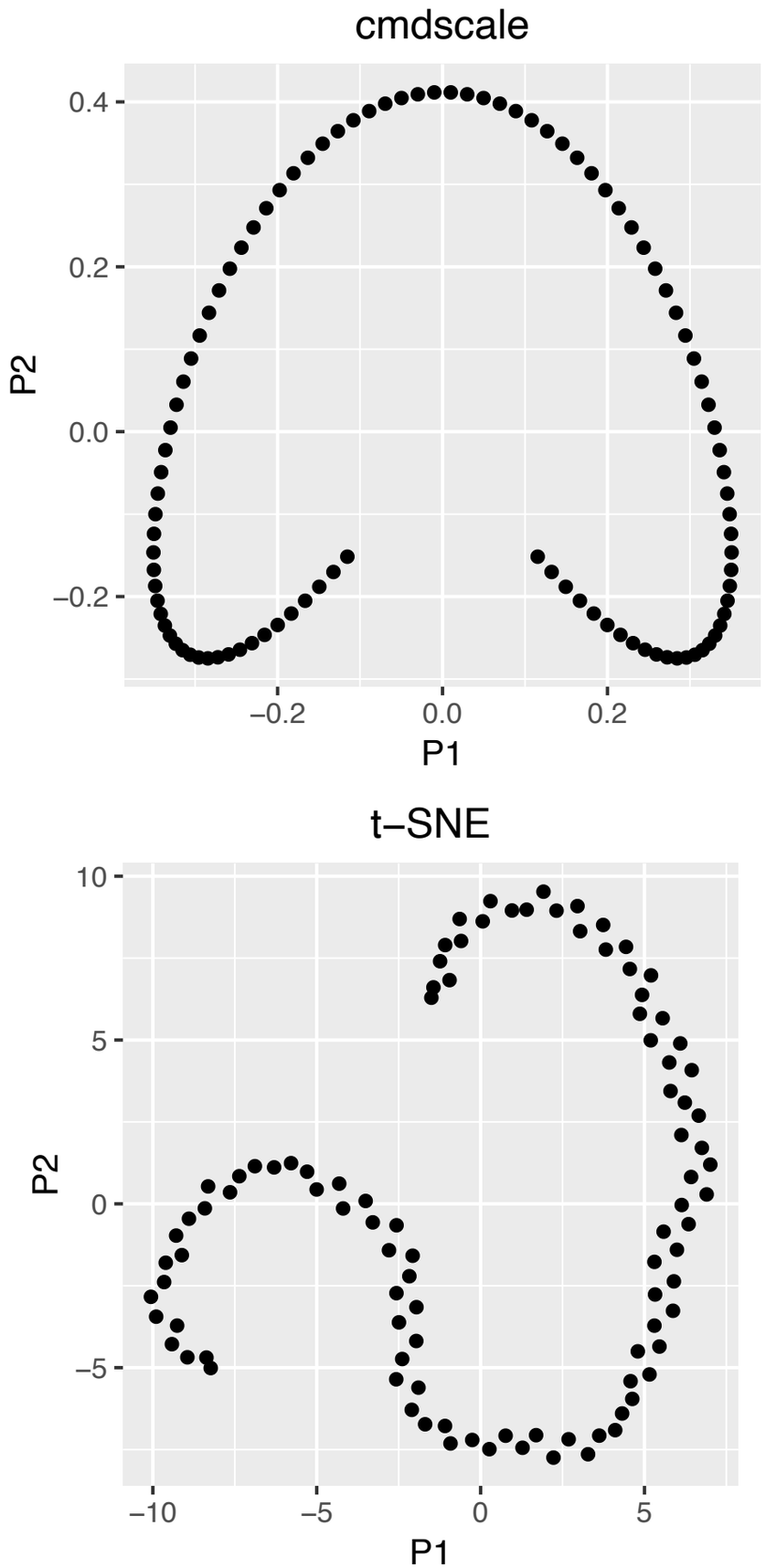


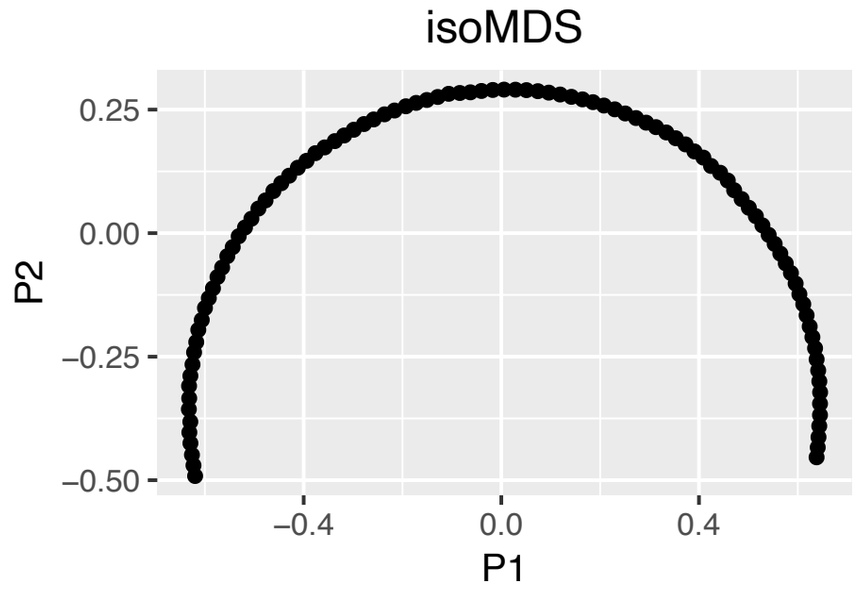
MDS of a 100-dimensional dataset

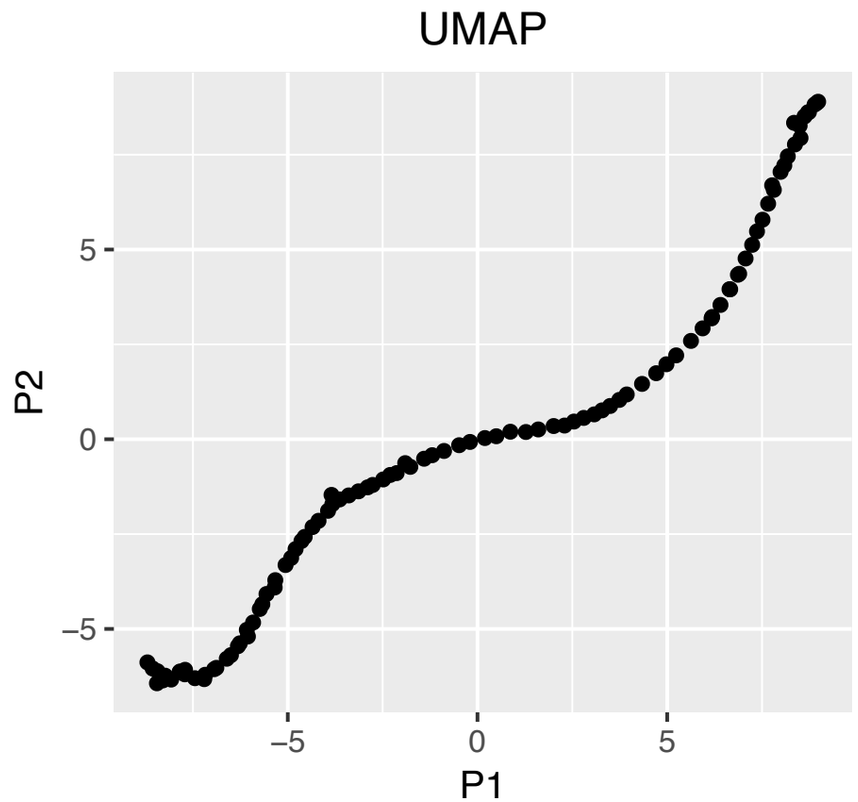


$$D_{ij} = 1 - e^{-\lambda|i-j|}$$

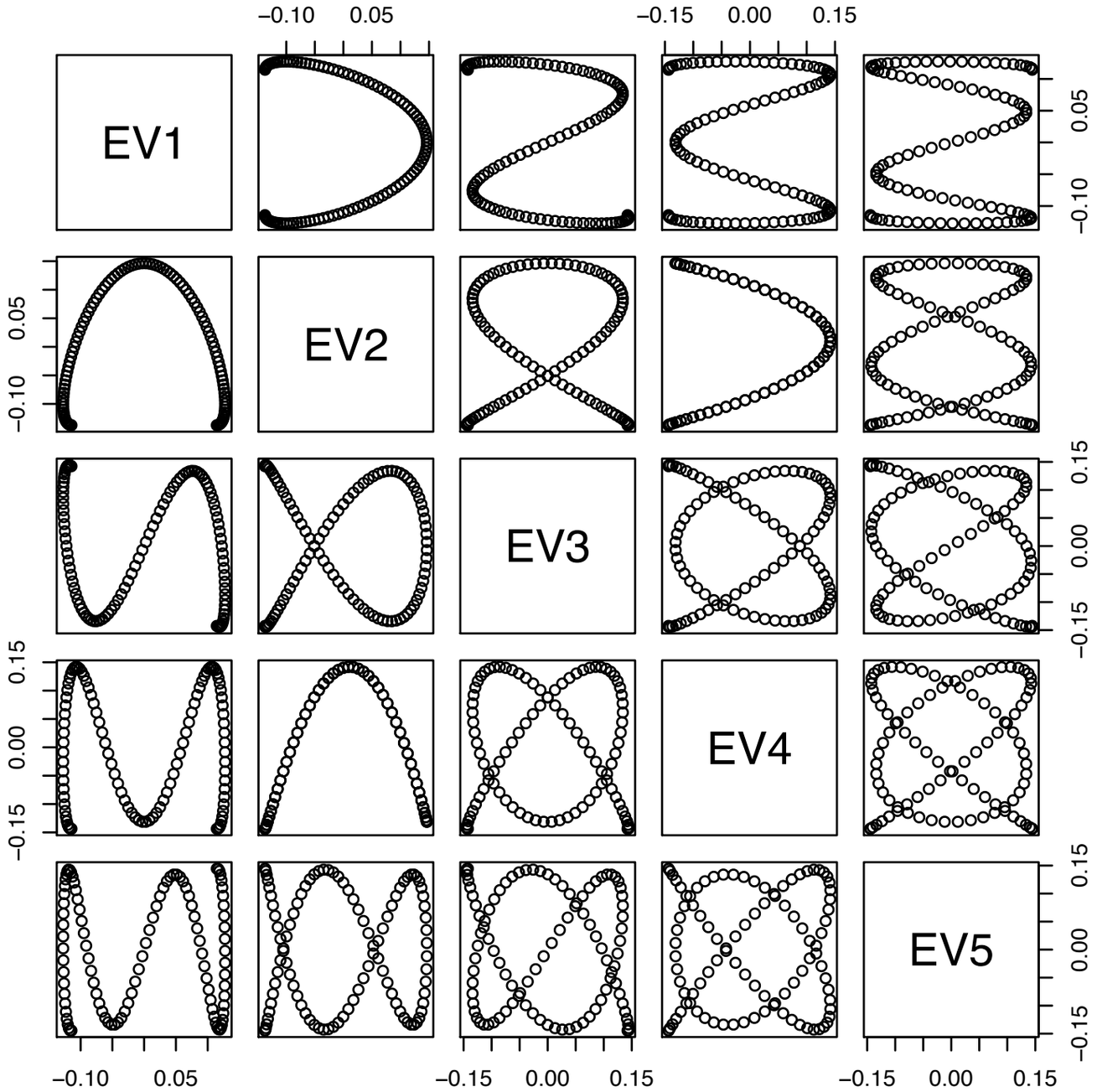
$$\lambda = 20$$

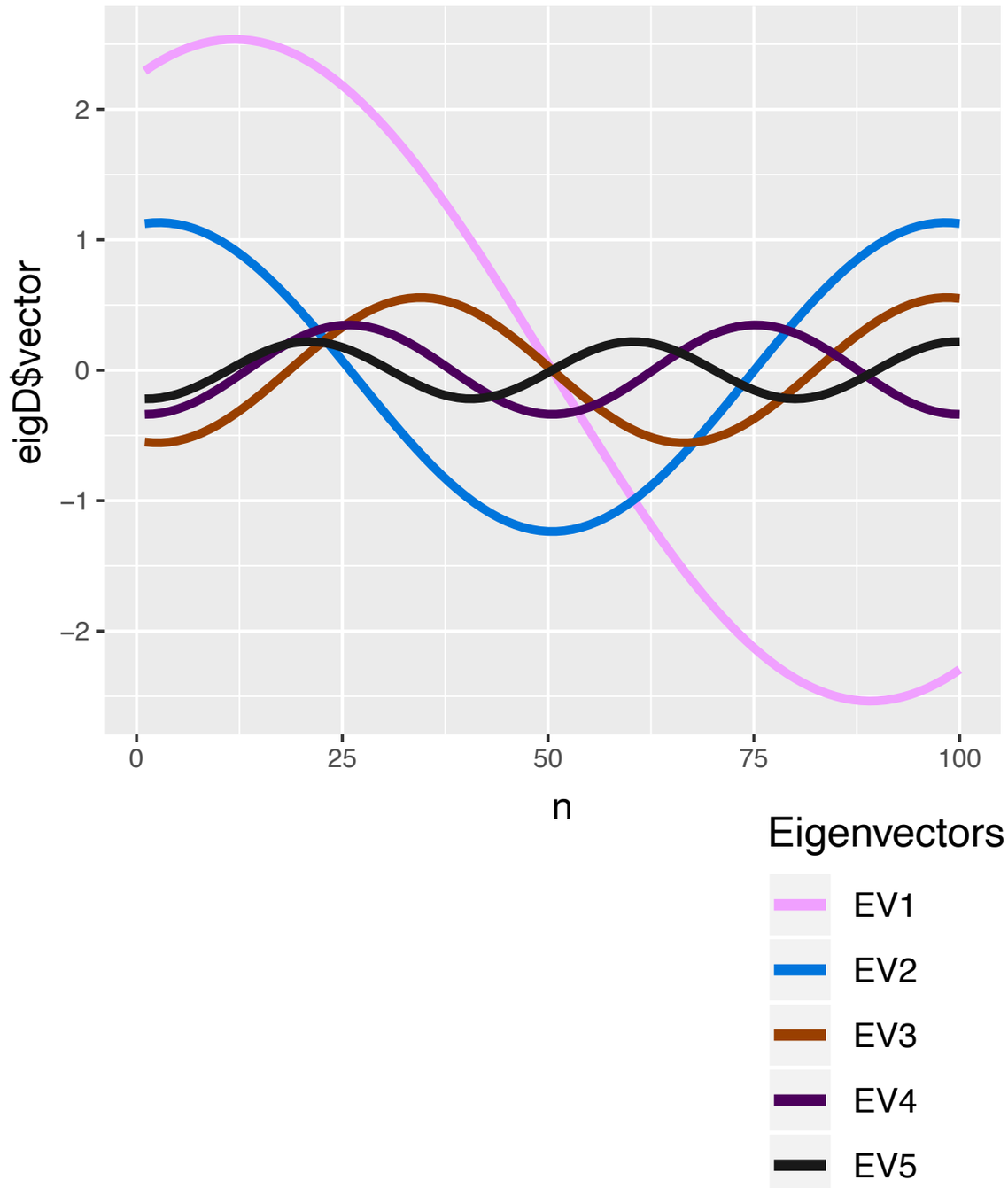






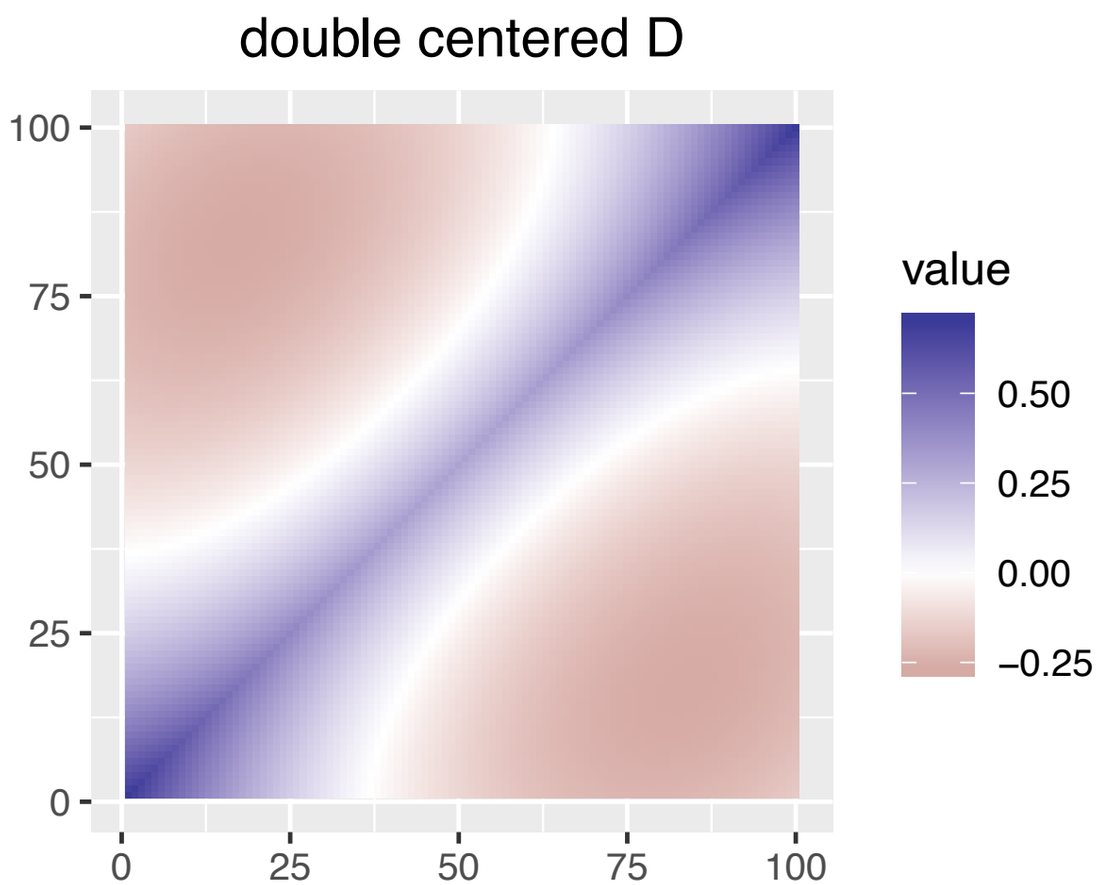
What is going on?






```
eigD ← eigen(double.center(D))
pairs(eigD$vectors[, 1:5])
```

For the mathematically inclined:



$$\frac{d^2}{dx^2} f(x) = \lim_{h \to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

$$\frac{d^2}{dx^2} f(x) = -k^2 f(x) \Leftrightarrow$$

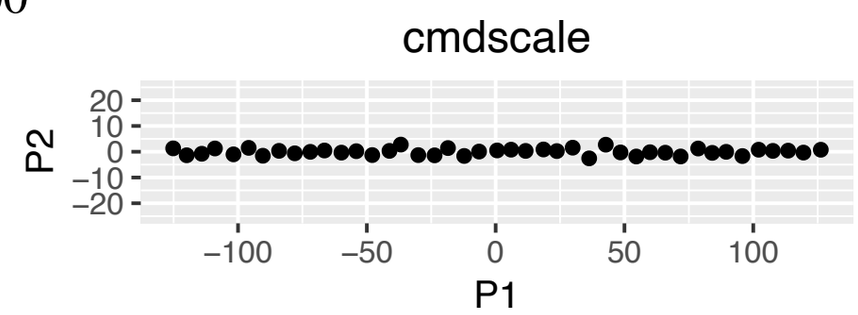
$$f(x) \propto e^{ikx} = \cos kx + i \sin kx$$

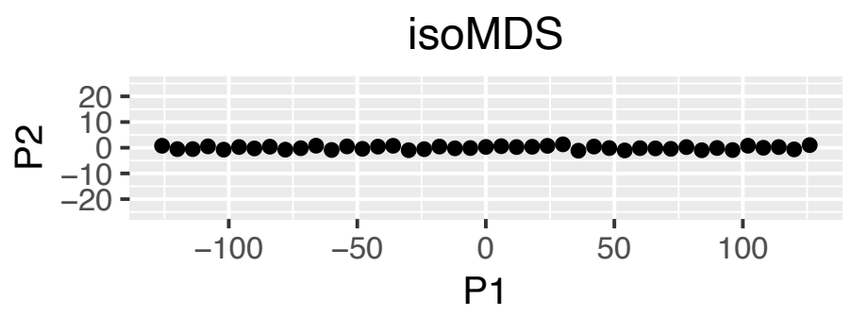
Multidimensional Scaling and Local Kernel Methods
Persi Diaconis, Sharad Goel and Susan Holmes
The Annals of Applied Statistics 2008

A straight line in $\mathbb{R}^{100}$

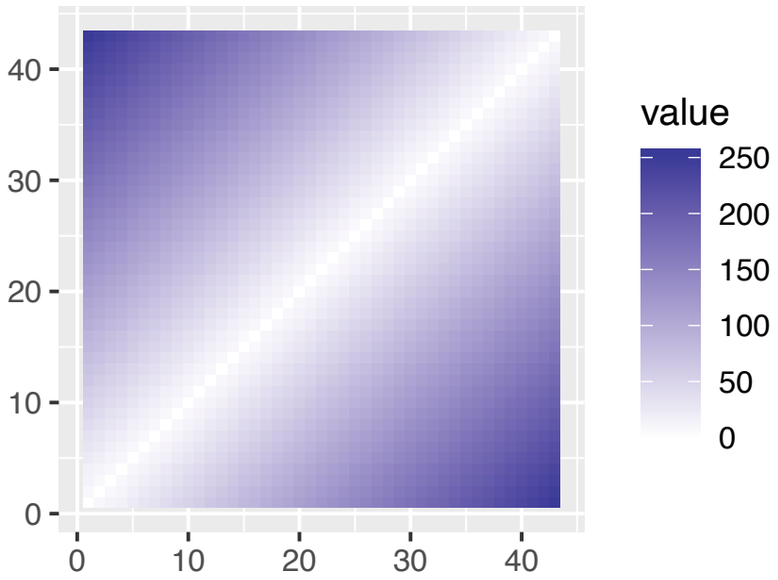
$$x = at + b$$

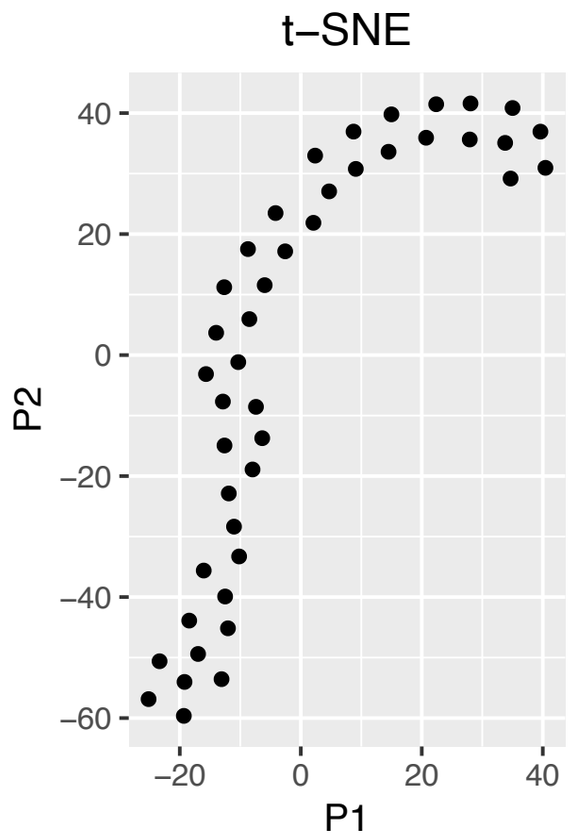
$$\text{for } t \in [t_1, t_2];\ a, b \in \mathbb{R}^{100}$$

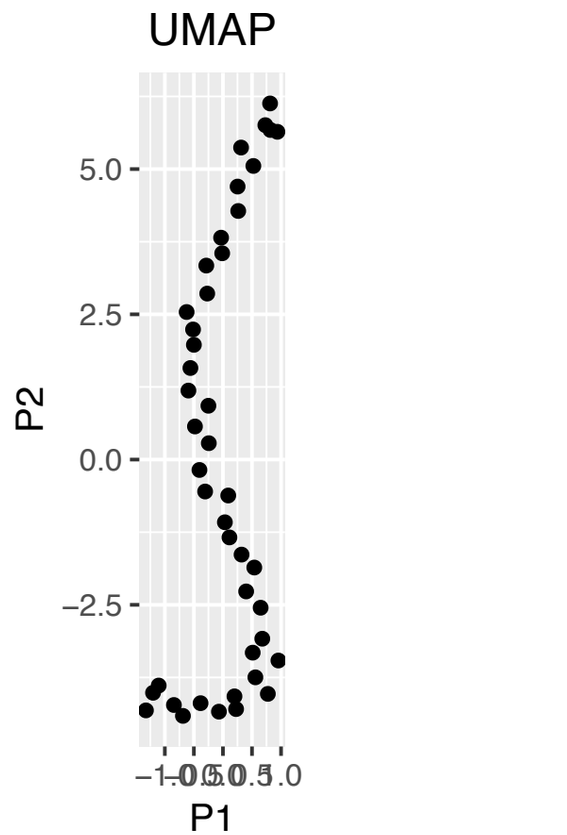




Distance matrix *D*



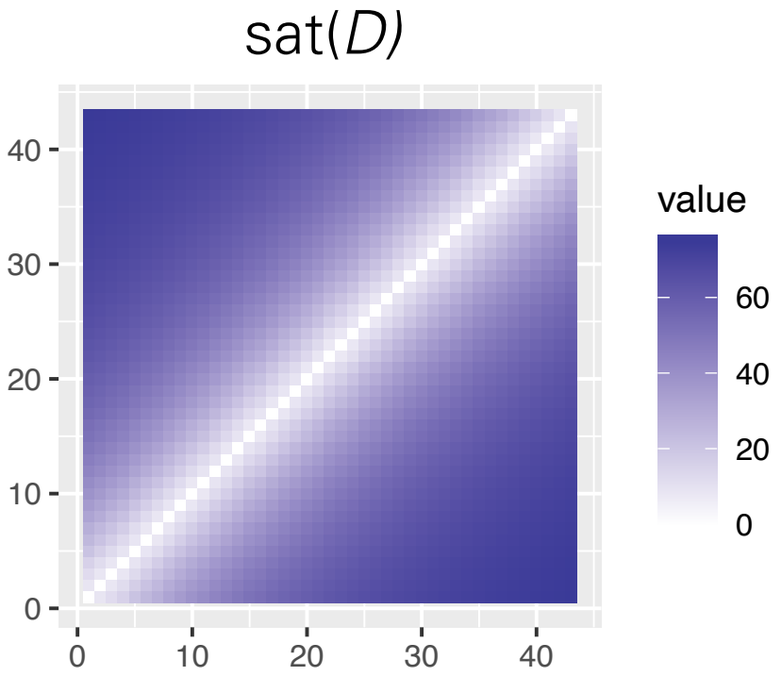


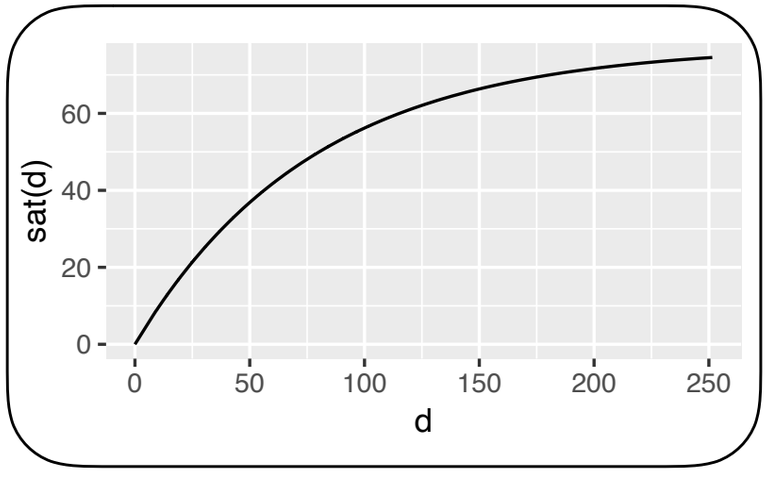


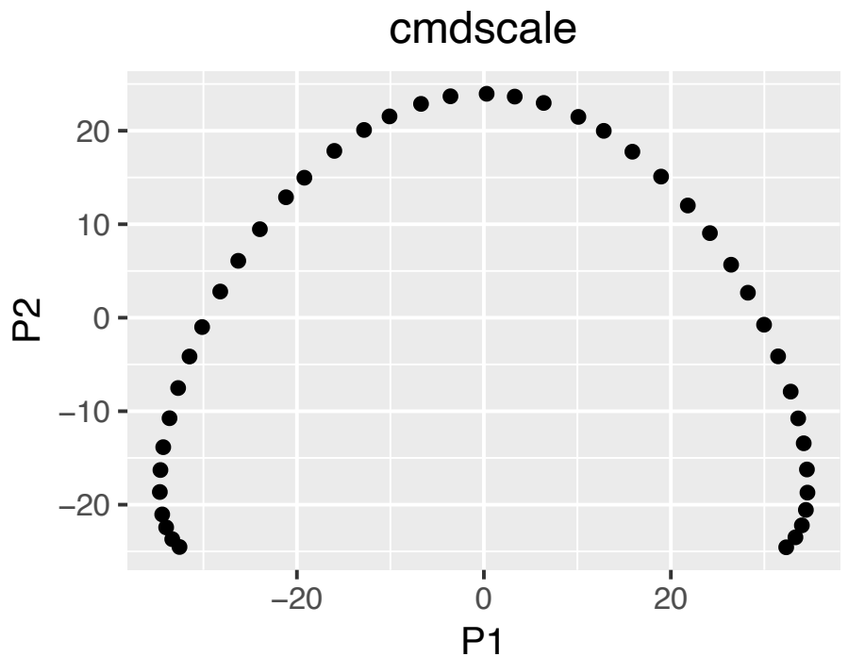
A straight line in $\mathbb{R}^{100}$ - with saturation of larger distances

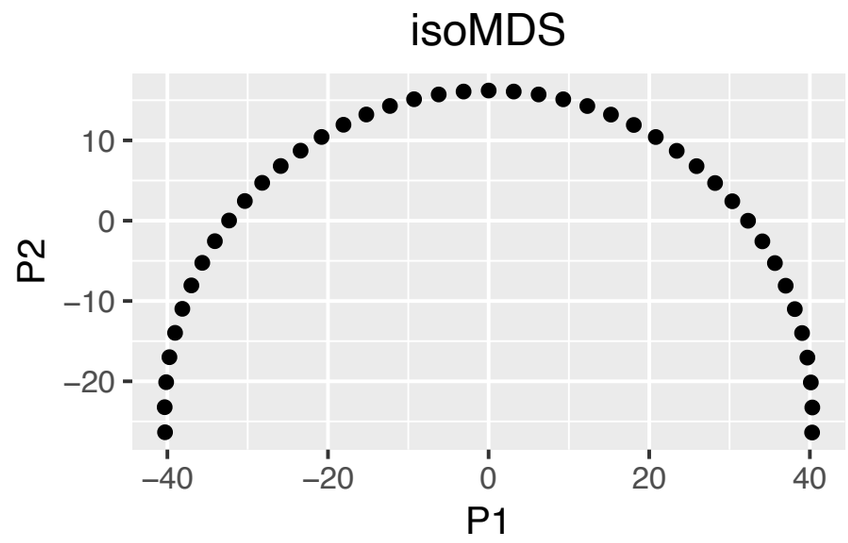
$$x = at + b$$

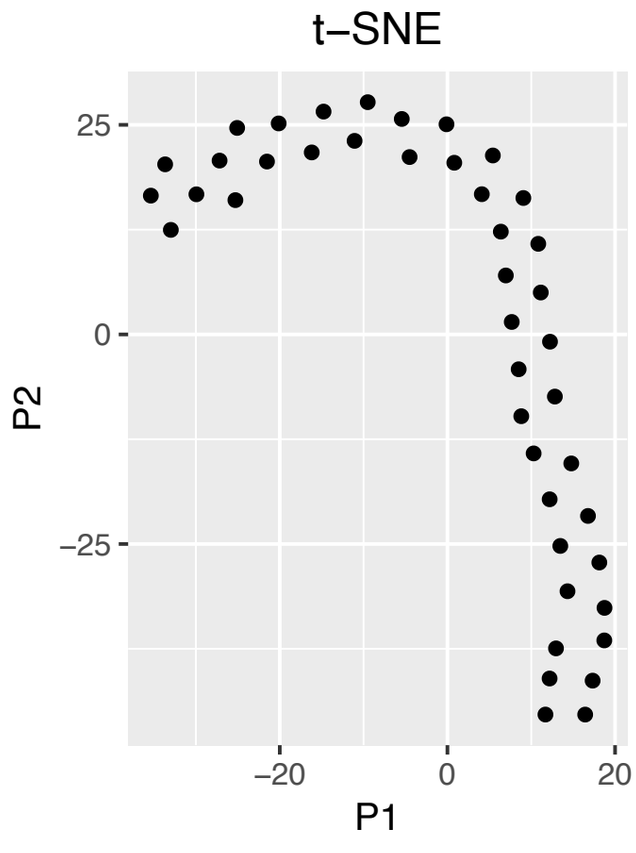
$$\text{for } t \in [t_1, t_2];\ a, b \in \mathbb{R}^{10}$$

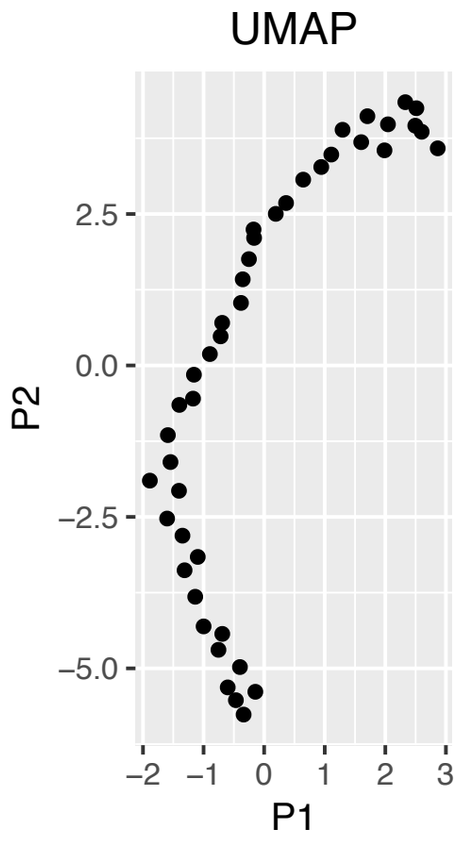




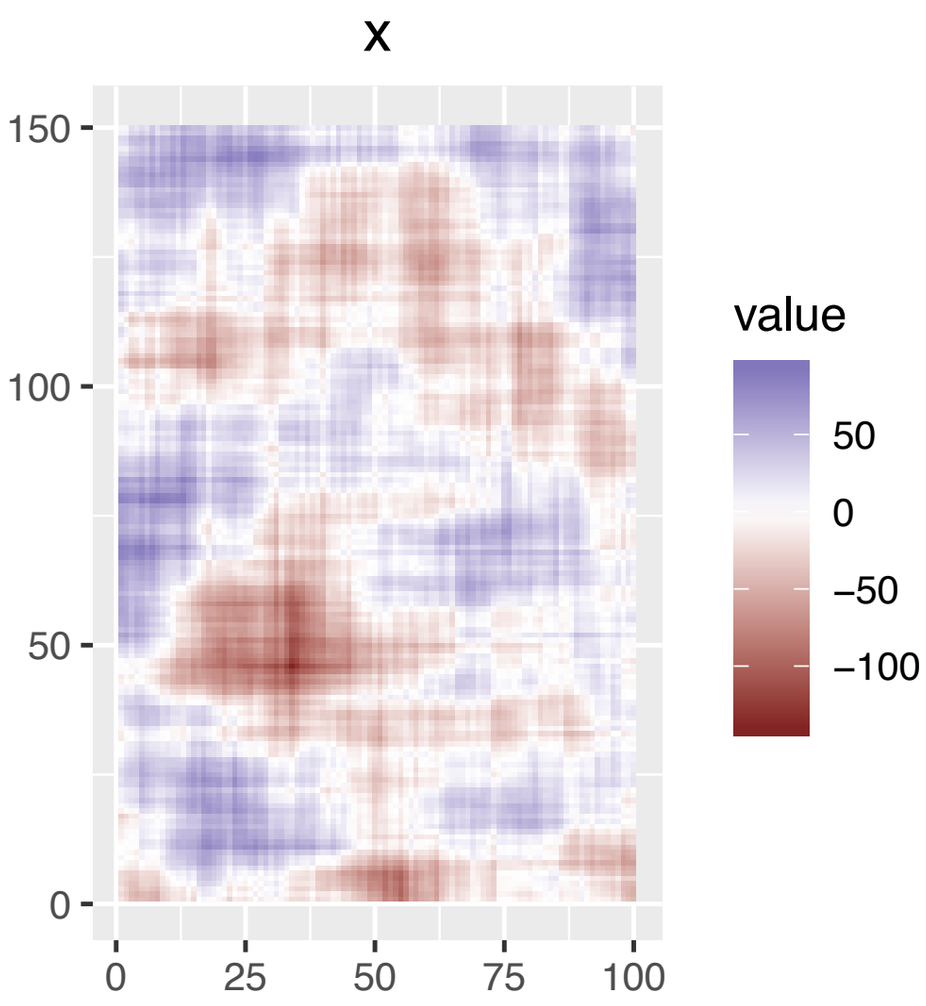


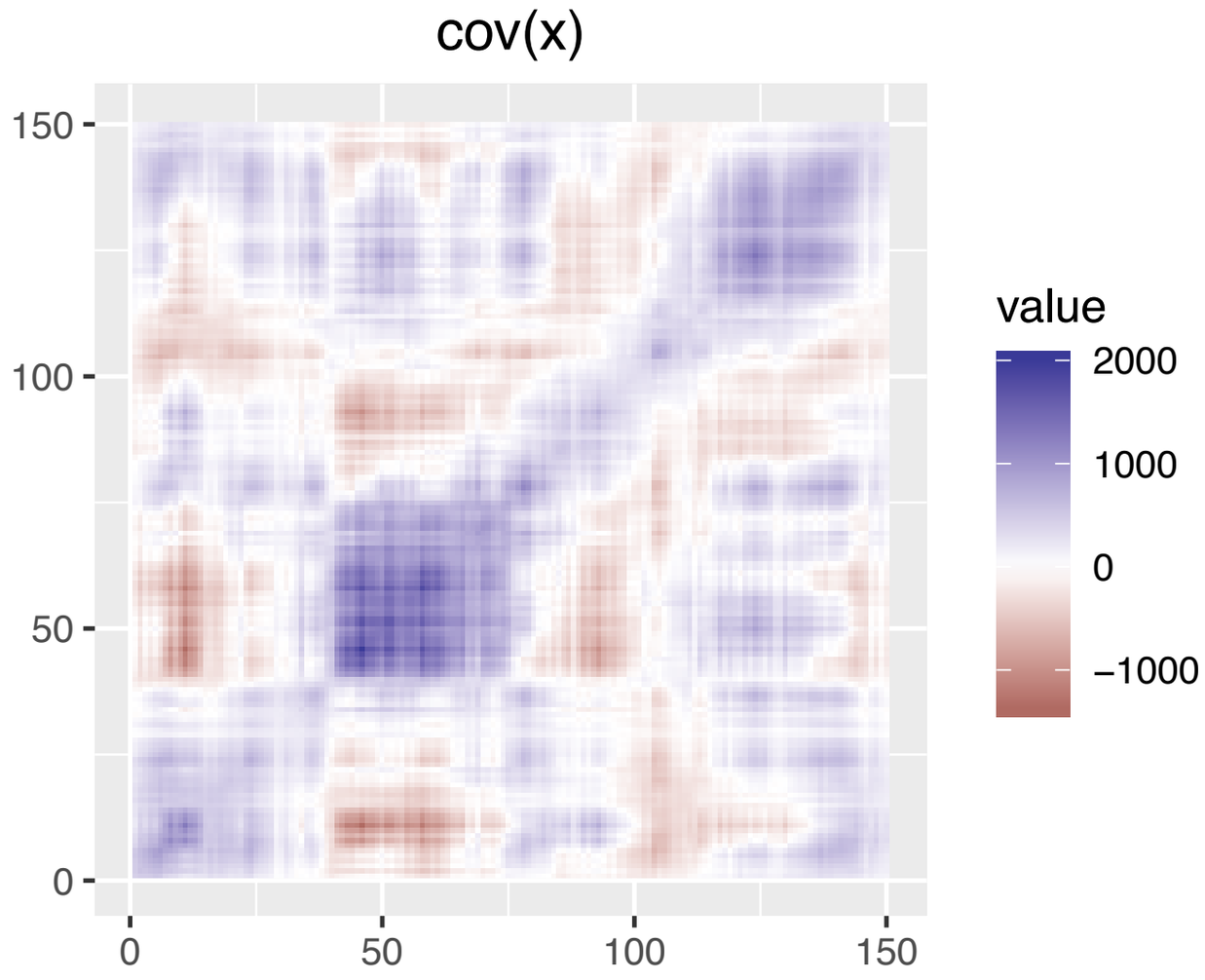




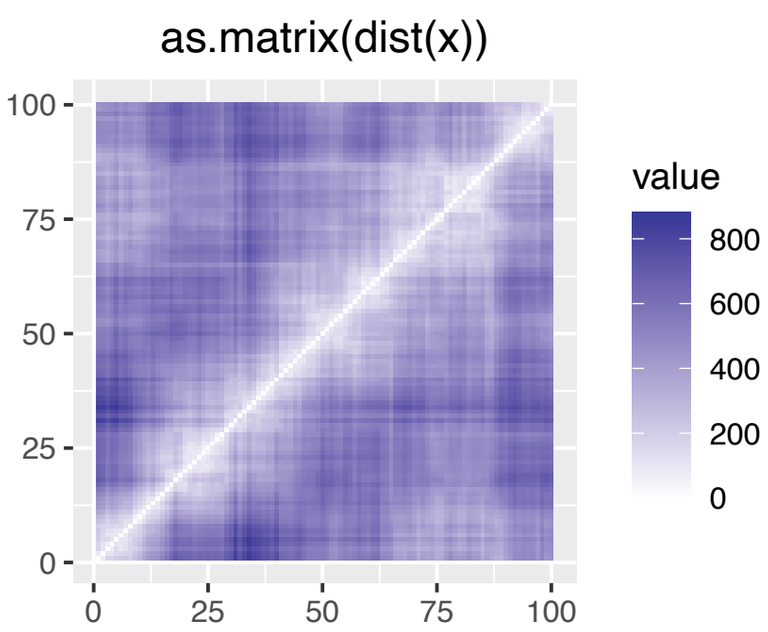


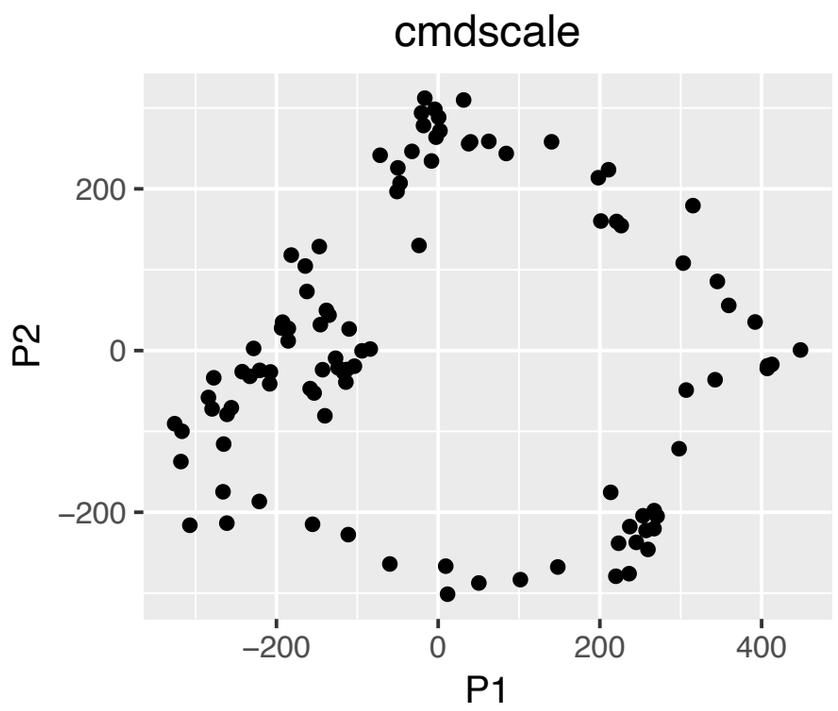
2D random field with spatial correlation

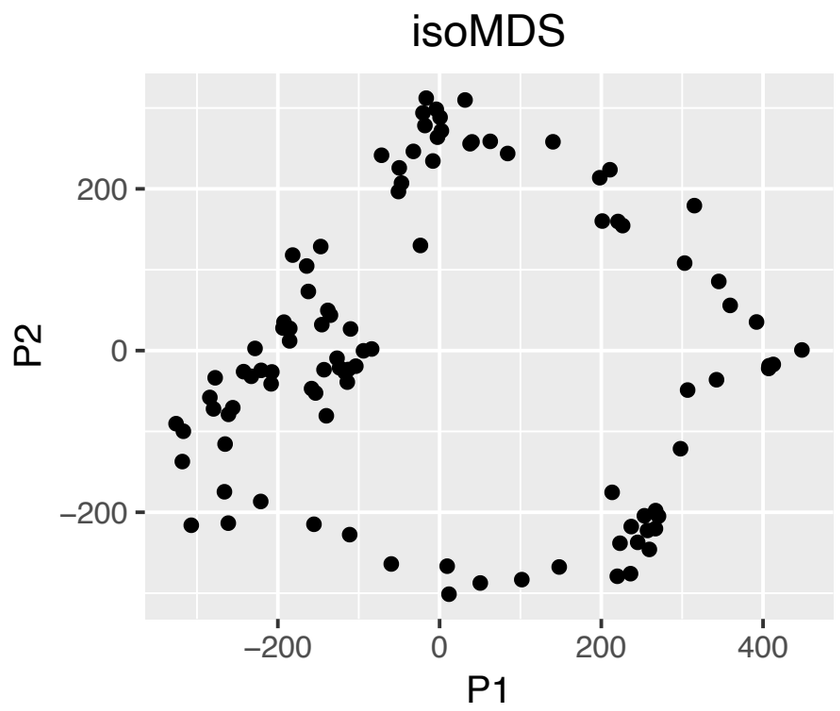


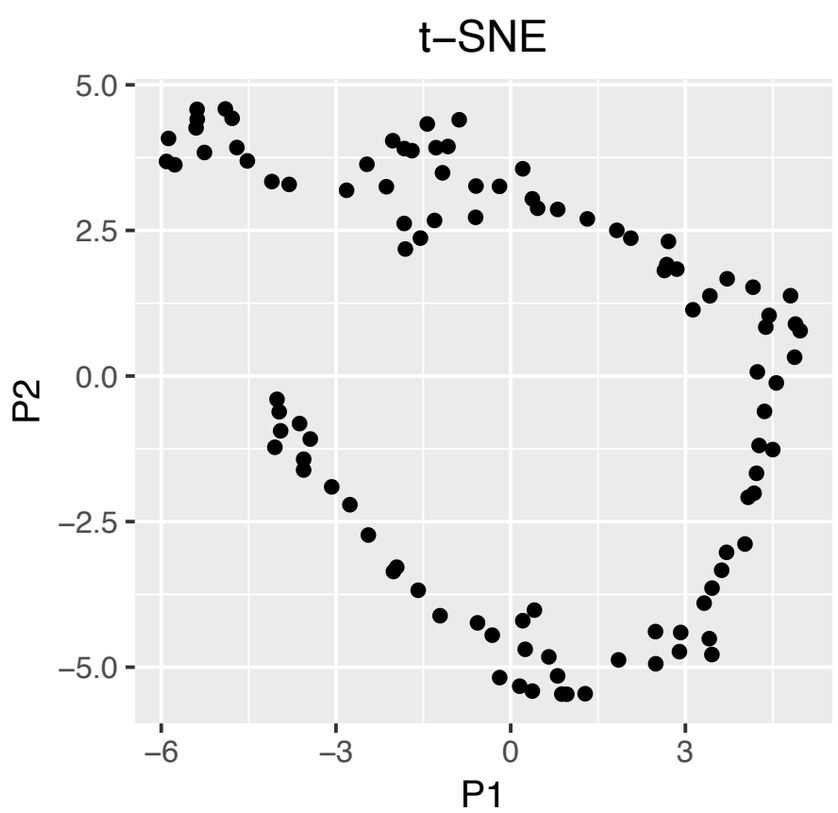


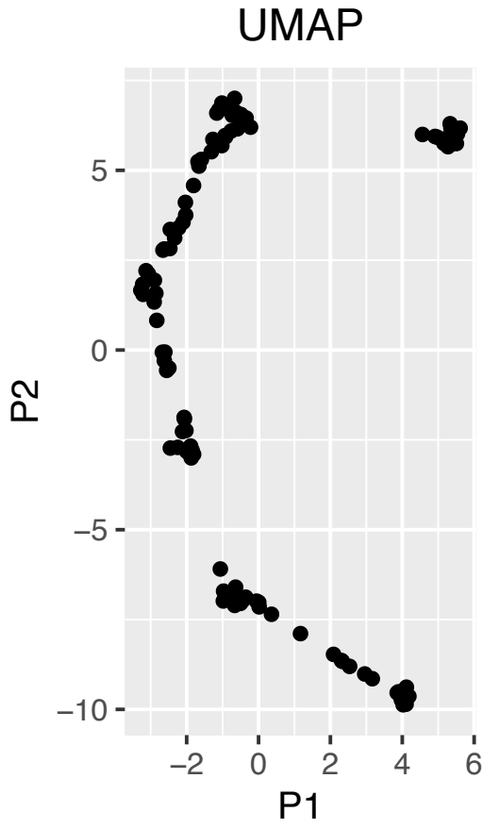
2D random field with spatial correlation



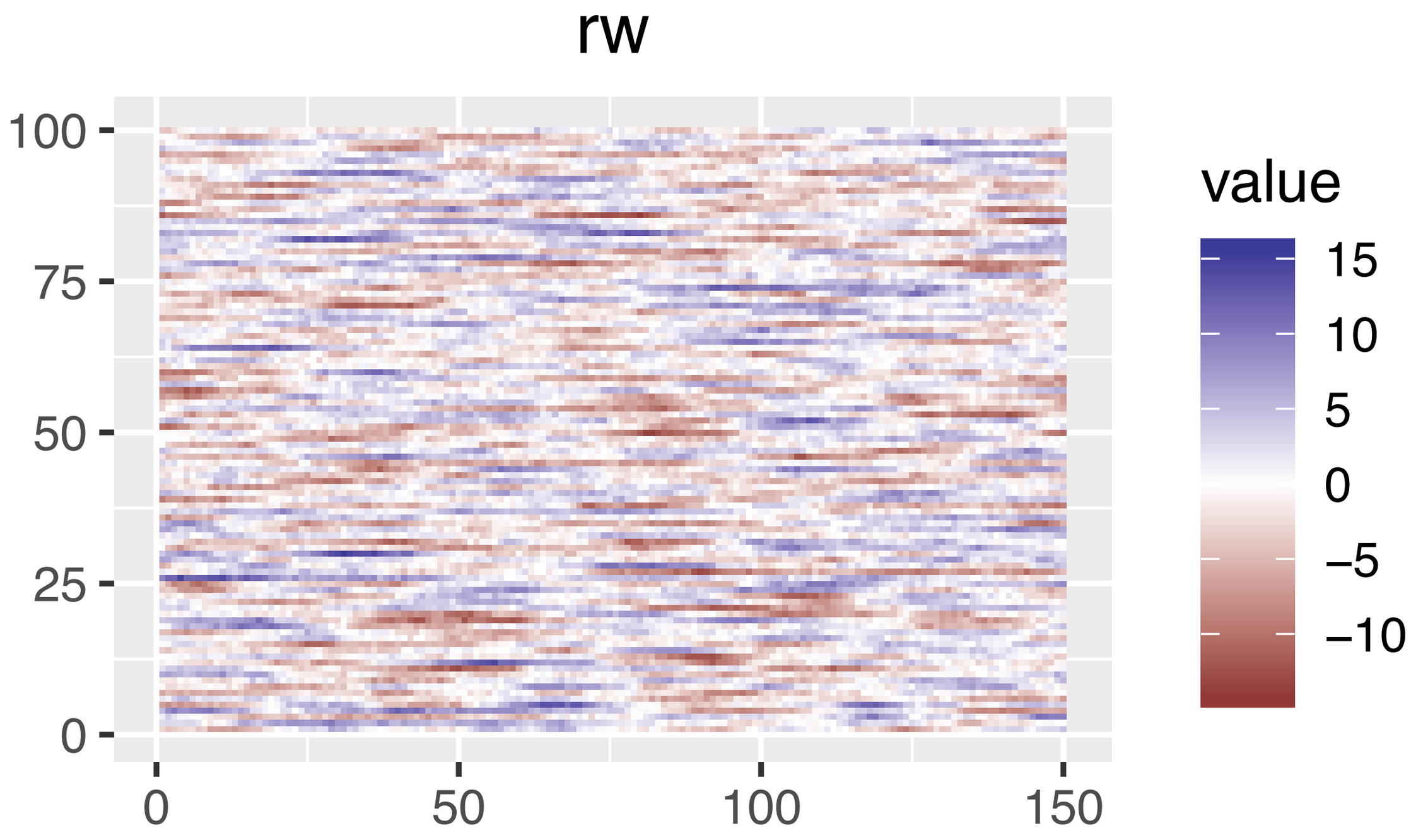




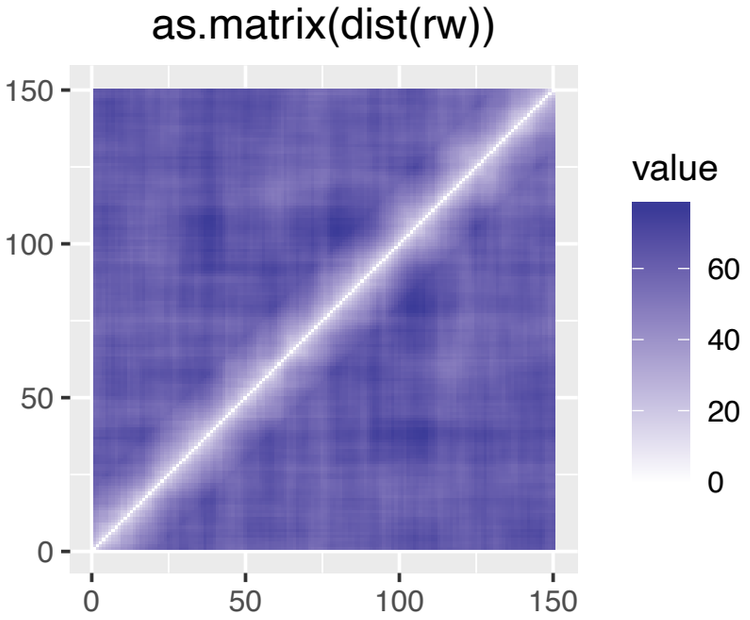


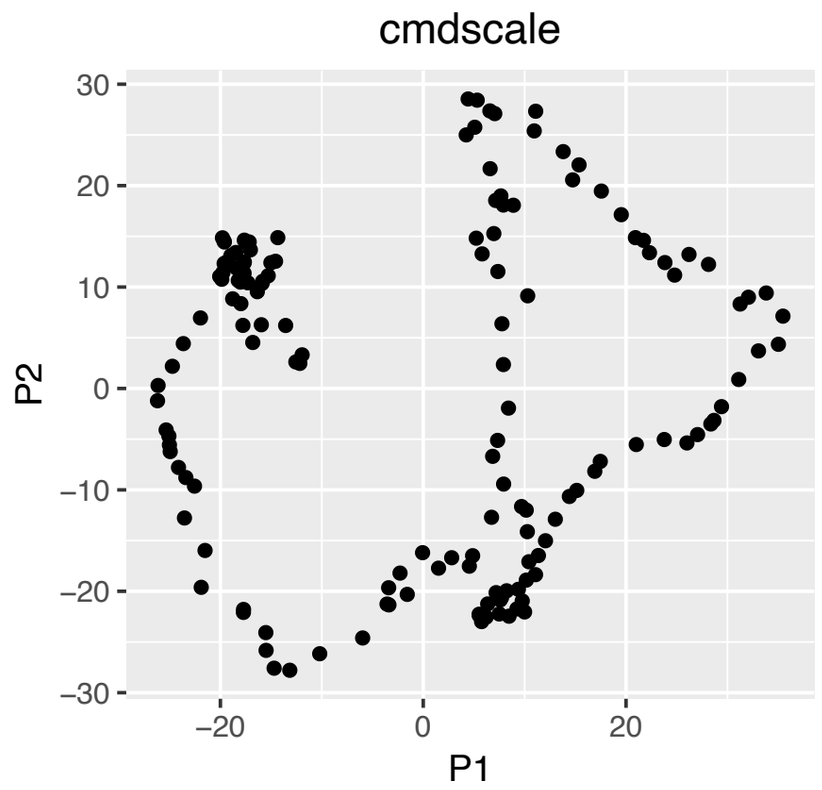


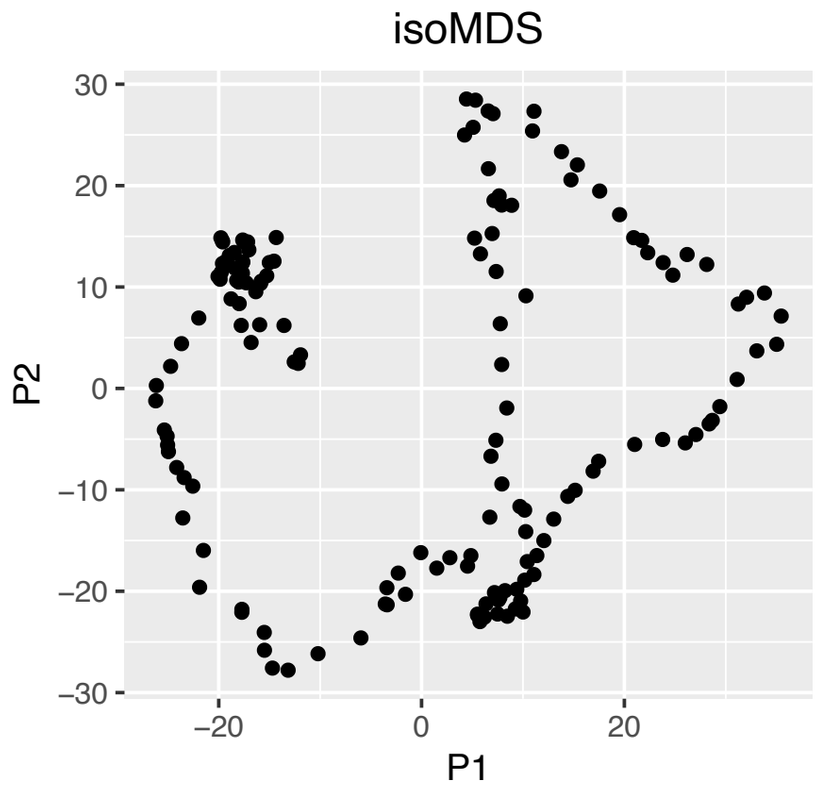
Matrix filled with random numbers with sequential correlation

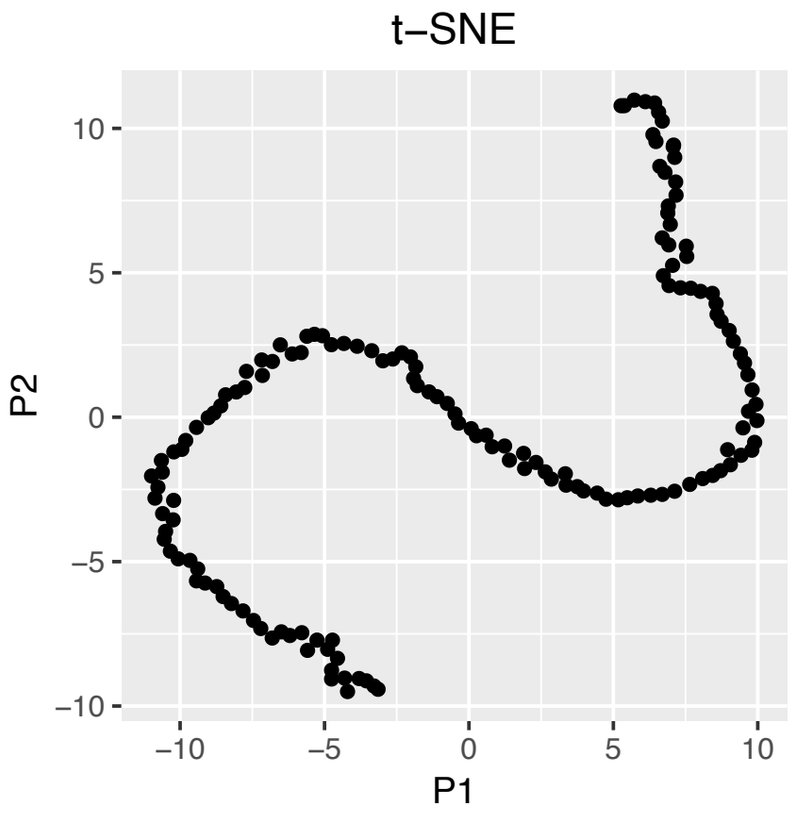


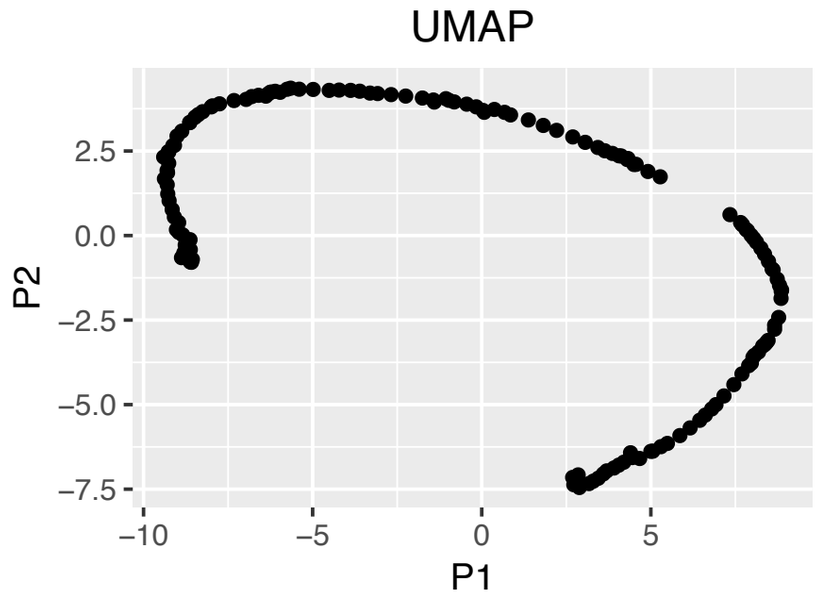
Matrix filled with random numbers with sequential correlation



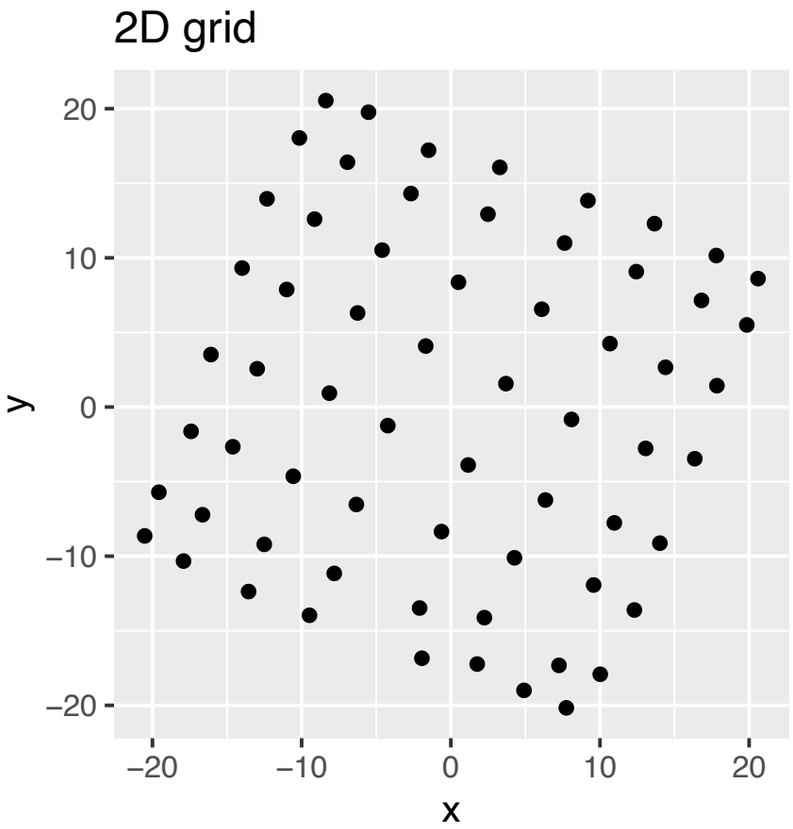


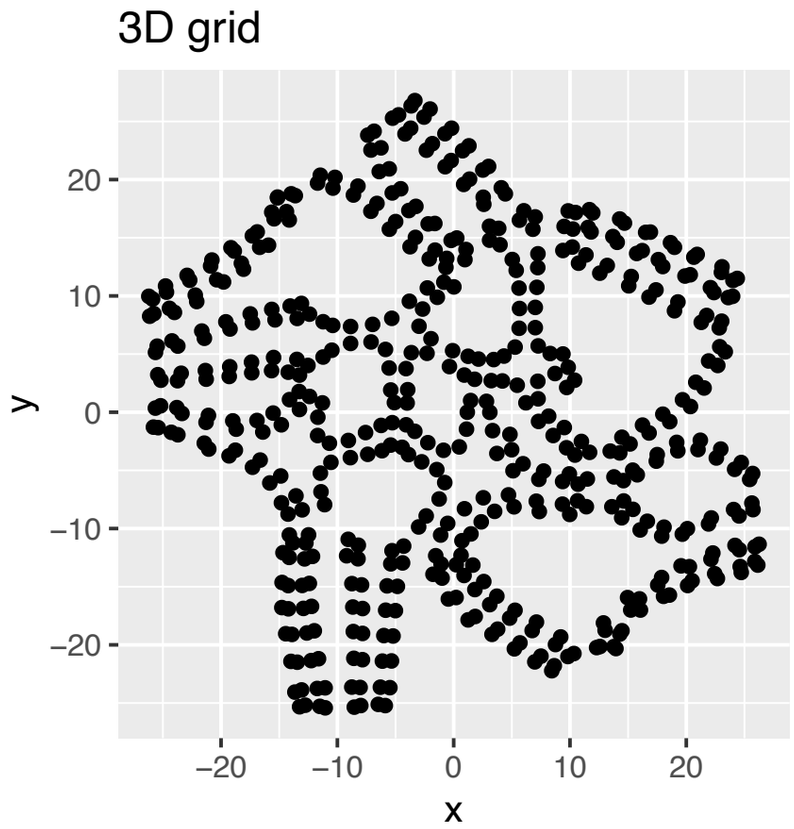


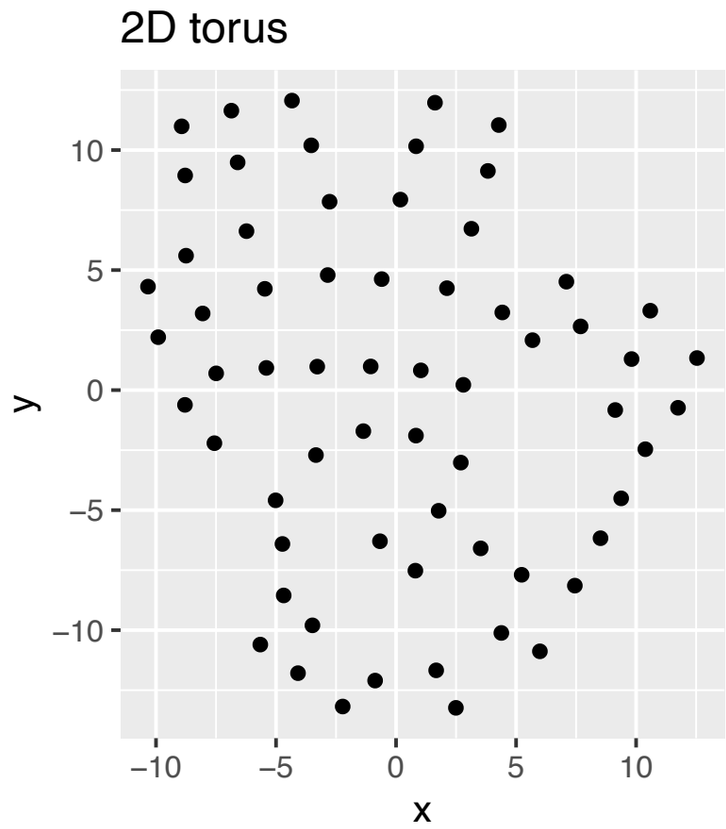


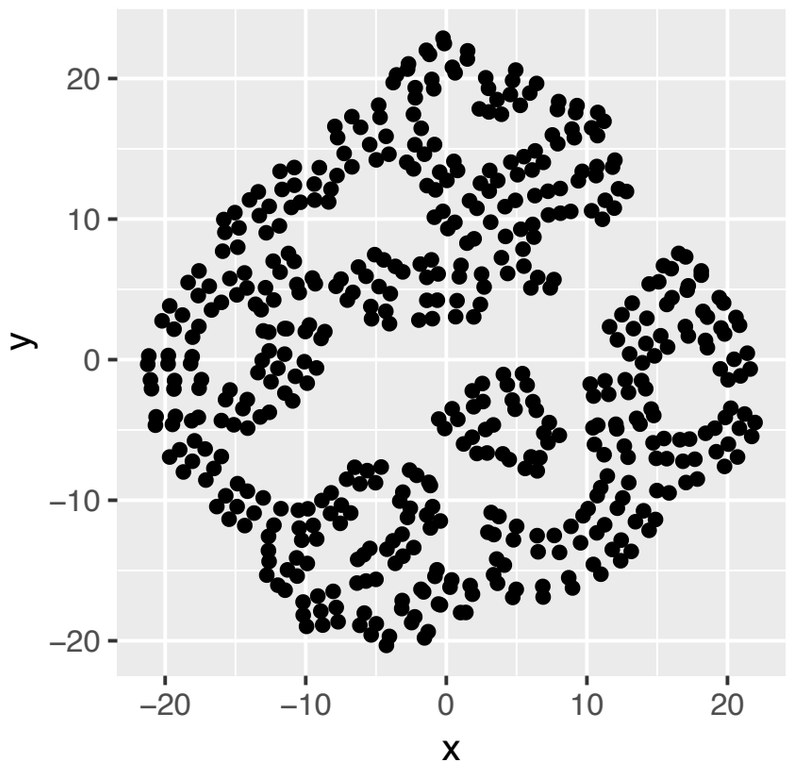


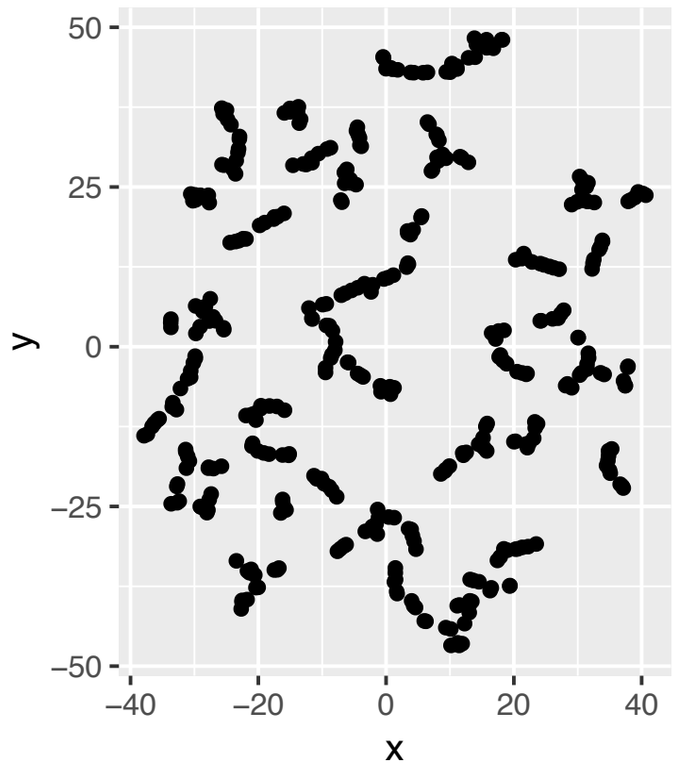
2D t-SNE on 'impossible' shapes

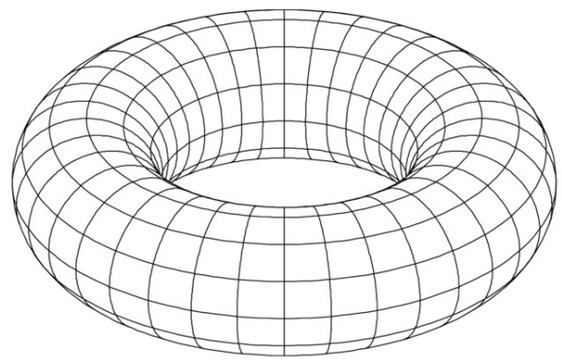




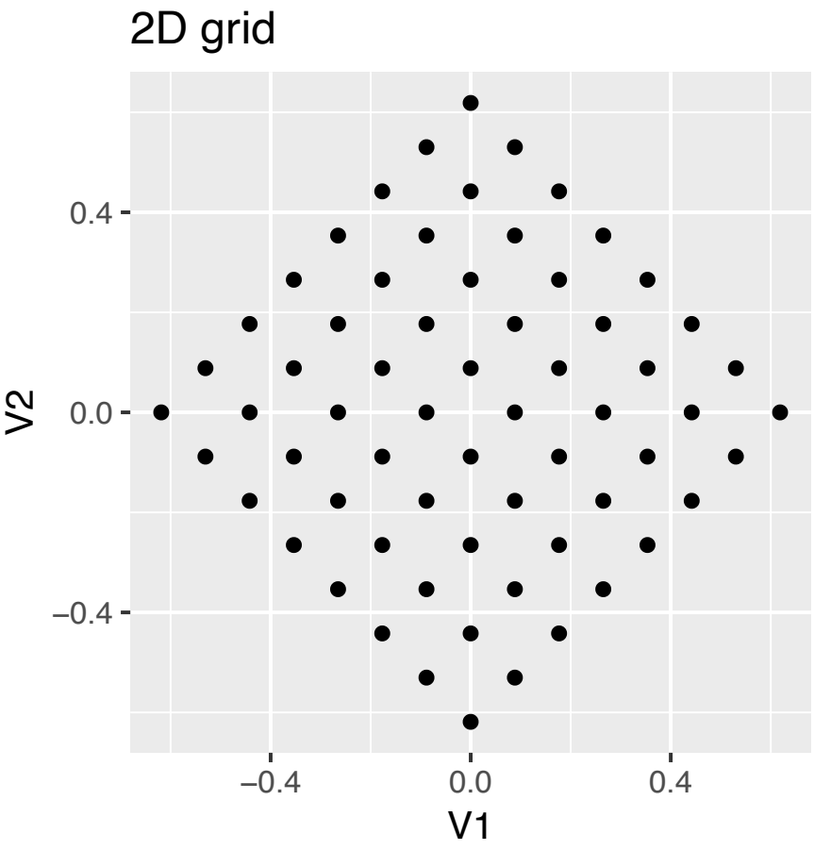


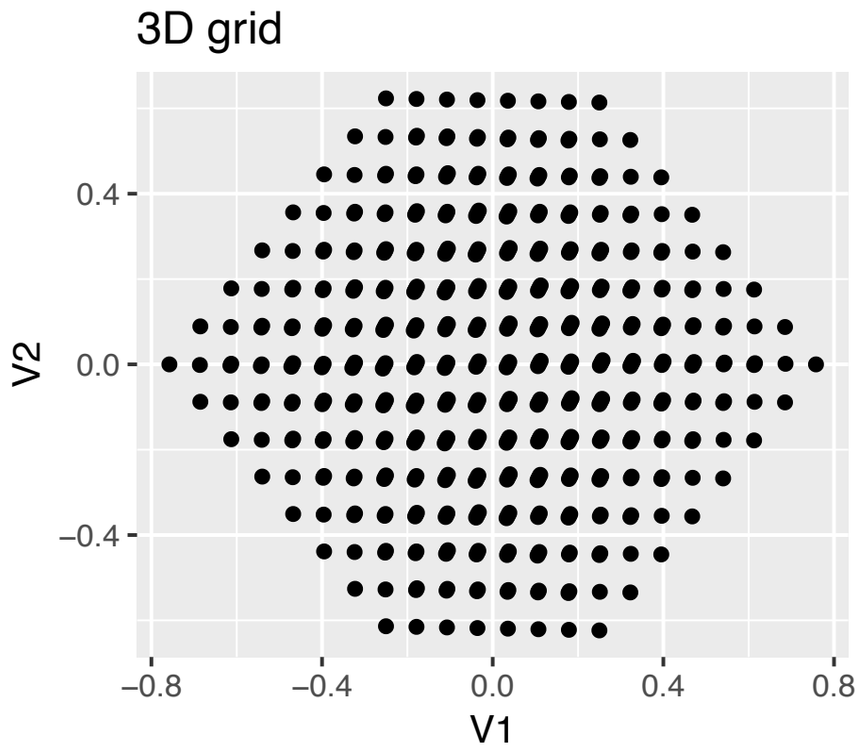


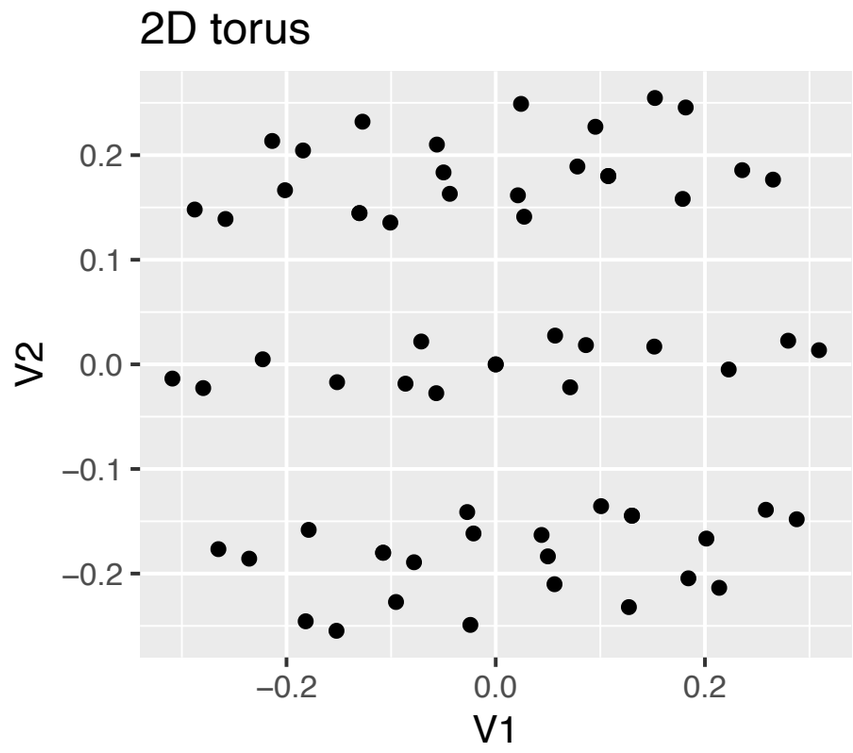


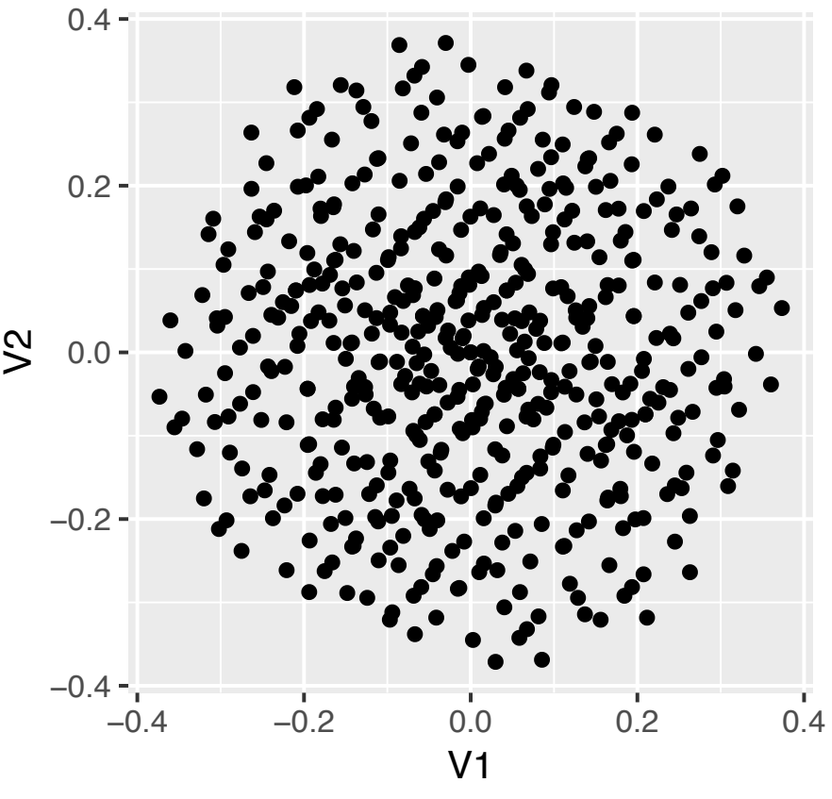


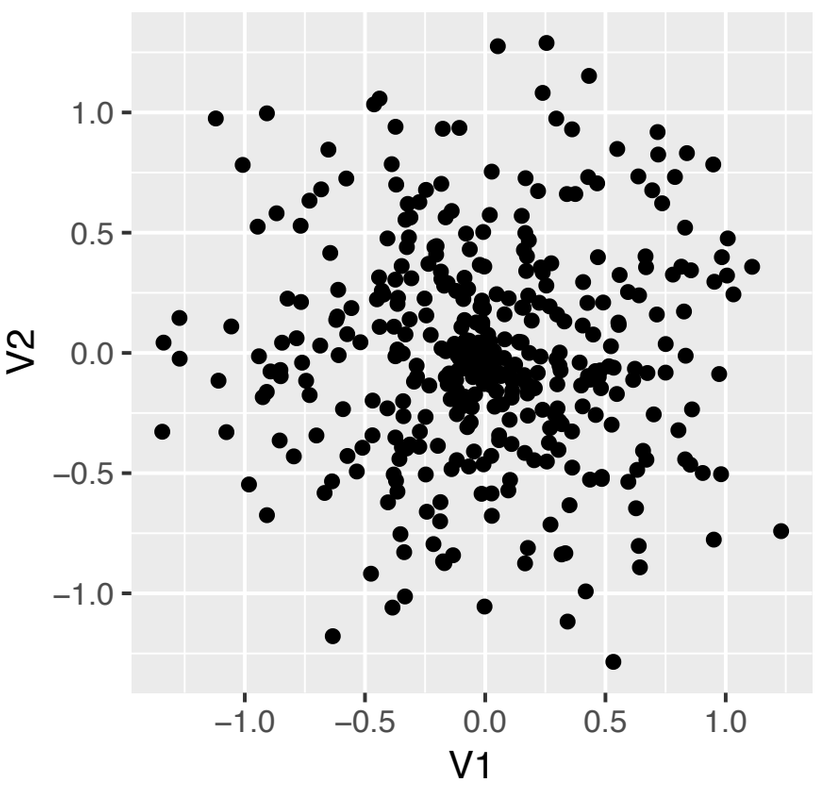
2D cmdscale on 'impossible' shapes

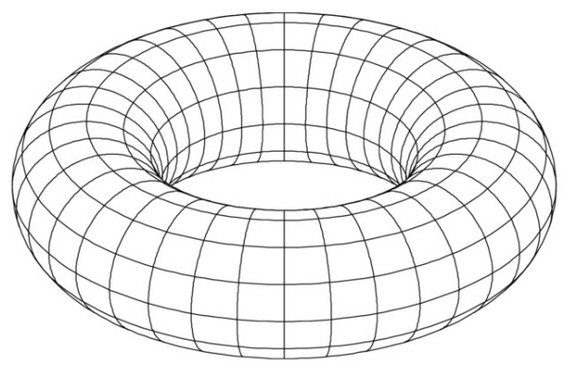












Take home messages

Embeddings of high-dimensional data into lower-dimensional space are useful

But they can create one-dimensional ("time-like") patterns that have little to do with the data-generating process

Sometimes, a faithful embedding is mathematically impossible

High-dimensional geometry is weird

Be aware