Machine Learning for biologie

V. Monbet

UFR de Mathématiques
Université de Rennes 1

Outline

Outline

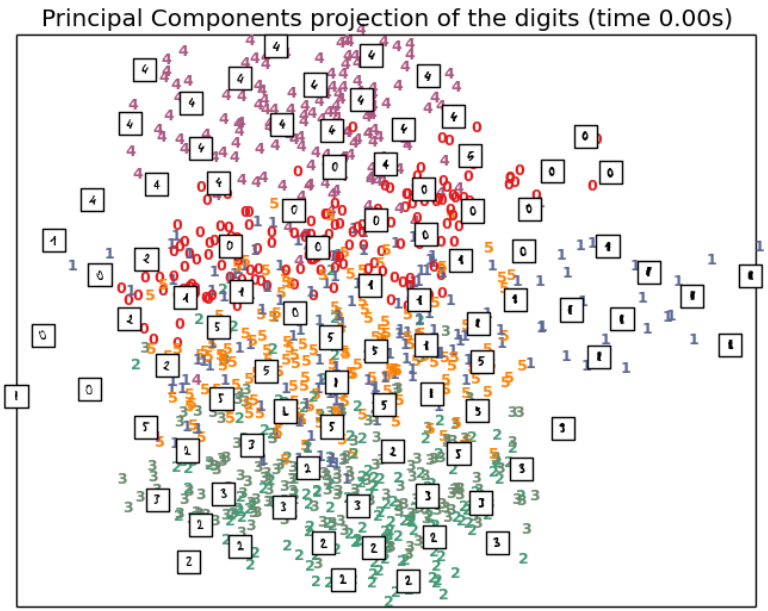
Outline

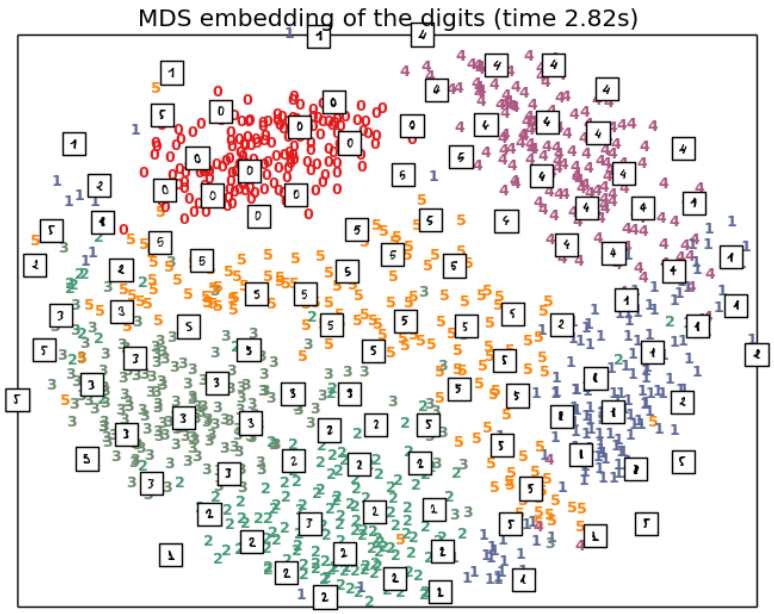
Other methods: multidimensional scaling (MDS)

- There are other methods for dimension reduction based on the searches of orthogonal/independent basis or factors.
- MDS
  Given a matrix of distances between the observations D between n individuals, MDS searches for euclidean coordinates $X_{mds} \in \mathbb{R}^q$ such that if D_{ij} is small, $X_{mds}(i)$ is close to $X_{mds}(j)$.
  In practice, it is obtained by the spectral analysis of the matrix B such that $b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2)$
- When D is defined by the euclidean distances, MDS is equivalent to PCA.

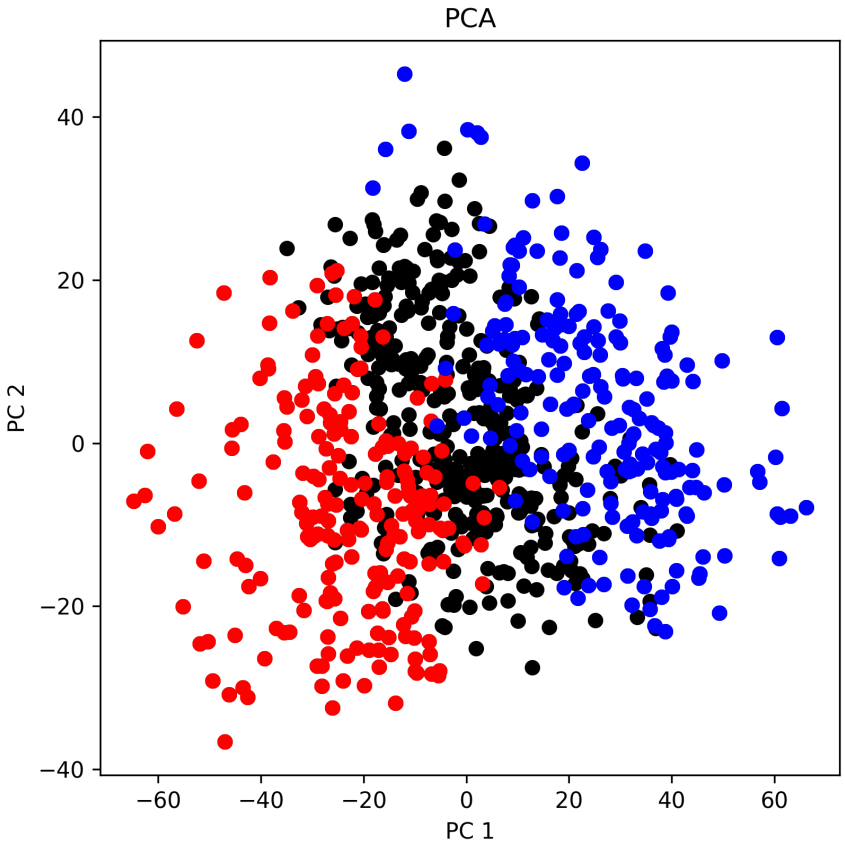
MDS, digits data

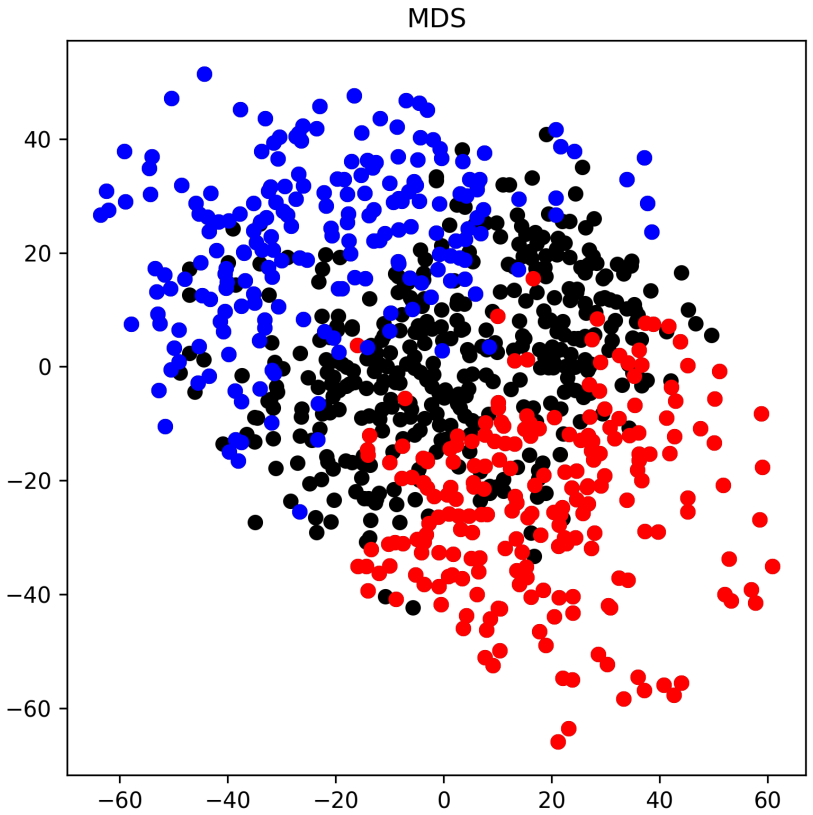
- MNIST digits, PCA (left) and MDS (right)





MDS, SST data

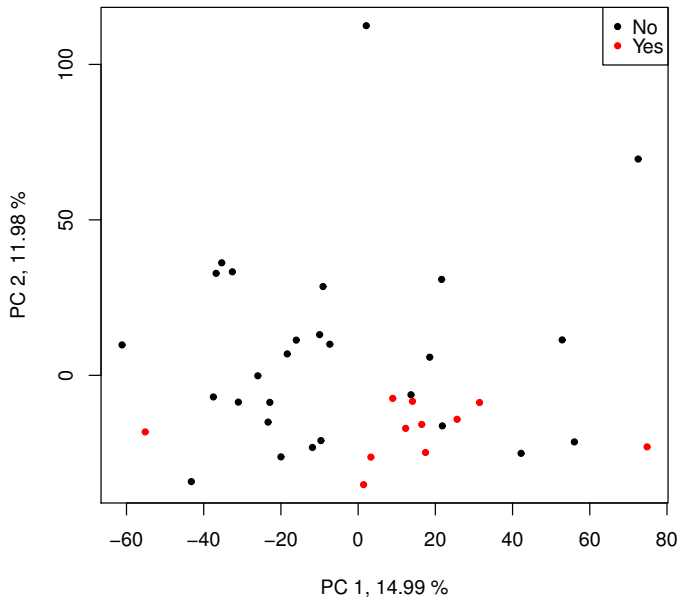

PCA
PC 2
PC 1
40
20
0
−20
−40
−60
−40
−20
0
20
40
60


MDS
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60

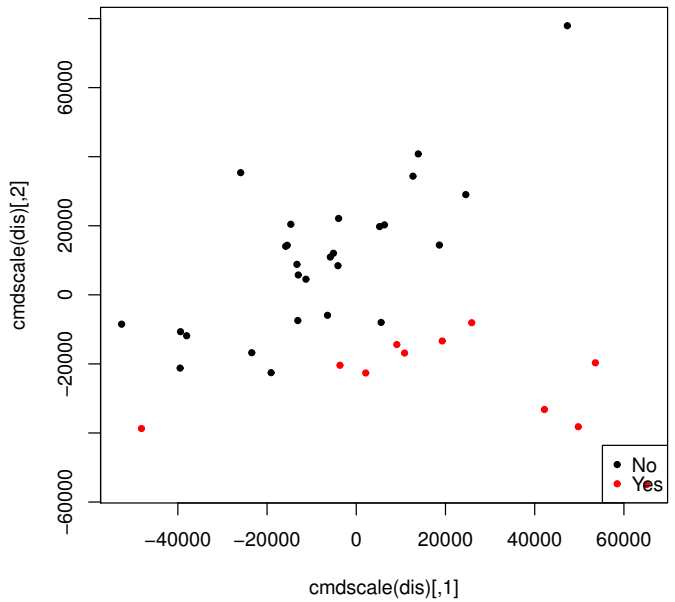
MDS

- Leukemia Gene Expression data

PCA



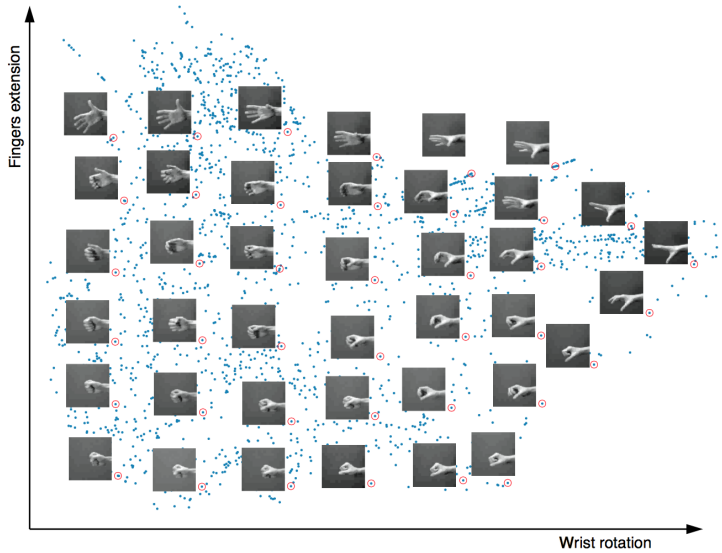
MDS



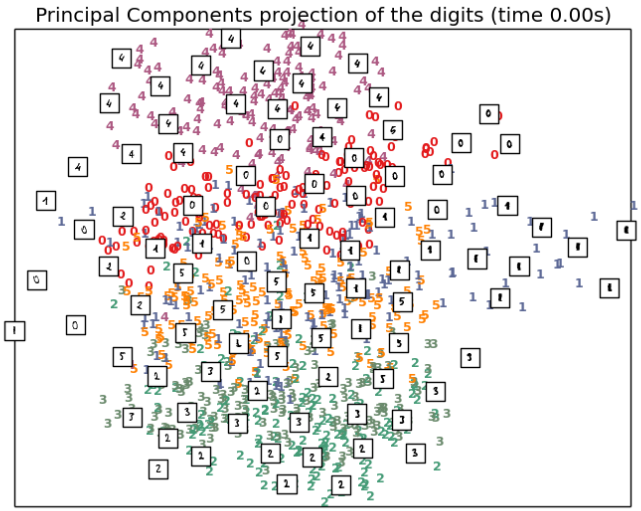
Other methods : Isomap

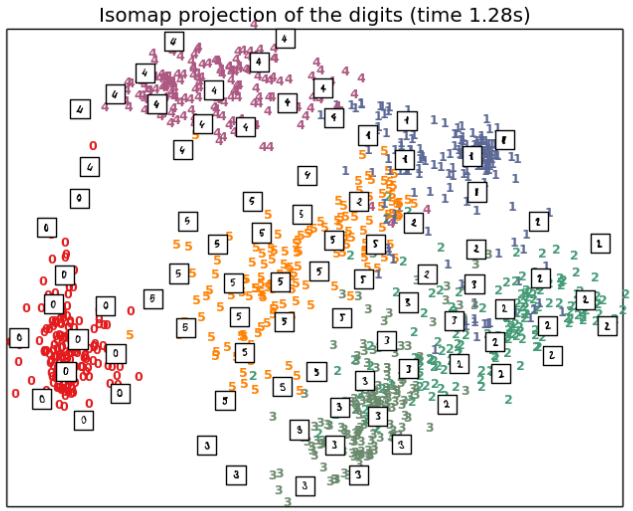
- Isomap is a version of MDS based on a distance on a neighborhood graph.
- d_{ij} is the number of edges between i and j
- Isomap is mostly used for image dimension reduction because it is able to well capture the "motions".

  Ref : Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500), 2319-2323.

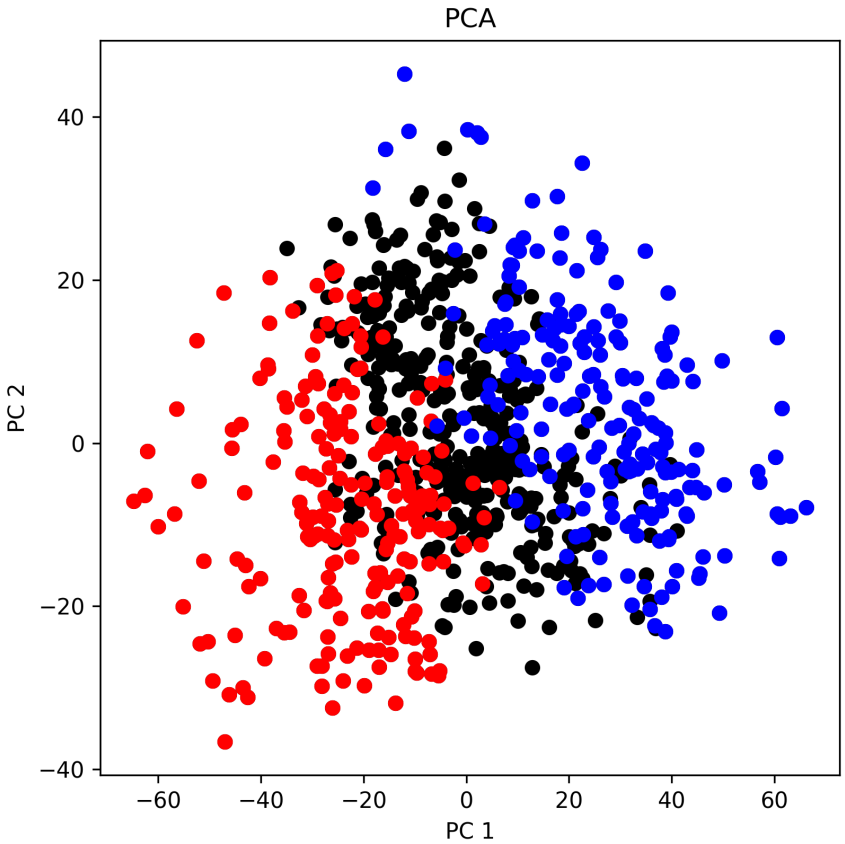


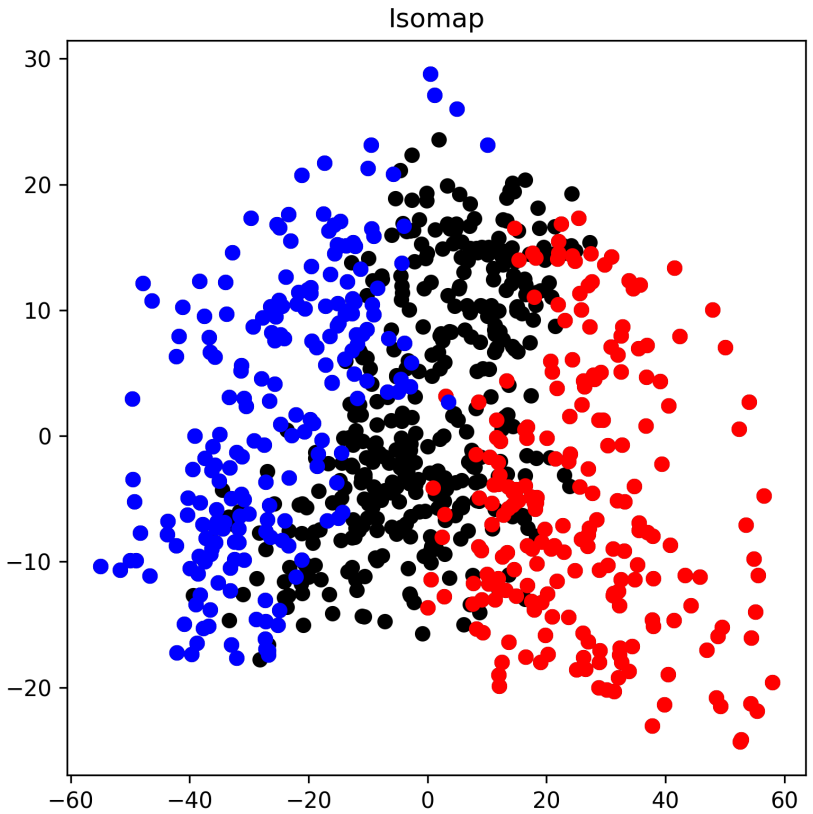
- MNIST digits, PCA (left) and Isomap (right)





Other methods : Isomap





Outline

Stochastic neighbor embedding

- SNE is a method of dimension reduction.
- SNE leads to a representation of the data in a low dimension space (typically 2). In this space, two samples with a high similarity will be close to each other.
- SNE is different from PCA because the similarity is not measured through correlation.
- In the original space of the data, the similarity between two samples x_j and x_i is defined by the conditional density of x_j given x_i

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i\right)}{\sum_{k\neq j}\exp\left(-\|x_i - x_k\|^2/2\sigma_i\right)}$$

  where σ_i is a parameter to be chosen. By convention, $p_{i|i} = 0$.
- In the reduced space, the similarity is measured by a conditional density as well

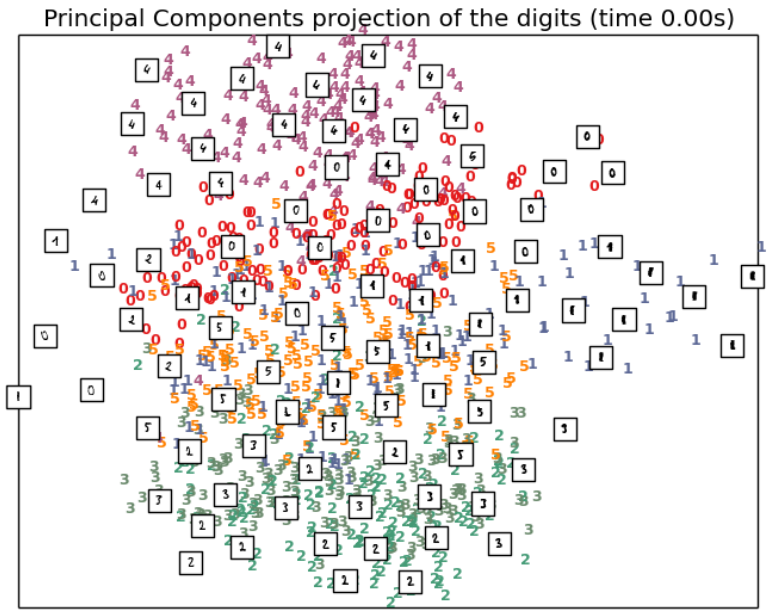
$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq j}\exp\left(-\|y_i - y_k\|^2\right)}$$

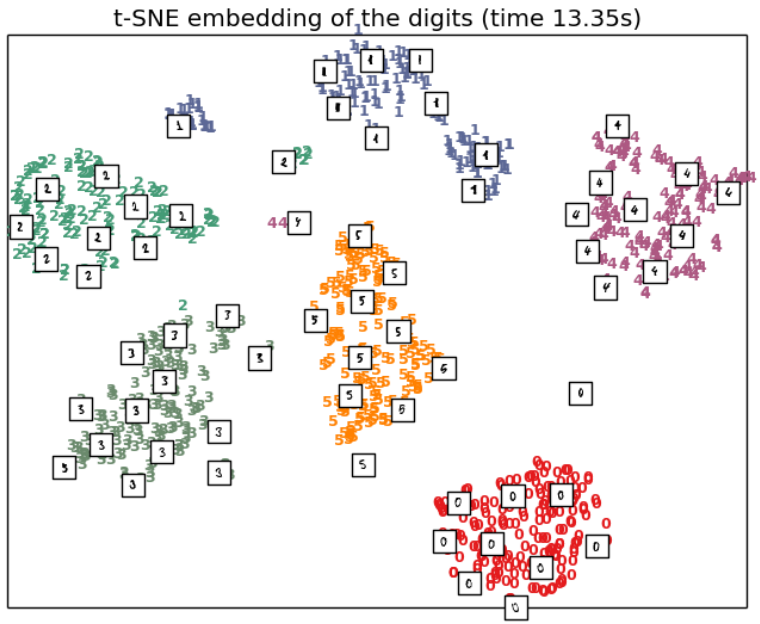
  with $q_{i|i} = 0$ by convention.

Ref : Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(2579-2605), 85.
Animations :
https://www.oreilly.com/learning/an-illustrated-introduction-to-the-t-sne-algorithm

t-distributed stochastic neighbor embedding (t-SNE)

- Digits MNIST, PCA (left) vs t-SNE (right)

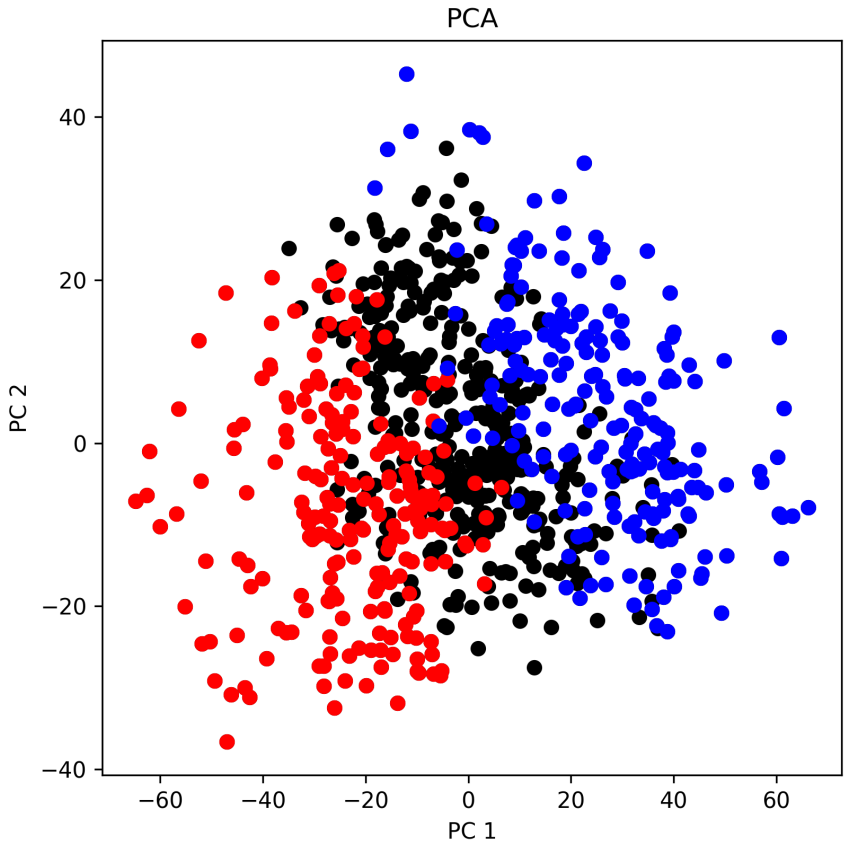


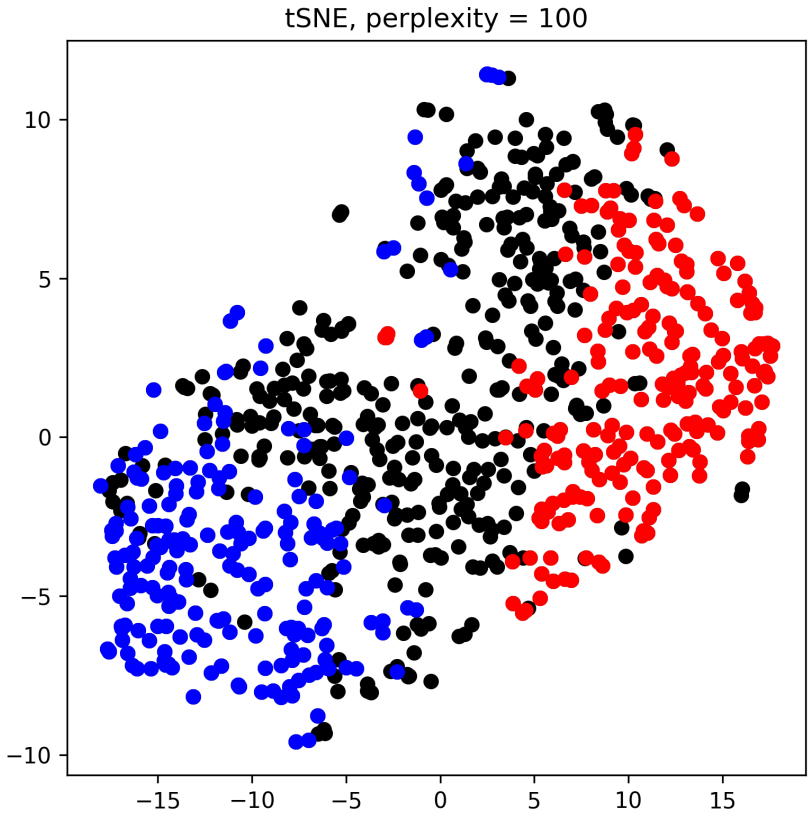


- t-SNE allows to better gather similar observations.
- But, a key parameters σ_i may be hard to choose and computation time is larger.

- Number of citations of the seminal paper : 2600

Other methods: t-SNE

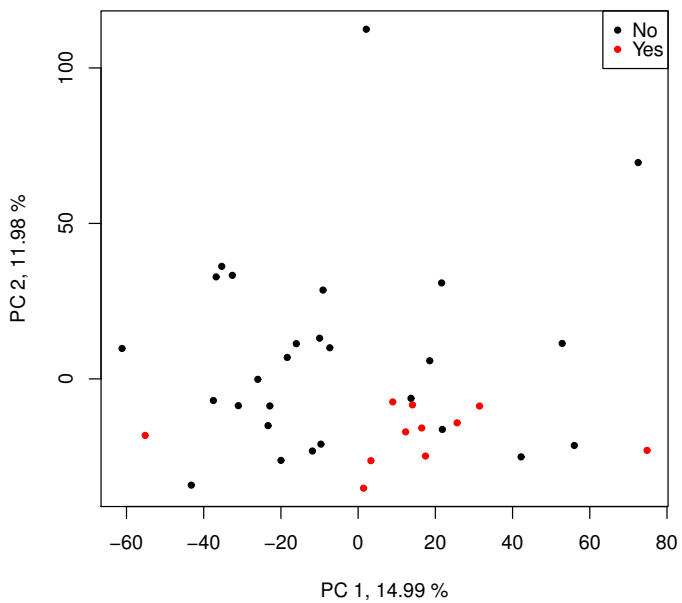




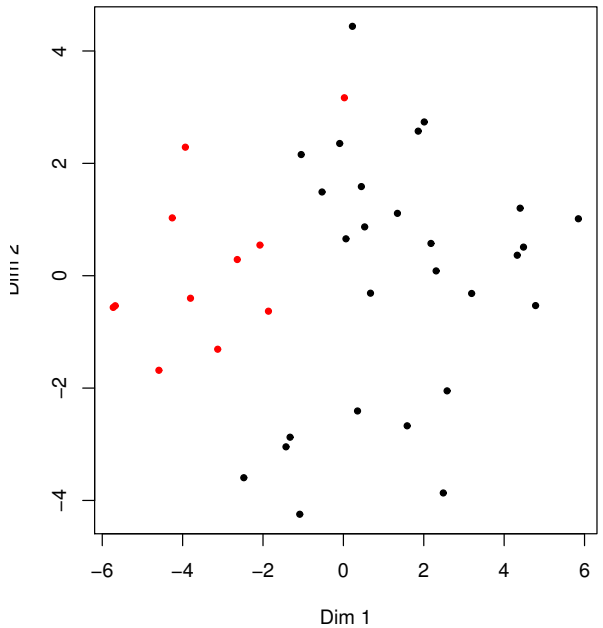
t-distributed stochastic neighbor embedding (t-SNE)

- Leukemia Gene Expression data

PCA



tsne



σ is chosen such that for each (i, j) the conditional probability $p_{j|i}$ is close to the chosen perplexity. Here perplexity=10.

Outline

Outline

Outline

Outline

Outline

Outline

Outline

Outline

Outline

Outline