Dimension reduction methods based on FINE algorithm for clustering patients from flow cytometry data.

Walid Laziri

Anne Gégout-Petit, Frédéric Allemand, Sophie Wantz-Mézières
INRIA, Université de Lorraine, EMOSIS

ENBIS 2023

Context

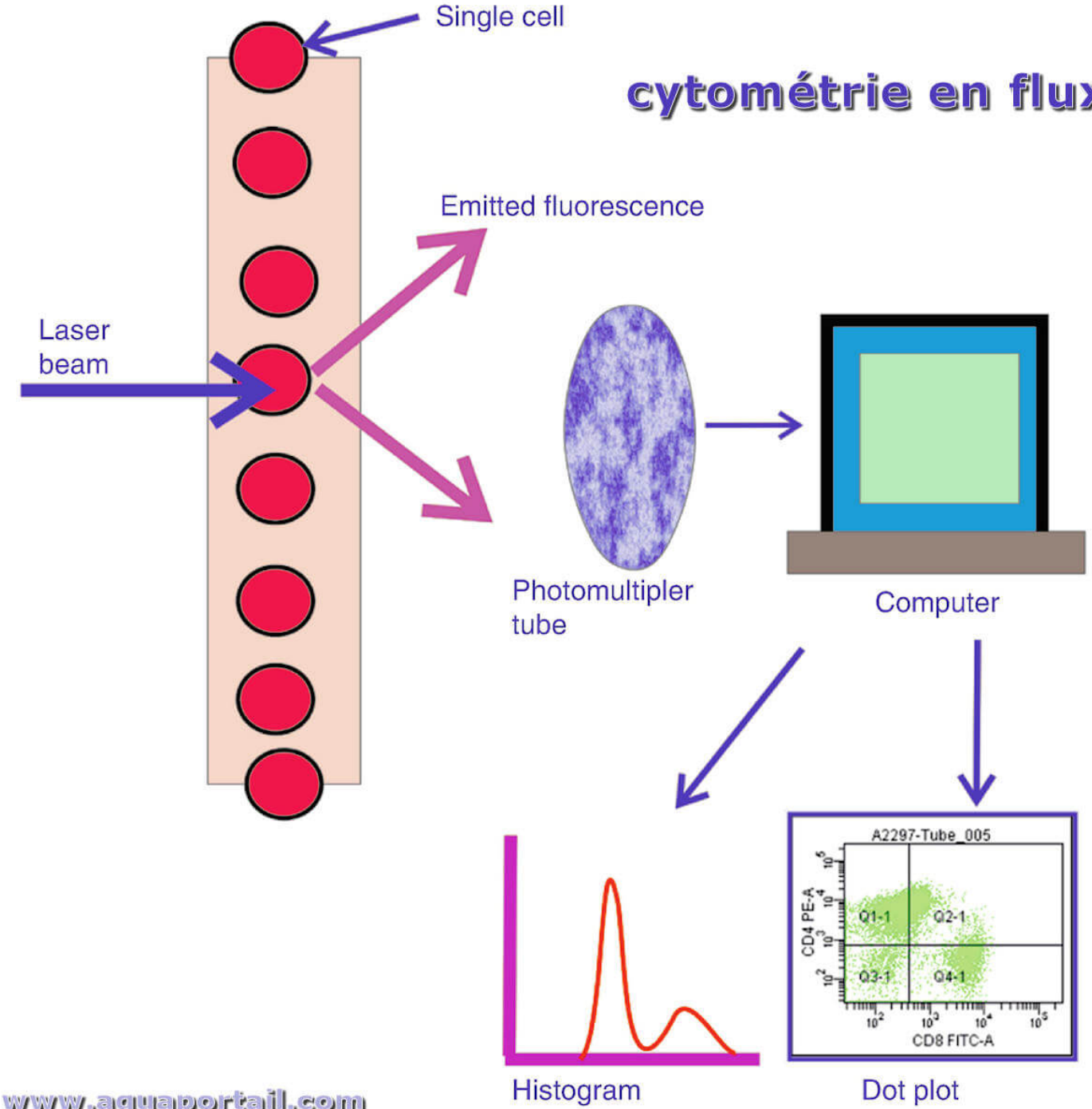
HIT

HIT (Heparin-Induced Thrombocytopenia) is a rare and potentially fatal pathology.

The aim is to be able to create clusters of HIT and non-HIT patients using **Flow cytometry** data from patient blood tubes.

The perspective is to use the results for diagnostic purposes to help the doctor make decisions.

cytométrie en flux



www.aquaportail.com

Flow Cytometry Data

With flow cytometry we measure a data set of p quantitative variables on N cells .
Here an example of a dataset consisting on 8 variables measured on 10000 cells that can be extracted from a patient blood sample.

FSC.A	FSC.H	SSC.A	SSC.H	FL1.H	FL1.A	FL2.H	FL2.A
2,38	2,29	2,93	3,31	0,82	1,12	3,21	3,31
2,23	2,43	2,87	2,74	1,67	1,32	3,02	3,28
3,01	2,98	2,63	2,83	3,03	2,8	2,98	3,04
2,78	2,89	3,02	3,01	2,42	2,2	3,43	3,50
2,67	2,47	3,01	2,87	2,89	2,6	3,28	3,29

FINE Algorithm

The Fisher Information non-parametric embedding (FINE[1] consists of these three steps :

- **1st step** Compute the estimated multidimensionnals densities for each patient tube dataset (n datasets).
- **2nd step** Compute the dissimilarity matrix D (nxn), where $D(i,j)$ represent the distance between the ith and the $j$ith densities.
- **3rd step** Use a classical Multi-Dimensional Scaling[2]on D to obtain our projection in a reduced space.

[1] Carter, K.M., Raich, R., Finn, W.G. and Hero III, A.O., (2009). Fine: Fisher information nonparametric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11), 2093-2098.

[2] Saeed, N., Nam, H., Haq, M. I. U., Muhammad Saqib, D. B. (2018). A survey on Multidimensional Scaling. *ACM Computing Surveys (CSUR)*, 51(3), 1-25.

Fisher Information

During the second step of the FINE algorithm the distance is based on the Fisher Information between two multiparametric distributions $p(x|\theta_1)$ and $p(x|\theta_2)$:

Multiparametric Fisher Information

$$D_F(\theta_1, \theta_2) = \min_{\Theta:\Theta(0)=\theta_1, \Theta(1)=\theta_2} \int_0^1 \sqrt{\left(\frac{d\Theta}{d\beta}\right)^T I(\Theta) \left(\frac{d\Theta}{d\beta}\right)} d\beta$$

Approximation

In most cases we don't know the parameters of our distributions, but we can approximate it with the Kullback-Leibler divergence :

Kullback Leibler formula

$$KL(p||q) = \int p(x) log(\frac{p(x)}{q(x)}) dx$$

Hence we can compute a dissimilarity matrix between all of our patient sample.

Alternatives of the FINE algorithm

First step alternative

- **1st step** Compute the n estimated densities of each patient tube.
- **2nd step** Compute the dissimilarity matrix D with a distance based on the Kullback-Leibler divergeance.
- **3rd step** Use a classical Multi-Dimensional Scaling (cMDS) on D to obtain our projection in a reduced space.

Gaussian assumption

- **1st step** Assume that our tube data are distributed as multivariate Gaussian distributions.
- **2nd step** Compute the dissimilarity matrix D with a distance based on the Kullback-Leibler divergence under the gaussian assumption.
- **3rd step** Use a classical Multi-Dimensional Scaling (cMDS) on D to obtain our projection in a reduced space.

Gaussian Hypothesis and KL Divergence

For two multidimensional normal distributions $f_1 = \mathcal{N}_d(\mu_1, \Sigma_1)$ and $f_2 = \mathcal{N}_d(\mu_2, \Sigma_2)$ the Kullback-Leibler divergence[3] is :

Gaussian Kullback Leibler

$$I_{KL}(f_1||f_2) = \frac{1}{2}\{(tr(\Omega_2\Sigma_1) + (\mu_2 - \mu_1)^T\Omega_1(\mu_2 - \mu_1) - d - ln|\Sigma_1| + ln|\Sigma_2|\}$$

with $\Omega = \Sigma^{-1}$.

[3] Kullback, S., Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1), 79-86.

KNN alternative

- ~~**1st step** Compute the n estimated densities of each patient tube with a gaussian kernel.~~
- ~~**2nd step** Compute the dissimilarity matrix D with a distance based on the Kullback-Leibler divergeance.~~
- **1st & 2nd step** Compute the dissimilarity matrix D directly from the data with a Kullback Leibler divergence using k-nearest neighbors using the package *RANN* on R.
- **3rd step** Use a classical Multi-Dimensional Scaling (cMDS) on D to obtain our projection in a reduced space.

Gaussian Wasserstein

- **1st step** Assume that our tube data are distributed as multivariate Gaussian distributions.
- **2nd step** Compute the dissimilarity matrix D with a distance based on the Wasserstein distance under the gaussian assumption.
- **3rd step** Use a classical Multi-Dimensional Scaling (cMDS) on D to obtain our projection in a reduced space.

Wasserstein distance

For two multidimensional normal distributions $f_1 = \mathcal{N}_d(\mu_1, \Sigma_1)$ and $f_2 = \mathcal{N}_d(\mu_2, \Sigma_2)$, the formula for Wasserstein's distance is :

Gaussian Wasserstein

$$W_2(f_1, f_2)^2 = ||\mu_1 - \mu_2||_2^2 + tr((\Sigma_1 + \Sigma_2) - 2 \times (\Sigma_2^{1/2} \Sigma_1 \Sigma_2^{1/2})^{1/2}).$$

Alternative on the dimension reduction

- **1st step** Compute the n estimated densities of each patient tube.
- **2nd step** Compute the dissimilarity matrix D with a distance based on the Kullback-Leibler divergeance.
- **3rd step** Use the ISOMAP algorithm[4] on D to obtain our projection in a reduced space.

[4] Tenenbaum, J. B., Silva, V. D., Langford, J. C. A global geometric framework for nonlinear dimensionality reduction (2000).

Alternative on the dimension reduction

- **1st step** Compute the n estimated densities of each patient tube.
- **2nd step** Compute the dissimilarity matrix D with a distance based on the Kullback-Leibler divergeance.
- **3rd step** Use the UMAP algorithm[5] on D to obtain our projection in a reduced space.

[5] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." (2018).

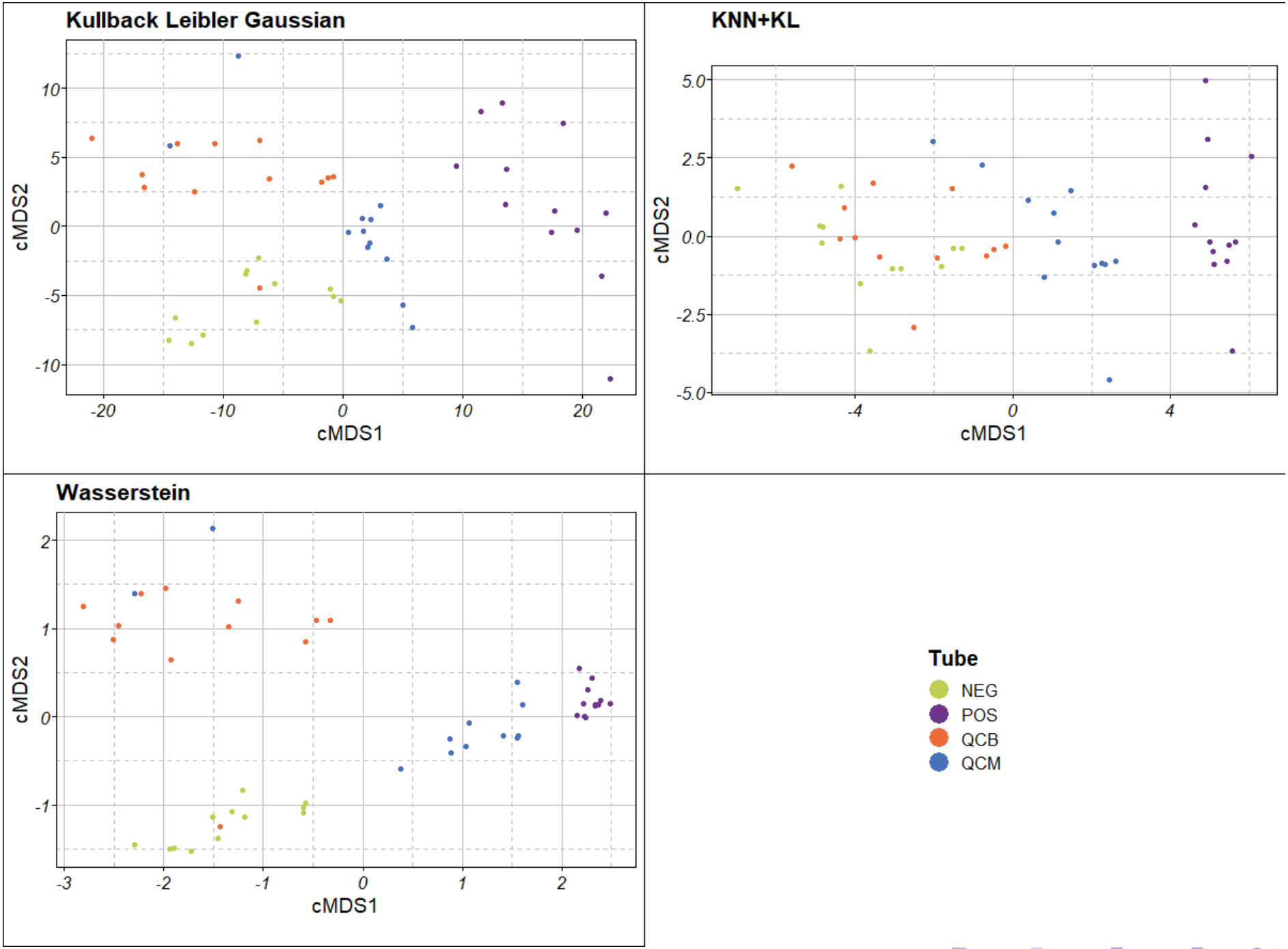
Applications on data

Control sample

Our dataset consists of 48 tubes:

- 12 HIT-positive control tubes **POS**
- 12 HIT-negative control tubes **NEG**
- 12 healthy patient tubes **QCB**
- 12 patient tubes with a priori HIT **QCM**

Our aim is to separate these groups into well-defined clusters.


Kullback Leibler Gaussian
cMDS2
cMDS1
KNN+KL
cMDS2
cMDS1
Wasserstein
cMDS2
cMDS1
Tube
NEG
POS
QCB
QCM

Dataset

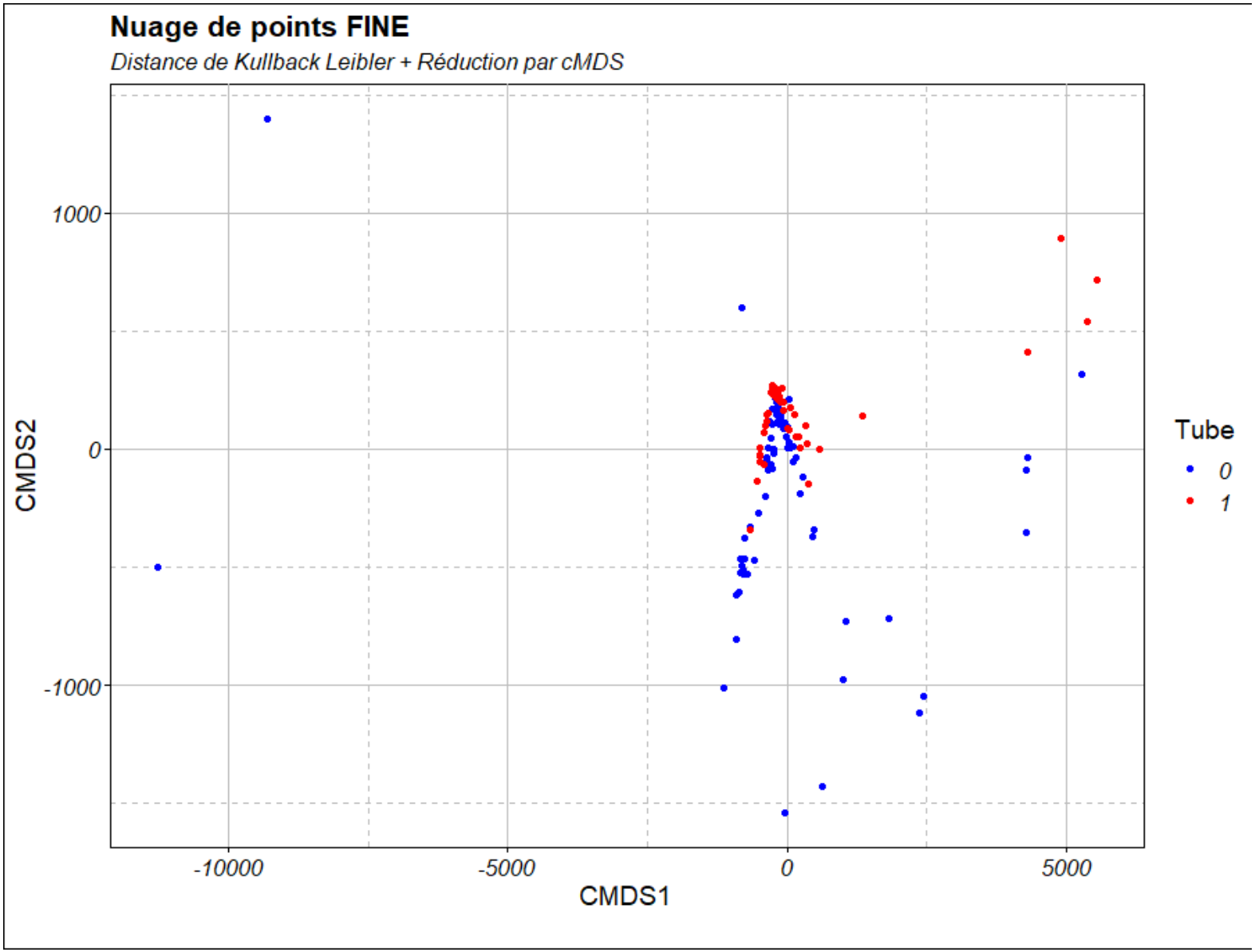
We're now going to test a few approaches on a new dataset based on a HIT study. The aim is to obtain clusters on our 141 patients.

- **0** If the patient is healthy.
- **1** If the patient has HIT.

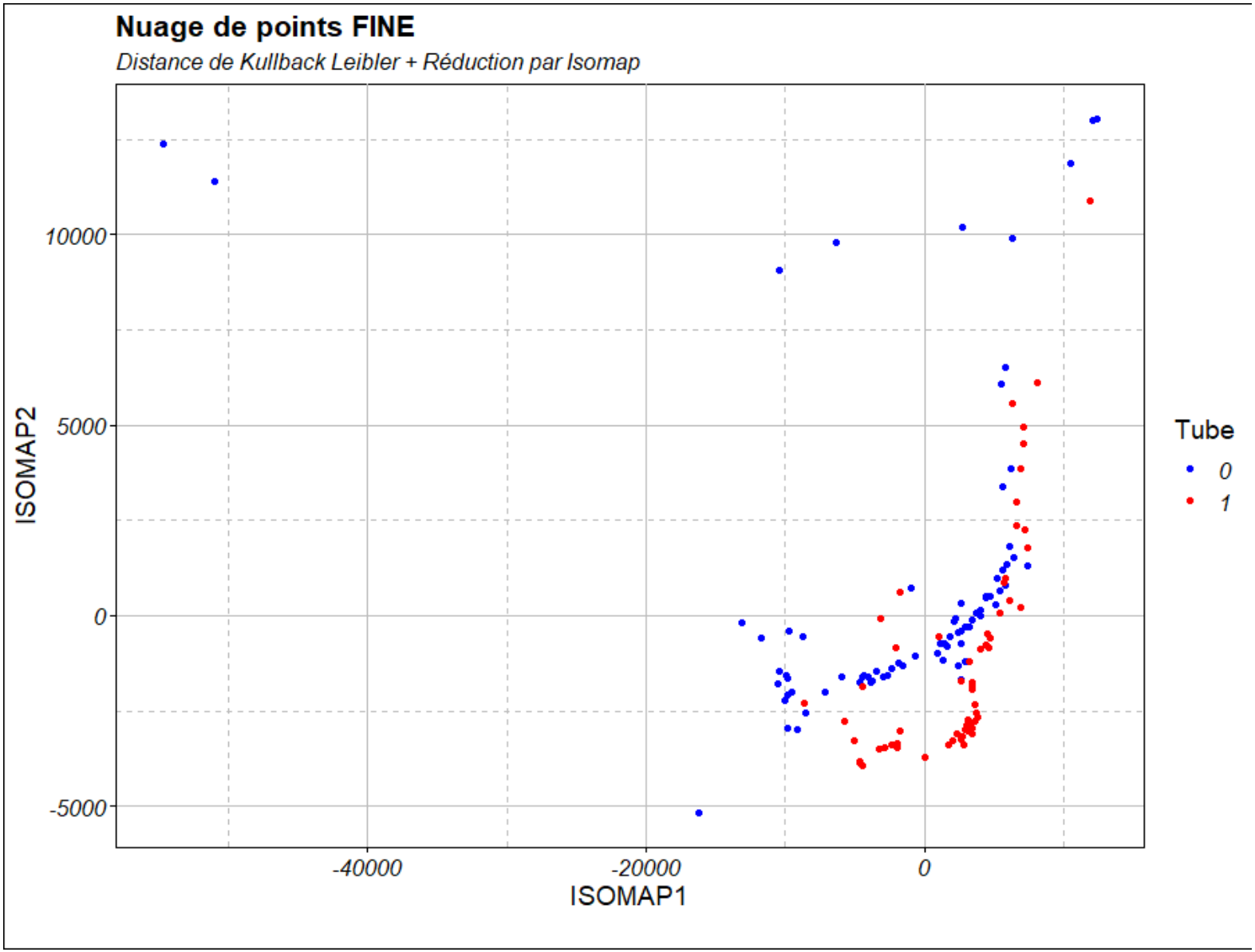
For each of the $n = 141$ patients, we have a dataset containing 8 variables for 10000 cells

The goal is to project our 141 tubes in a reduced space and to visualize two differents clusters of patients

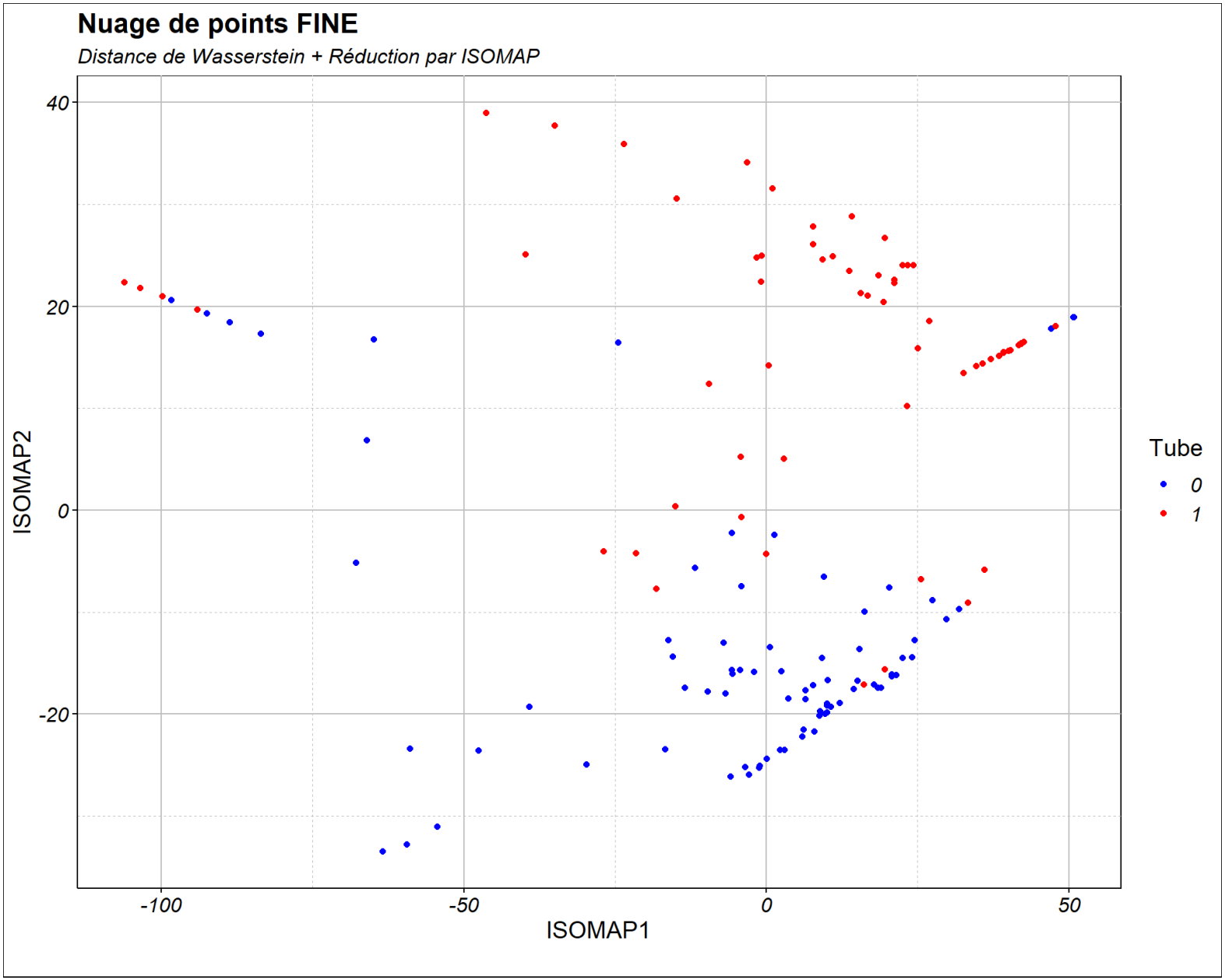
Visualisation classic FINE



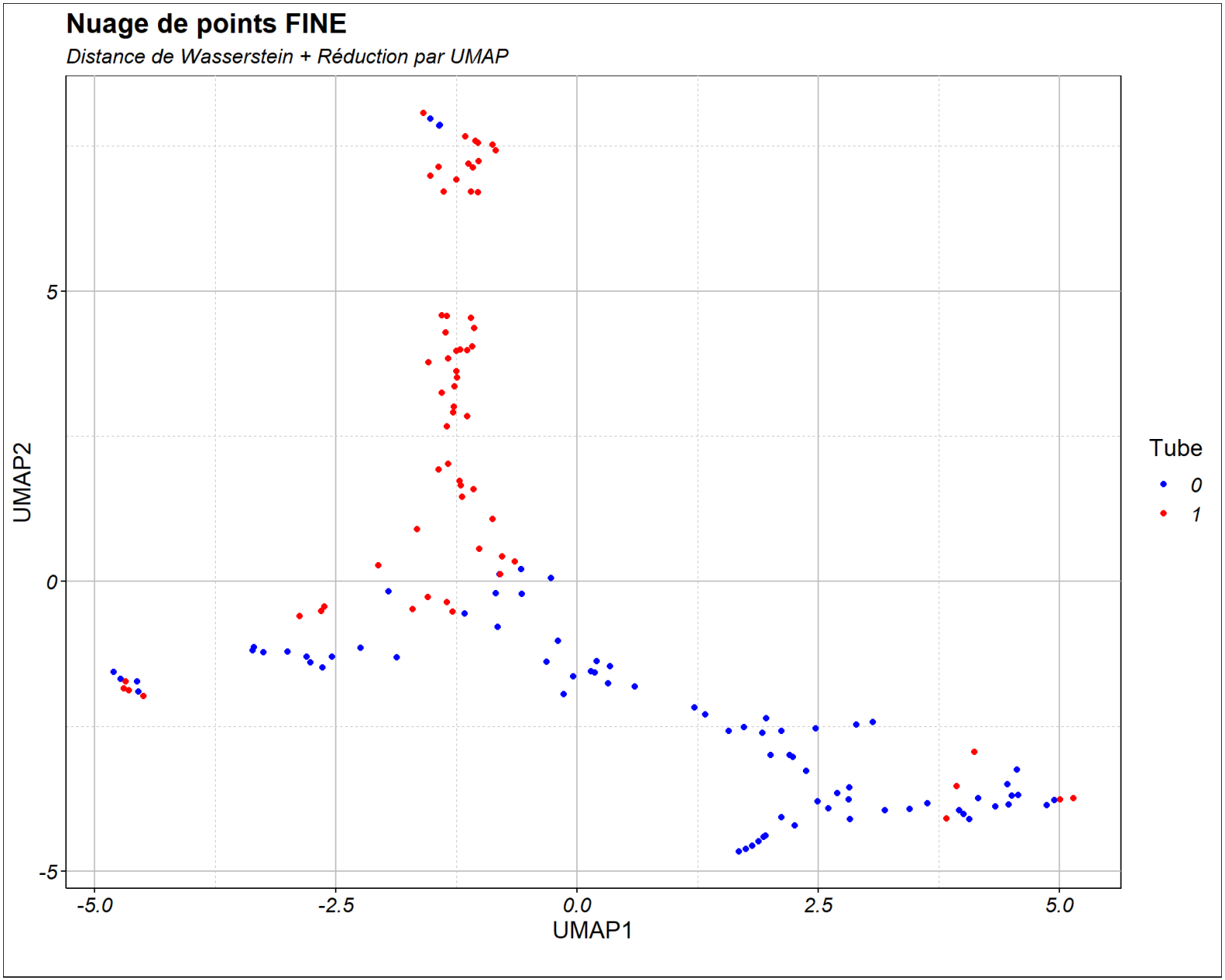
Visualisation FINE + ISOMAP



Visualisation Wasserstein + ISOMAP



Visualisation Wasserstein + UMAP



Performance

Using a SVM algorithm on our projections we measured the performances of our differents methods.

Performance of FINE				
Method	**Sensitivity**	**Specificity**	**Precision**	**Time**
Classical FINE	52%	58%	55%	30 mn
FINE+ ISOMAP	75%	81%	69%	30 mn
KNN + cMDS	65%	66%	64%	100h
Wass - G + ISOMAP	87%	84%	86%	3 mn
Wass - G + UMAP	86%	84%	85%	3 mn

Perspective

- Test our versions of FINE on other data.

- To be able to project a new tube into the reduced space to observe where it fits in relation to the clusters already created.

Thank you for your attention