Machine Learning for biology

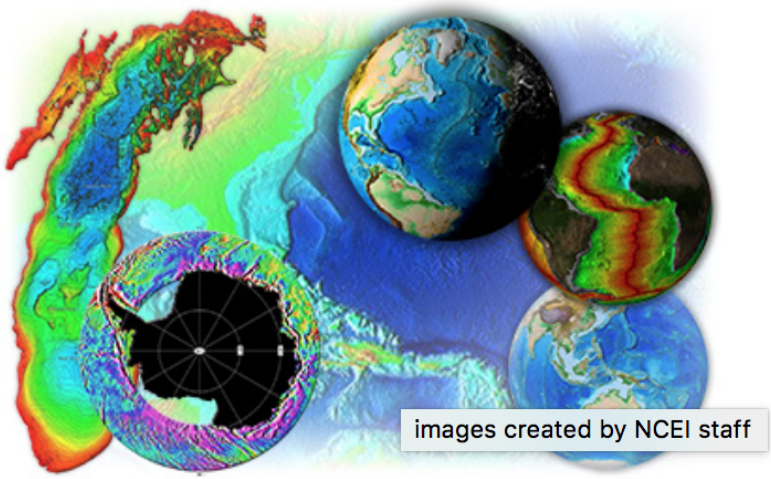
V. Monbet

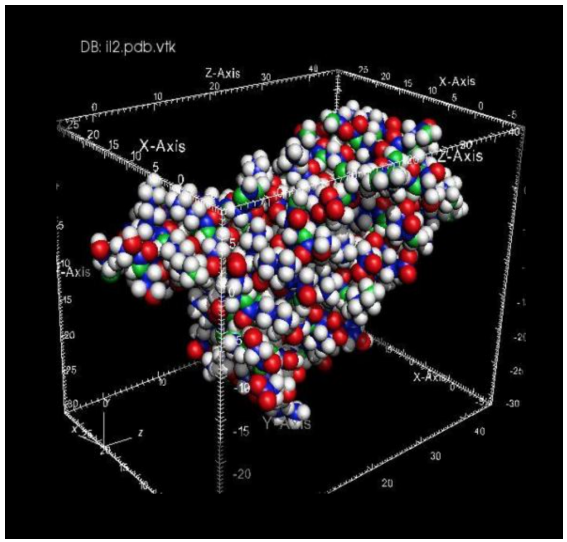
UFR de Mathématiques
Université de Rennes 1

Outline

Outline

- When dealing with huge volumes of data, problems naturally arise. How do you whittle down a dataset of hundreds or even thousands of variables into an optimal model? How do you visualize data with countless dimensions?
- Dimension reduction techniques
  - Principal Component Analysis (PCA) or Empirical Orthogonal Function (EOF)
  - Multidimensional Scaling (MDS) or Principal coordinates analysis and Isomap
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)



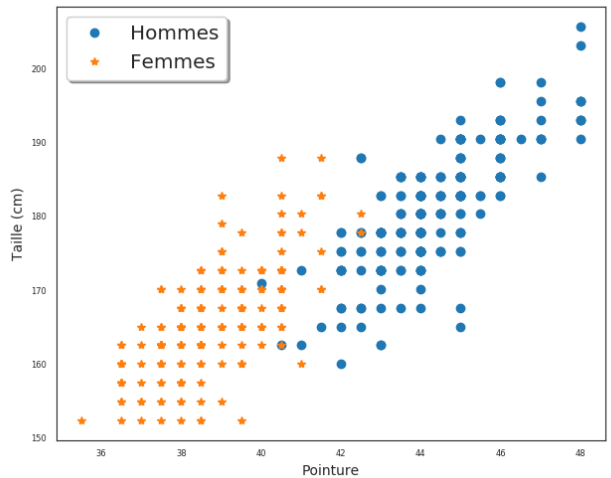


Outline

Principal Component Analysis, introduction

- First example
  When only 2 quantitative variables are observed, it is easy to plot them in a plan.
  Each axis of the plan represents a variable.

Taille	Pointure	Genre
155.1	36.5	F
167.6	40.0	F
⋮	⋮	
182.9	44.0	H



- Relationships between the 2 variables are visible (about linear).
- Two different groups (male/female)

Principal Component Analysis, introduction

- Second example: How to deal with more than 2 variables?
  Composition of 45 potteries found in Great Britain dating from the Roman period 5 different ovens.

PAl2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO	Oven
18.8	9.52	2	0.79	0.4	3.2	1.01	0.077	0.015	1
16.9	7.33	1.65	0.84	0.4	3.05	0.99	0.067	0.018	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19.1	1.64	0.6	0.1	0.03	1.75	1.04	0.007	0.018	5

- Material composition analysis is an important tool for the study of trade in ancient economies. Objects of distinct origins have generally different chemical signatures that identify their origin.
  In order to identify these signatures it is necessary to be able to group together objects of similar composition.

Principal Component Analysis or Empirical Orthogonal Functions

- Consider a set of p variables observed for n samples.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

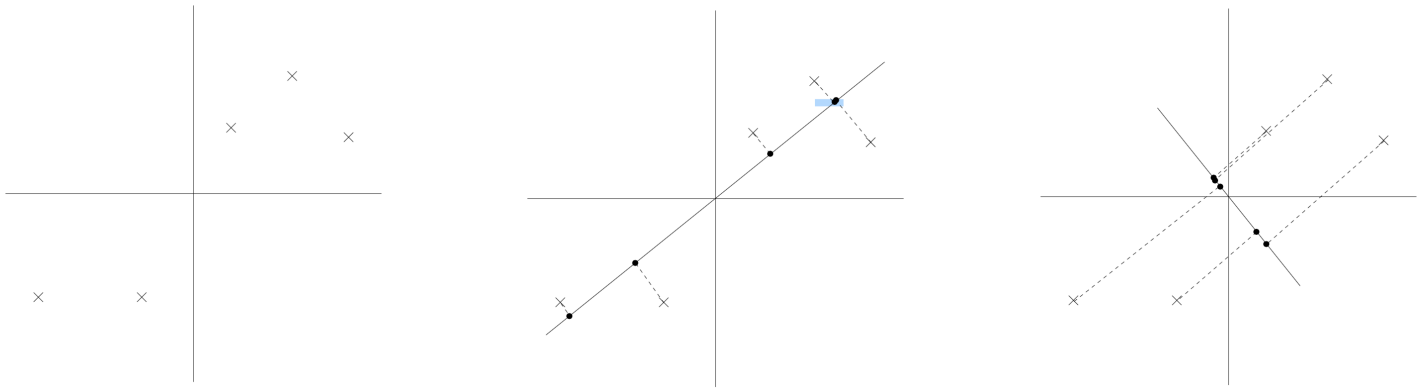
- $\mathbf{X} \in \mathbb{R}^{n,p}$
- A line x_i represents one individual (or sample) ; it is a point of $\mathbb{R}^p$
- A column x_j represents one variable (or feature)

Representation/visualisation for more than 2 variables

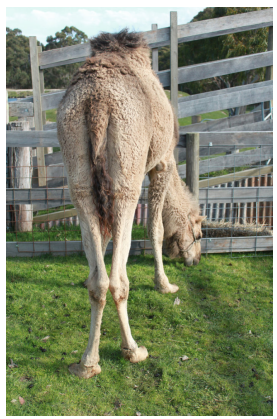
- More than 2 variables: how to represent individuals?
- Remark: If the individuals are concentrated in a $\mathbb{R}^2$ plane,
  $\rightarrow$ represent in this plan.
  From a mathematically point of vue, a basic change is performed (rotation) and the first two variables are represented, the others (from 3 to p) being zero.
- General idea: If the individuals are all close to a plane,
  $\rightarrow$ find the best plane (in the sense that the sum of the distances of the points to the plan is as small as possible)
  $\rightarrow$ represent the data by projection onto this plane.
- Required quality: The distances between projected individuals must reflect at the better their distances in $\mathbb{R}^p$ ie *Respect the geometry of the data.*

Principal Component Analysis

- **Recap**: For a given dataset, PCA builds the projection in a space of dimension $q < p$ which gives the best overview of the data, preserves the distance between individuals and does not deform the "image".
- Simple example ($p = 2$, $q = 1$)



- What is the best 2D representation of the camel?


Representation/visualisation for more than 2 variables

Miscellaneous remarks

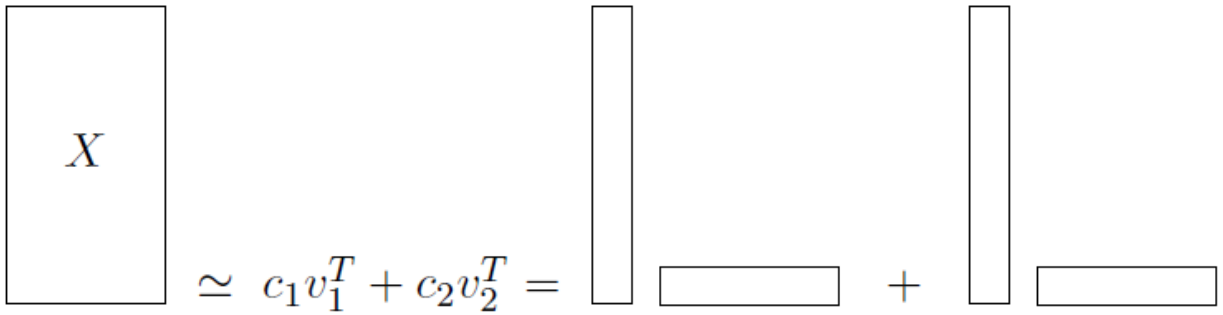
- Projecting in dimension two (ie in a plane) may be too coarse. You can look for the best sub-space of dimension three (or more)
  $\rightarrow$ data representation is more difficult: in practice we will represent 2 or 3 projections in dimension 2.
  $\rightarrow$ interpretation may be hard.

- Difficulty related to the standardization of variables.
  If, for example, $p = 3$, multiplying the 3rd dimension by a small factor will concentrate the individuals on the plane containing the two first.
  "individuals are almost in a plane" depends on the scales.

- The outcome will be the creation of new variables called *principal components*.

Mathematical aspects of PCA

From a matricial point of vue, X is approximated by a matrix with rank equal to 2



where $X \in \mathbb{R}^{n,p}$ is the matrix of the observations, $c_1 \in \mathbb{R}^{n,1}$, $c_2 \in \mathbb{R}^{n,1}$, $v_1 \in \mathbb{R}^{n,1}$ and $v_2 \in \mathbb{R}^{n,1}$ are column vectors
c_1 and c_2 are the coordinates of the individuals in the PCA plan.
v_1 and v_2 are the vectors which define the PCA plan.

Each individual (or sample) is projected on the plan spanned by the principal axis v_1 and v_2.

The relation $X \simeq c_1 v_1^T + c_2 v_2^T$ also says that each variable can be approximated by a linear combination of the principal coordinates c_1 and c_2.

Notation : • v^T stands for "transpose of x". A matrix v undergoes transposition when its rows and columns are interchanged.

Inertia

- In the multidimensional setting, the dispersion of the set $\mathcal{S} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is measured by inertia

$$I = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i||^2$$

  For quantitative variables, inertia is also the sum of the empirical variances.

- If P_E is an orthogonal projector on a vectorial space $E \subset \mathbb{R}^p$; the inertia of E is defined by

$$I_E = \frac{1}{n} \sum_i ||P_E x_i||^2 = \frac{1}{n} \sum_i ||x_i||^2 - \frac{1}{n} ||x_i - P_E x_i||^2$$

  from Pythagore.

- Hence, the PCA is equivalent to the projection on the space of dimension q which better preserve the inertia.

  Minimize the errors $x_i - P_E(x_i)$ is equivalent to maximize the variability of $P_E(x_i)$

Notations : • $||x||$ stands for "norm of x" and, in most of the cases occuring in machine learning, $||x - y|| = d(x, y)$, the distance between x and y
• $\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$

Practical aspects of PCA

- PCA provides axes v_j and coordinates c_j such that

$$\mathbf{x} = c_1 v_1^T + \cdots + c_q v_q^T + \epsilon \tag{1}$$

where $v_j \in \mathbb{R}^p$ are the factors, $v_j \perp v_\ell$ and $||v_j|| = 1$,
$c_j \in \mathbb{R}^n$ are the principal components
ϵ is a residual ; it can be interpreted as an approximation error.

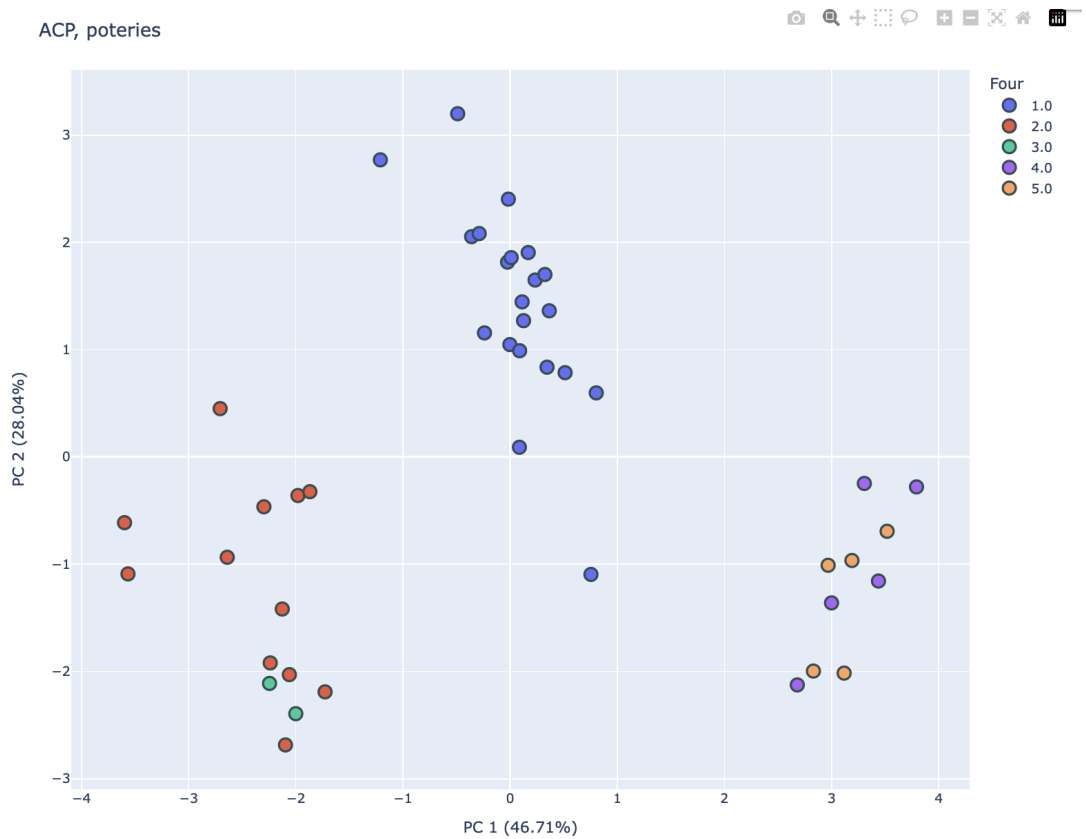
- Factors v_j are the q eigenvectors of the covariance matrix $\hat{\Sigma} = \frac{1}{n}\mathbf{x}^T\mathbf{x}$ associated with the q largest eigenvalues.
It means that the space generated by $v_1, \cdots, v_q$ is the space of dimension q for which the inertia of the projection of $\mathbf{x}$ is the largest.

- Coordinates (= principal components) of individual i

$$c_{ij} = \mathbf{x}_i^T v_j$$

The coordinates allow to plot the individuals (see next slide).

Représentation des poteries

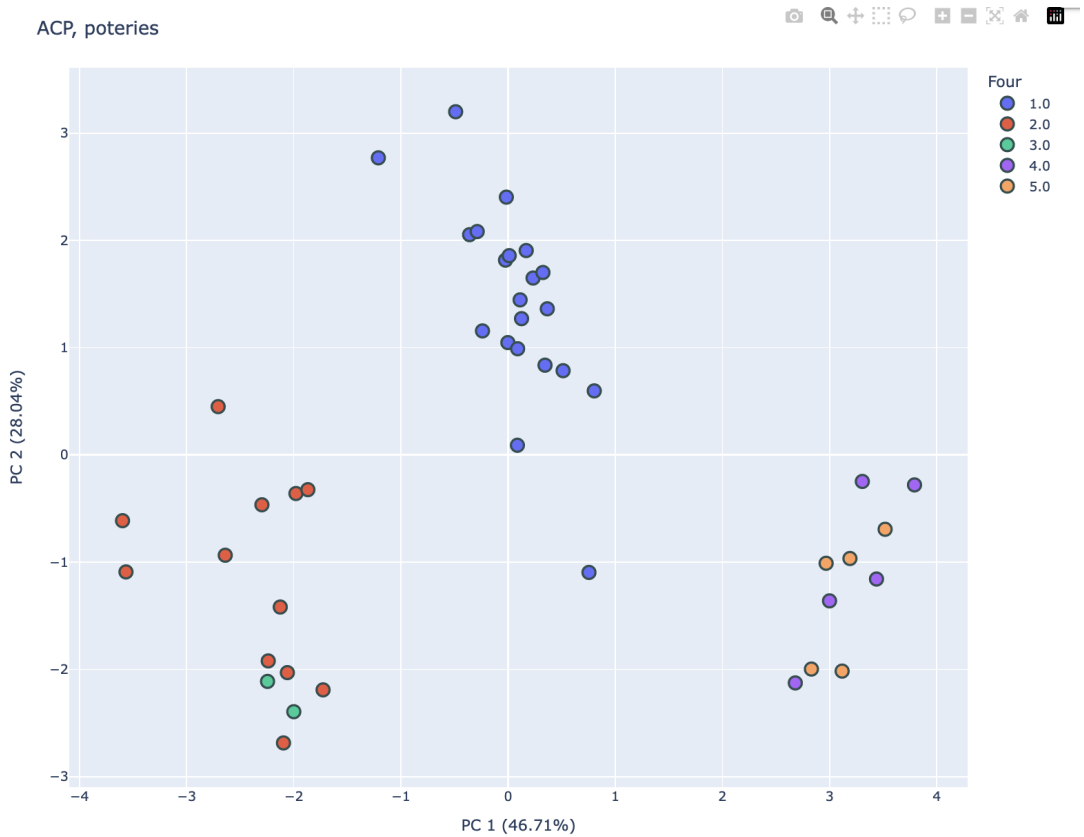


Sur les légendes d'axe, on reporte le pourcentage d'inertie expliqué par chaque composante principale (voir ci-dessous).

La variable «four» n'a pas servi à l'ACP: apprentissage non-supervisé.

code

Example of potteries



Note that the *oven* variable has not been used for the analysis but only for the plot. However, the poteries are grouped according to the oven in the first plan of the PCA.

code

Poetries, principal axes interpretation

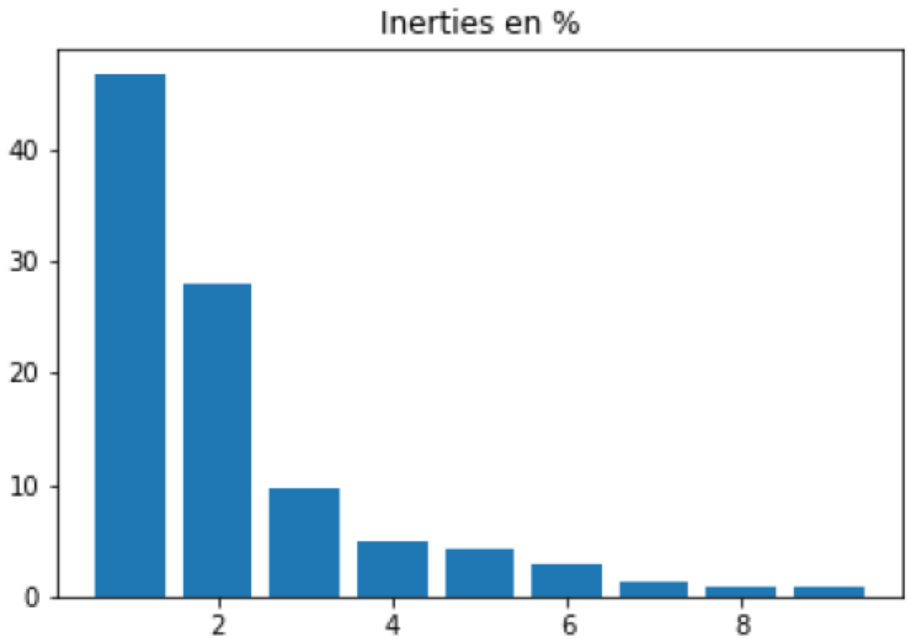
PCA axes can be interprete by plotting the coordinates of the original variables on principal axes (ie SVD's eigen vectors).

	PAL2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
PC1	0.35	-0.33	-0.43	-0.06	-0.22	-0.46	0.34	-0.46	-0.02
PC2	0.33	0.40	-0.19	0.50	0.46	-0.02	0.30	0.09	0.38

code

Poteries, inertia

The largest the inertia in the new space E the better the representation of the initial dataset. It is common to represent the decrease in intertia as the following barplot.

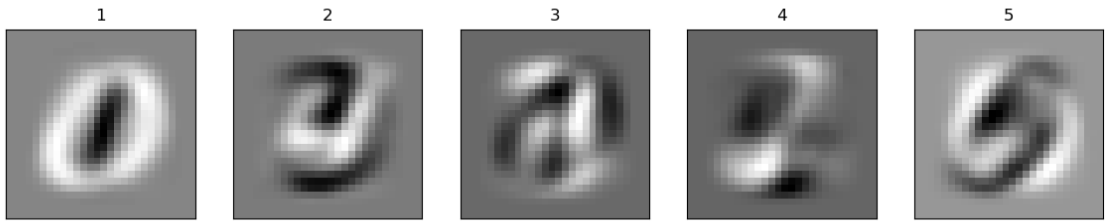


In practice, the approximation including q components is considered to be good if the gain of inertia obtain by adding a $q+1$th component is low.

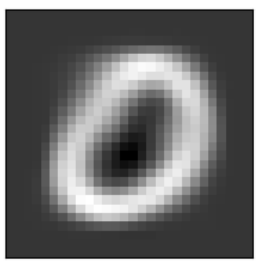
PCA reconstruction, example with handwritten digits images

Original images

Factors (5 firsts)



Reconstruction



$$= c_{11} v_1^T + c_{12} v_2^T + c_{13} v_3^T + c_{14} v_4^T + c_{15} v_5^T$$

PCA reconstruction

PCA 5 first factors

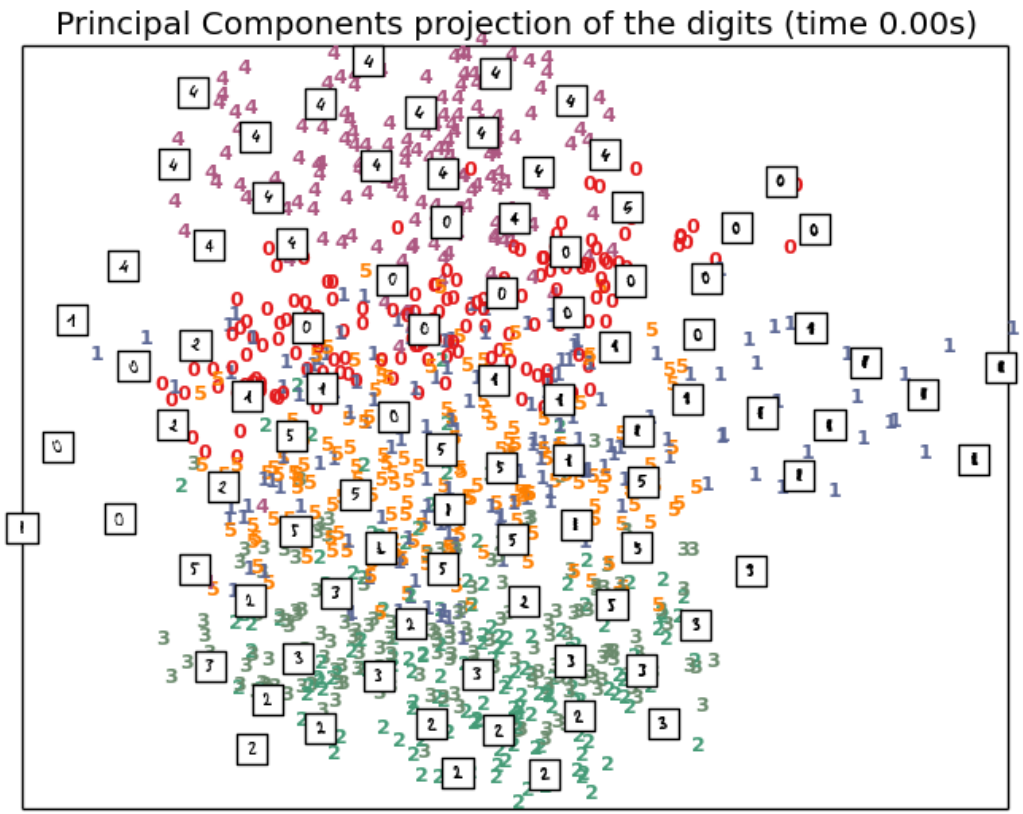
Reconstructed images (based on 5 factors)

with **weights**

image 0	2.44	-0.74	-0.60	0.02	0.01
image 1	-1.38	-0.62	-1.18	0.96	-0.05
image 2	-0.68	0.13	0.43	1.66	0.36
image 3	-0.23	-0.95	0.27	-1.55	0.63
image 4	0.25	1.03	0.12	-1.88	0.23

Principal Component Analysis, visualization example

- PCA is often used to help plot data in high dimensions. For the digits example, it helps to "discriminate" the numbers.
- For $q = 2$, the 1s are on the right side of the first axis while numbers 0, 2, 3, 4 and 5 are on the left side. The 2s are on the bottom of the second axis and the 4s at the top.
- Similar-shaped numbers are close to each others: if 7s were added they would be close to the 1s.



Outline

PCA for multiple imputation

- In biology and other experimental sciences, some data can be missing in a data set.
- Sometimes people impute the missing data with the mean of the concerned variable. However it is not a good idea.
- An iterative PCA algorithm can be used to impute missing data.
- It allows to take into account the multivariate correlations.

Algorithm
Initialization: replace missing values by the mean of the variables
and compute $\bar{\mathbf{X}}^{(0)}$
Repeat until convergence
(a) Compute a PCA on $\mathbf{X}^{(\ell-1)}$ to find $\mathbf{V}^{(\ell)}$ and $\mathbf{c}^{(\ell)}$
(b) Missing values are replaced by

$$\mathbf{X}^{(\ell)} = \mathbf{V}^{(\ell)}\left(\mathbf{c}^{(\ell)}\right)^T + \bar{\mathbf{X}}^{(\ell-1)}$$

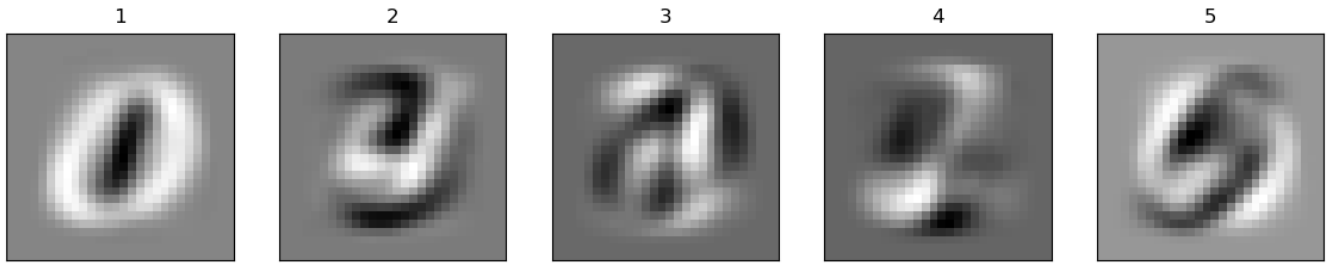
$$\mathbf{X}^{(\ell)} = W\mathbf{X} + (1 - W)\mathbf{X}^{(\ell)}$$

where $w_{ij} = 0$ if observation is missing and 1 otherwise.
(c) Compute $\bar{\mathbf{X}}^{(\ell)}$

Ref: Josse, J., Pages, J., & Husson, F. (2011). Multiple imputation in principal component analysis. Advances in data analysis and classification, 5(3), 231-246.

Principal Component Analysis vs other projection methods

- PCA is a projection on a basis of orthogonal vectors (see example below).



- Analogy with Fourier, wavelets, etc.
- Particularities of PCA: the basis is data driven.
- Used for visualization, description, dimension reduction.

- PCA is not appropriate for every data because it is based on euclidean distance. For example, ecological data may contain a lot of 0s that are ignored by euclidean distance.
  Other methods for reduction dimension can work with other distances.

Outline

Other methods: multidimensional scaling (MDS)

- There are other methods for dimension reduction based on the searches of orthogonal/independent basis or factors.
- MDS or Principal coordinates analysis PCoA
  Multi-Dimension Scaling is a learning method which allows to represent samples on a low dimension space preserving the distances between samples.

  Given a matrix of distances D between the n observations, 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 minimizing the following cost function

$$J(X_{mds}) = \sum_{i<j} \omega_{ij}(d_{ij}(X_{mds}) - D_{ij})^2$$

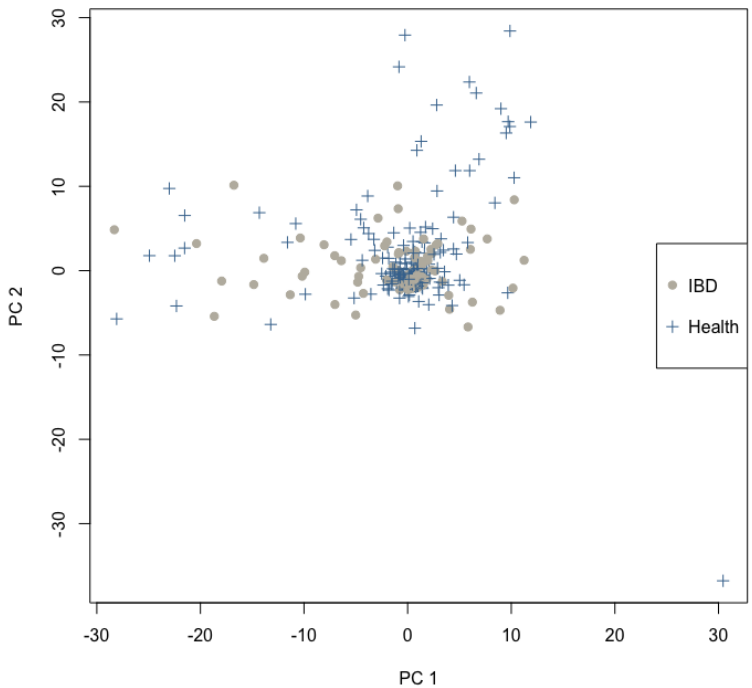
  where $d_{ij}(X_{mds})$ denotes the euclidean distance in the low dimension MDS space and ω_{ij} are weights.

- When D is defined by the euclidean distances, MDS is equivalent to PCA.

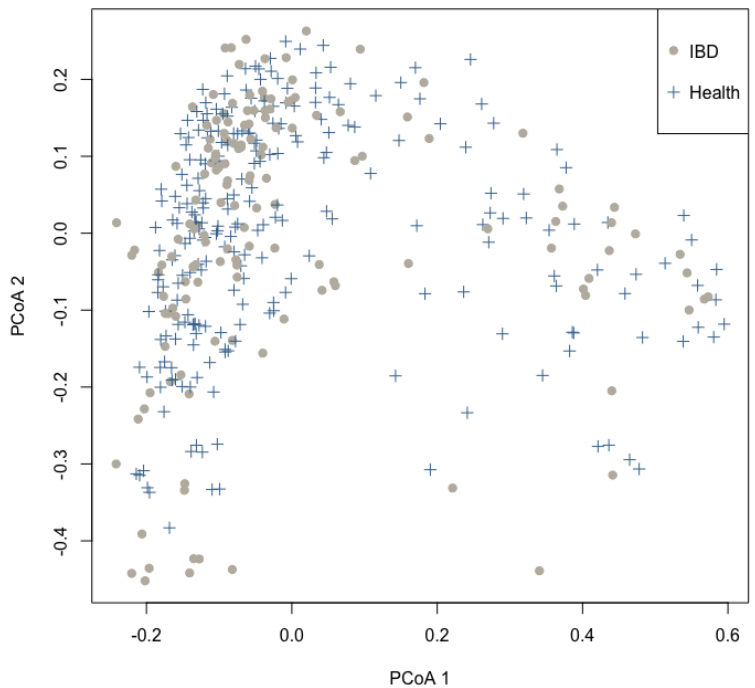
Example : gut microbiota

- Irritable Bowel Disease (IBD) and gut microbiota
  Can we distinguish IBD pateint from control if we only knew the genetic differences between communities of their gut micro-organisms?

PCA (on clr transformed data)



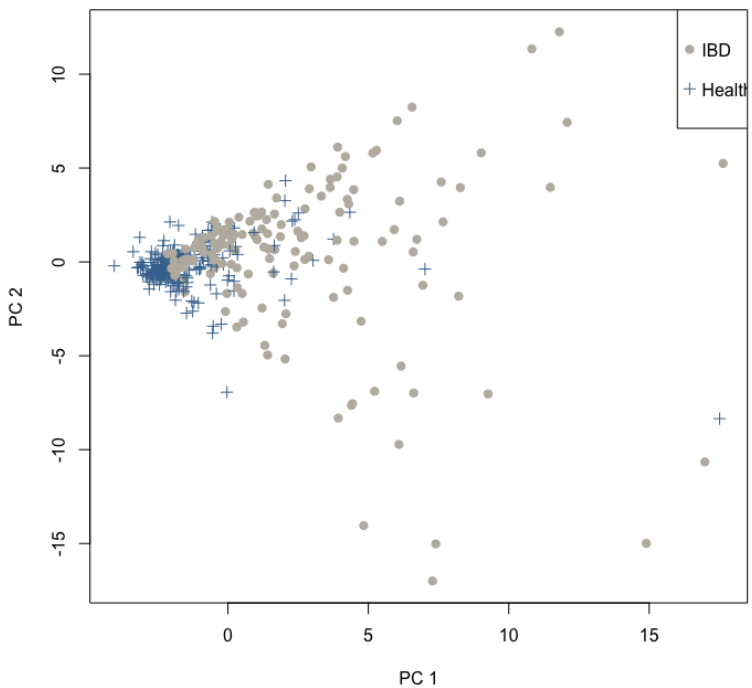
MDS with Bray distance



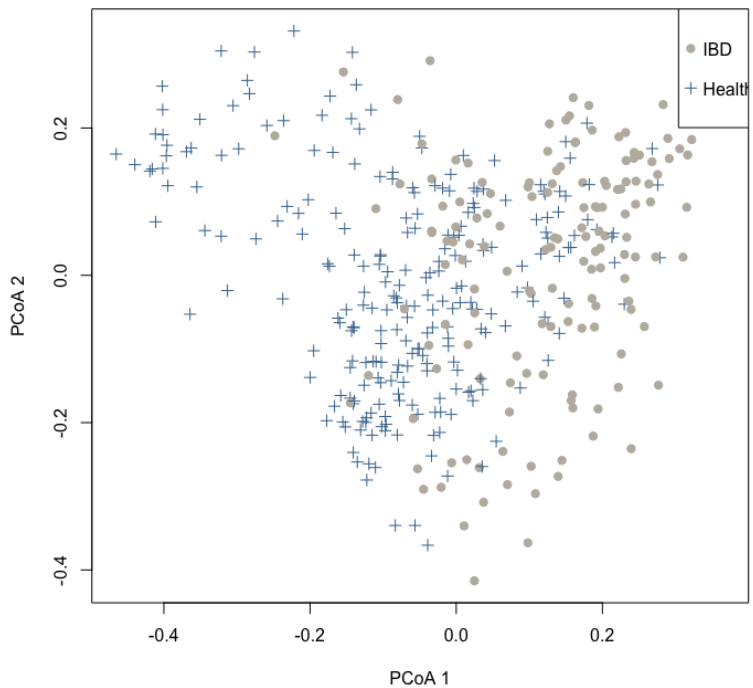
PCA or MDS do not provide direct answers to the question, but they can help to have an idea of the respons (see next page).

- Gut microbiota comprises approximately 600 OTUs, which is quite extensive. Preprocessing the data may enhance the clarity of distinguishing between the two groups.
  Note that, most of the OTUs are probably not linked with the disease.
  Here, Student tests were used to select a subset of "important" variables and improves the plots.

PCA

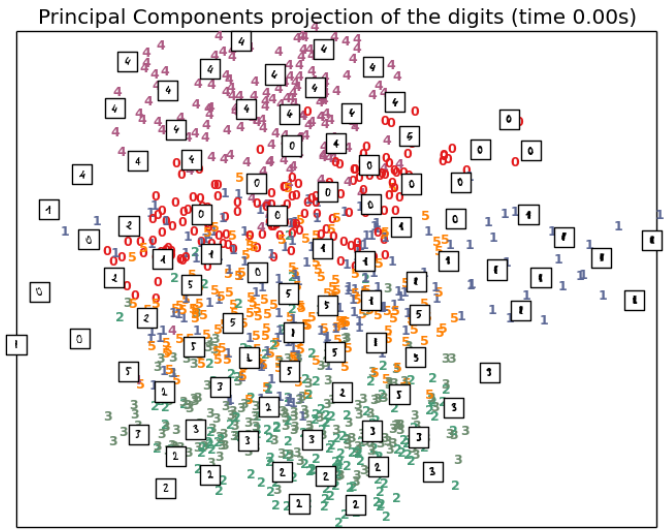


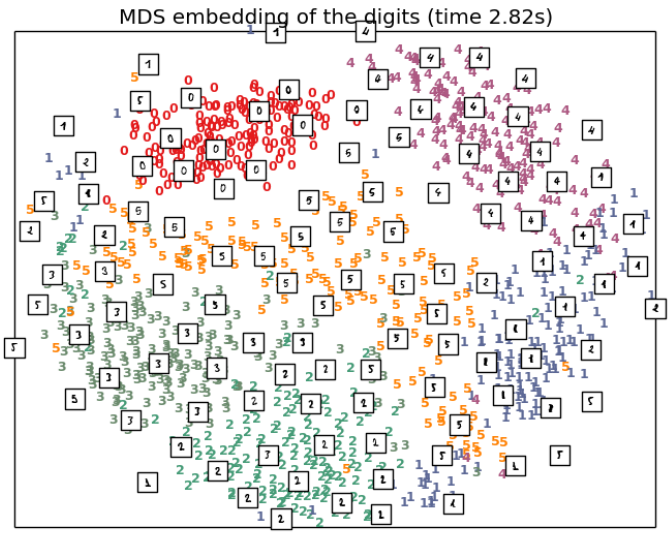
MDS (Bray distance)



MDS, digits data

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





Optimization algorithm

- The solution of

$$\min_{X_{mds}} \sum_{i<j} \omega_{ij}(d_{ij}(X_{mds}) - D_{ij})^2$$

  is usually approximated by iterative algorithms

- Majoration-Minimization (MM) approach

Find a function $g(x, x_m)$
easy to minimized
Such that $J(x) \leq g(x, x_m)$
And $f(x_m) = g(x_m, x_m)$

Steps of MM

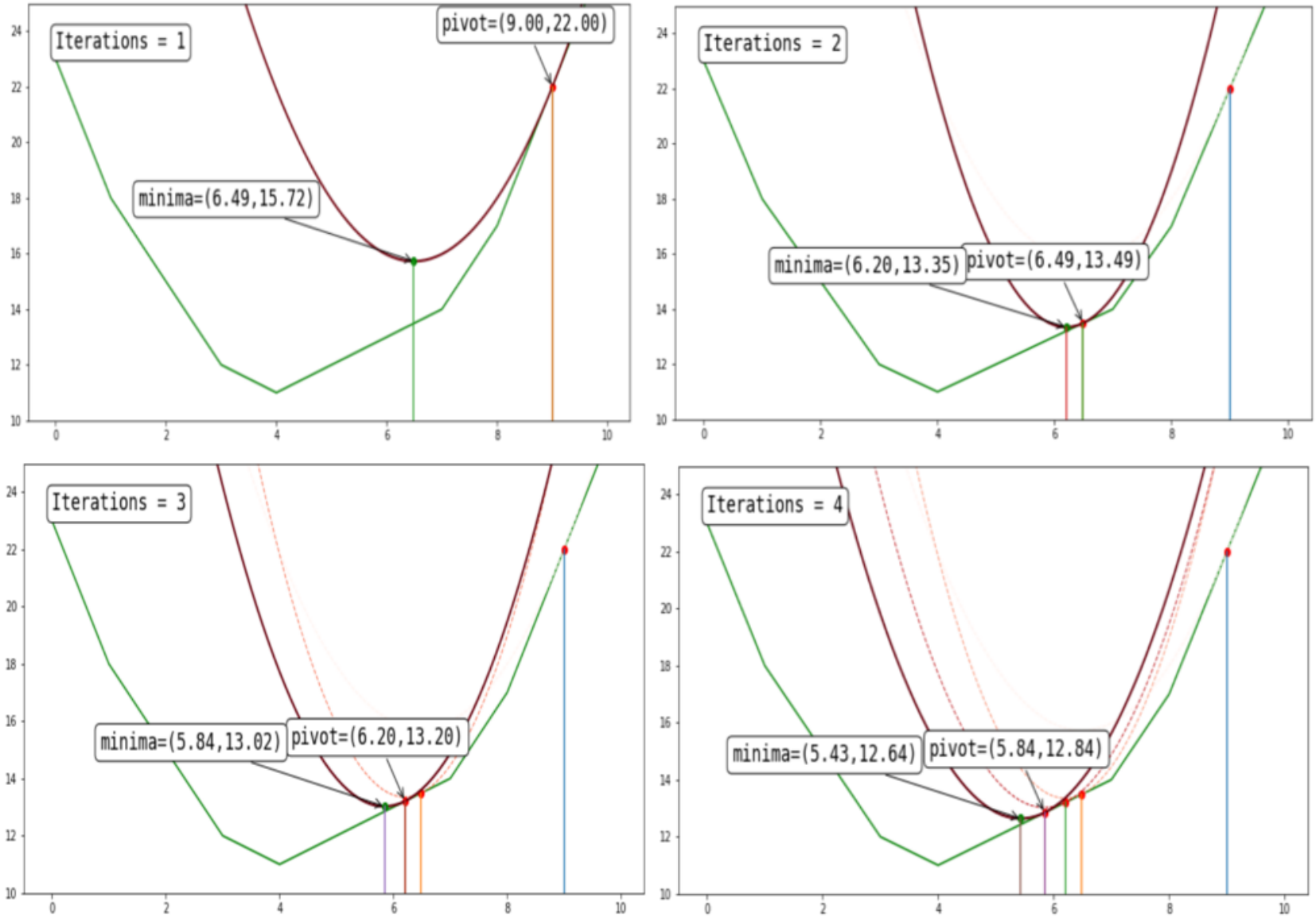
1. choose a random support point x_m
2. find $x_{min} = \arg\min_x g(x, x_m)$
3. if $|f(x_{min}) - f(x_m)|$ small then break
else go to step 4
4. set $x_m = x_{min}$ and go to step 2

- But it can also be transformed in a spectral analysis problem (computation of eigen vectors).

Notations : • $\min_{X_{mds}} J(X_{mds})$ means find the minimum of the function $J : X_{mds} \mapsto J(X_{mds})$

• $x_{min} = \arg\min_x g(x, x_m)$ means that $x_m in$ is the point x for which $g(x, x_m^)$ is minimum.

Optimization algorithm



https://blog.paperspace.com/dimension-reduction-with-multi-dimension-scaling/

Limitations of PCA and MDS

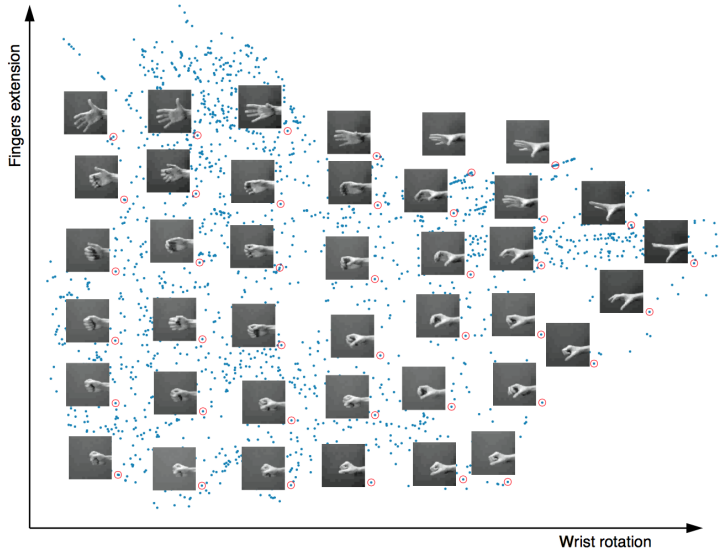
One of the limitations of methods such as MDS and PCA is that their effectiveness is limited by the fact that they are globally linear methods : if the original data is inherently non-linear these methods will represent the true reduced manifold in a subspace of higher dimension than necessary in order to cover non-linearity.

Among other methods : principal curves (Hastie and Stuetzle 1989, Tibshirani 1992), multi-layer auto-associative neural networks (Kramer 1991), local PCA (Kambhatla and Leen 1997), and generative topographic mapping (Bishop et al. 1998), isomap method (Tenenbaum et al. 2000).

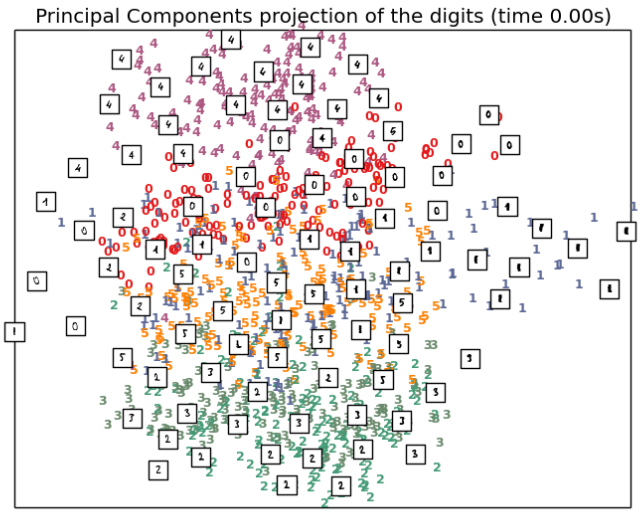
Other methods : Isomap

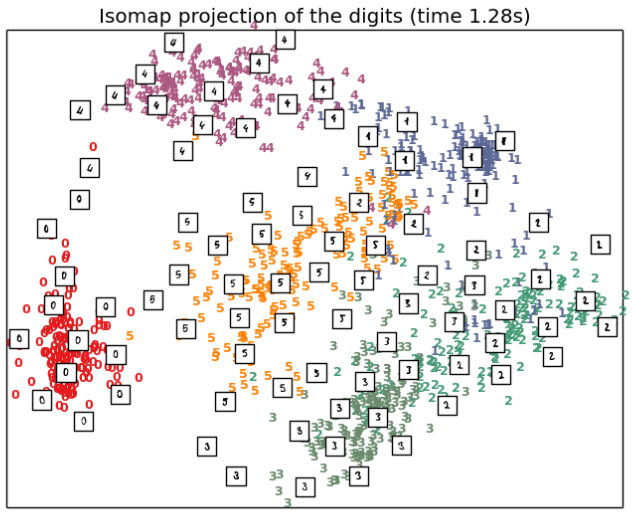
- Isomap is a version of MDS based on a distance on a neighborhood graph.
- D_{ij} is the number of edges between i and j
- Isomap is mostly used for image dimension reduction because it is able to well capture the "motions".

  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)





Outline

Non-negative Matrix Factorization

Non-negative matrix factorization (NNMF) is a tool for dimensionality reduction of datasets in which the values, like the rates in the rate matrix are constrained to be non-negative (ex: approximation of micro-array data, approximation of soil composition, etc.).
$\rightarrow$ easier to interpret.

NNMF

$$X \simeq HW$$

(or $X_j \simeq h_{j1} W_1 + \cdots + h_{jq} W_q$)
where $V \in \mathbb{R}^{n,p}$,
$W \in \mathbb{R}^{n,q} \rightarrow$ basis
$H \in \mathbb{R}^{q,p} \rightarrow$ weights
constraints : $W \geq 0, H \geq 0$.

(H, W) solution of

$$\min_{W \geq 0, H \geq 0} ||X - WH||_F^2$$

PCA

$$X \simeq Vc$$

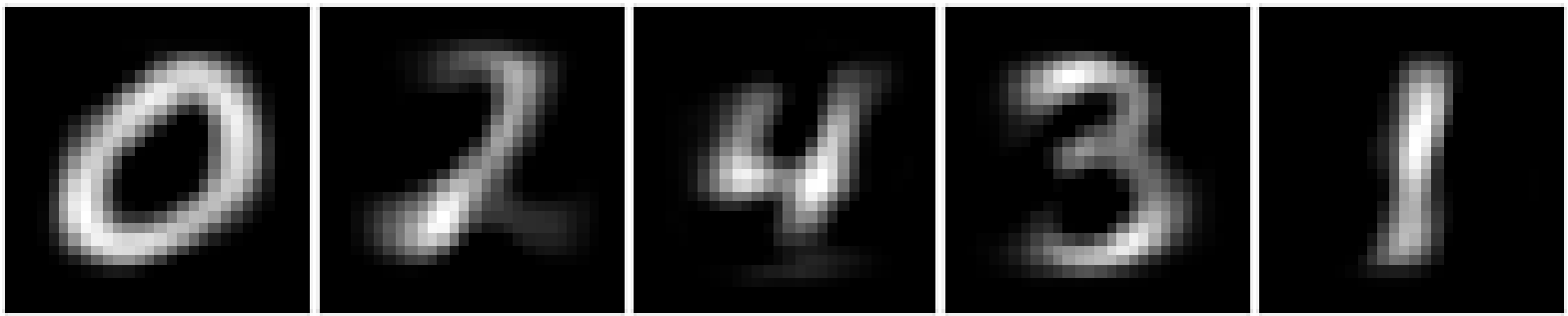
(or $X_i \simeq c_{i1} v_1 + \cdots + c_{iq} v_q$)
where $X \in \mathbb{R}^{n,p}$,
$V \in \mathbb{R}^{n,q} \rightarrow$ basis
$C \in \mathbb{R}^{q,p} \rightarrow$ weights
no constraints

- $||X - Y||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (X_{ij} - Y_{ij})^2}$ is the Frobenius norm of $X - Y$ is may be read as a distance between the 2 matrices

NMF loadings

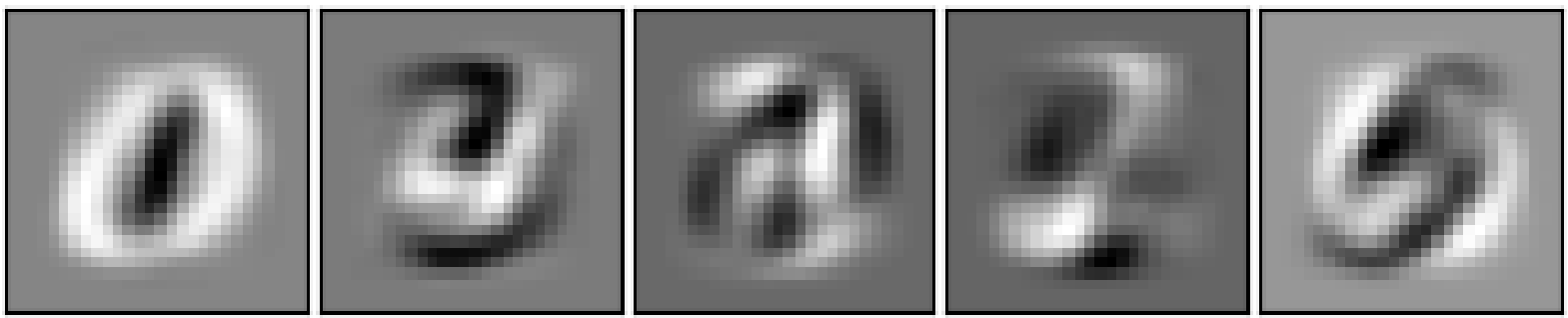
Comparison of NMF and PCA for the handwritten digits

NMF 5 first factors



One of the advantages of NMF is that the components may be interpretable in the original space.

PCA 5 first components



NMF reconstruction

NMF 5 first factors

Reconstructed images (based on 5 components)

with **weights** that can be interpreted as pourcentages

image 0	0.61	0.	0.01	0.23	0.
image 1	0.	0.24	0.	0.	0.46
image 2	0.	0.56	0.03	0.09	0.13
image 3	0.02	0.	0.04	0.37	0.16
image 4	0.11	0.	0.5	0.06	0.22

PCA reconstruction

PCA 5 first factors

Reconstructed images (based on 5 components)

with **weights**

image 0	2.44	-0.74	-0.60	0.02	0.01
image 1	-1.38	-0.62	-1.18	0.96	-0.05
image 2	-0.68	0.13	0.43	1.66	0.36
image 3	-0.23	-0.95	0.27	-1.55	0.63
image 4	0.25	1.03	0.12	-1.88	0.23

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
  - t-SNE performs different transformation on different regions of the space.
- 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. It drives the size of the neiborghood of x_i. By convention, $p_{i|i} = 0$.
- In the reduced space, the similarity is measured by a conditional density as well

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

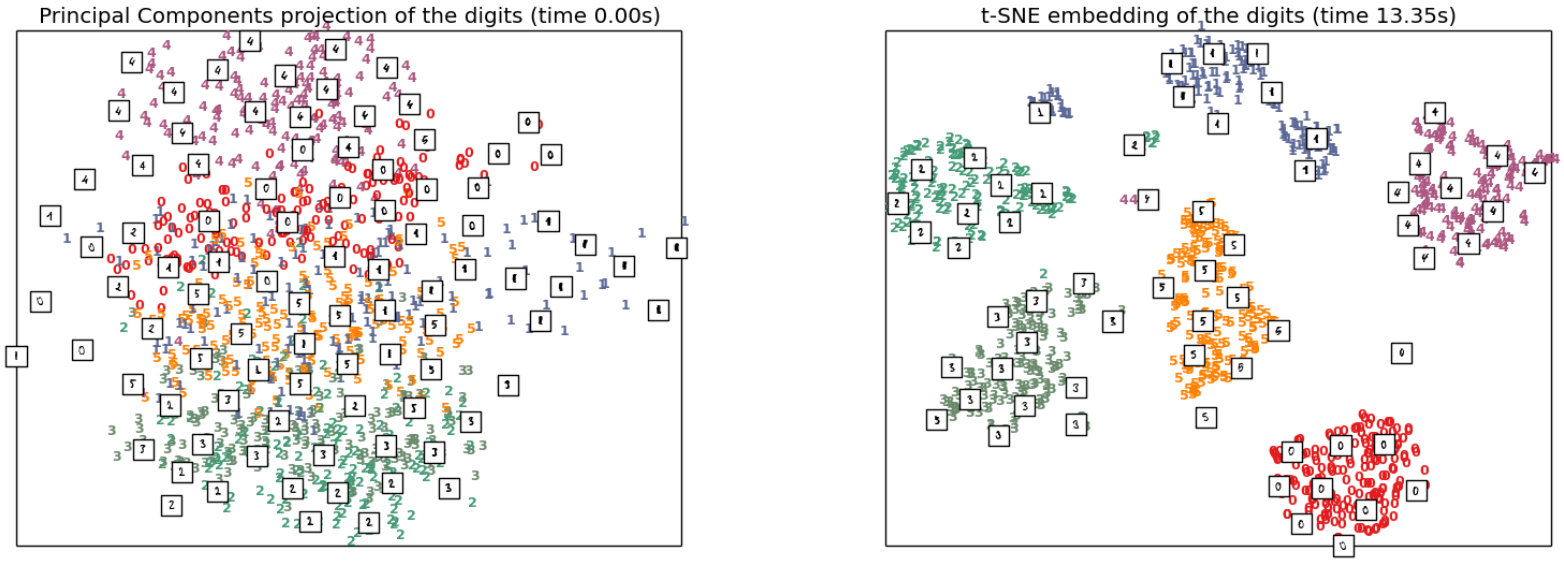
  with $q_{i|i} = 0$ by convention.

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)

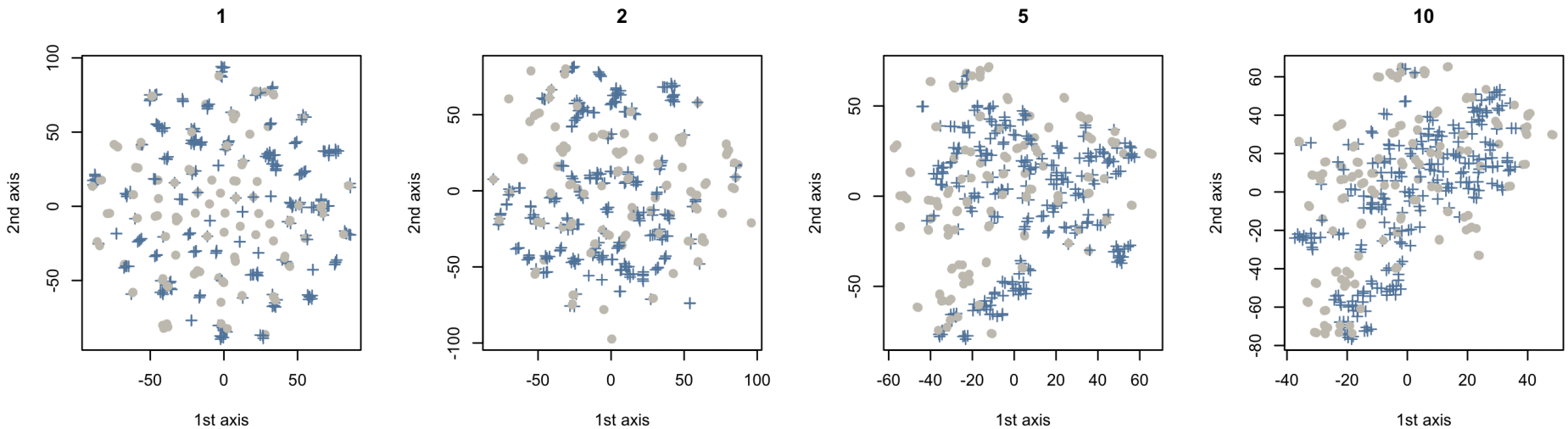


- t-SNE allows to better gather similar observations.
- But, a key parameters may be hard to choose
- The tuneable parameter σ_i which says (loosely) how to balance attention between local and global aspects of the data.
  It is referred implicitly to as the "perplexity" in the solftwares and it is link the the number of close neighbors each point has.
  It may be hard to choose!
  For more detailled examples and discussion see
  `https://distill.pub/2016/misread-tsne/`
- Computation time is large.

- Number of citations of the seminal paper : 2600

Other methods: t-SNE, IBD

To capture the structure, it is usually useful to plot the t-SNE result for various values of the perplexity (perp<n).
t-SNE used with the Bray-Curtis distance.

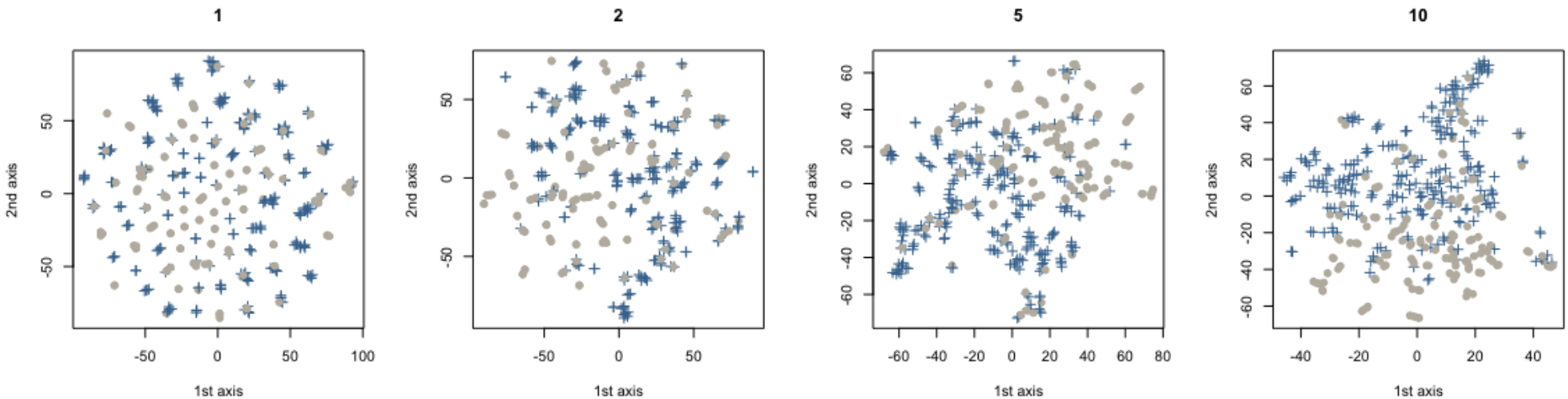


$$BC_{jk} = 1 - \frac{2\sum_{i=1}^{p} min(N_{ij}, N_{ik})}{\sum_{i=1}^{p}(N_{ij} + N_{ik})}$$

where N_{ij} denotes the number of samples of OTU j for patient i.

Other methods: t-SNE, IBD

Now, with the subset of variables selected using a Student test.



Concluding remarks

- There are many algorithms for dimension reduction: PCA, MDS, t-SNE, etc.
  The common idea is to find a representation of the samples in a low dimension space which preserves as much as possible some distance between the points.
- They are used for
  - vizualization of data of high dimension or
  - dimension reduction (pre-processing)
- The most common is to use PCA (or one of its extension).
  PCA leads to a linear approximation of the initial dataset.
- NMF should be prefered when data are positive (images, genes expressions, etc).
  NMF can be interpreted as a "PCA" under constraints.
- MDS and Isomap are useful for data when distances are defined as a neighborhood (microbiota data, ecological data, etc).
  MDS can be interpreted as a "PCA" based on a non euclidean distance.
- t-SNE is a local method which is powerfull but with parameters to choose. It can be used with chosen distances.
  t-SNE can be interpreted as a local "PCA".