DATASCIENCE, LEARNING AND APPLICATIONS

DALAS - EDA

8 février 2024

Laure Soulier - Nicolas Baskiotis

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

Introduction

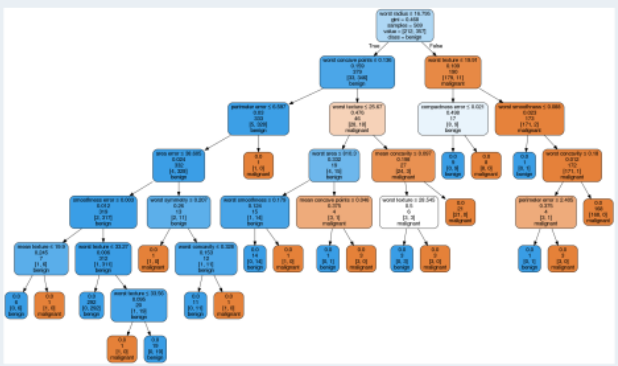
Context

- Data : high number of variables (and individuals)
- Objective : synthesize data
  - Identify relationships between individuals (notion of distance)
  - Identify relationships between variables (notion of correlation)
  - Identify the most informative/important variables
  - Identifying whether classes are well separated

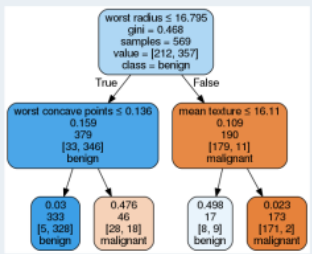
Dimensionality Reduction : Why ?

Interpretability

- Less features leads to better explainability
- Redundant and irrelevant features degrades performance of ML algorithm



Real risk $= 0.92 \pm 0.05$

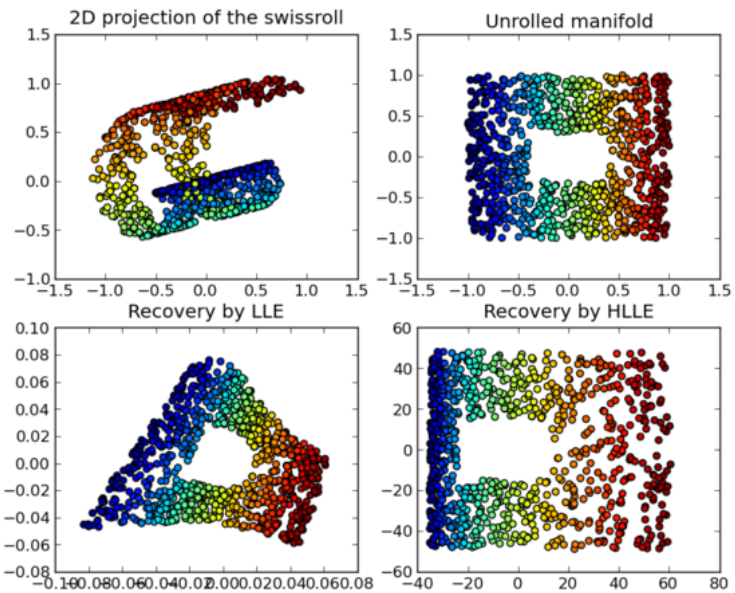


Real risk $= 0.92 \pm 0.06$

Dimensionality Reduction : Why ?

Often data lies on a lower dimensional manifold

- You don't loose information by reducing the number of dimensional
- Curse of Dimensionality :
  $d^{-1/2}(max\|X_i - X_j\| - min\|X_i - X_j\|) \to 0$, $\frac{max\|X_i - X_j\|}{min\|X_i - X_j\|} \to 1$
  when $d \to \infty$ All points are almost equidistant . . .



Dimensionality Reduction : How ?

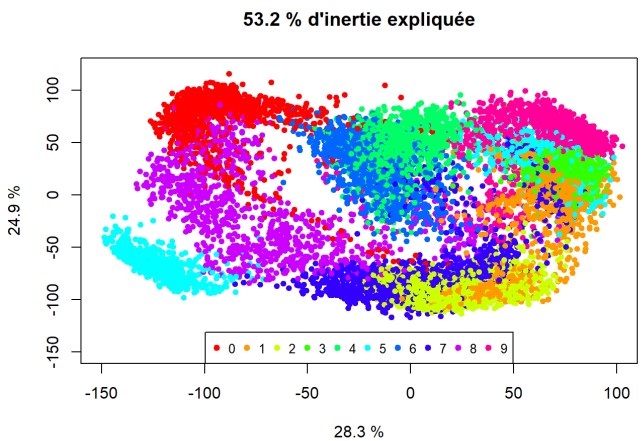
Two main approaches

- Feature selection : keep a subset of the original features
  - The algorithm itself can restrict the features considered (penalization, expressivity)
  - Filtering : appropriate evaluation measures provide the interest of each feature
  - Posthoc : eliminating some features and observe the performance variability
- Build new features !
  - Based on reconstruction error
  - Or try to keep the distances between examples
  - Visualizing data using dimensionality reduction :
    - Linear projections : ACP, AFC, ...
    - Non linear projections : MDS, LLE, t-SNE

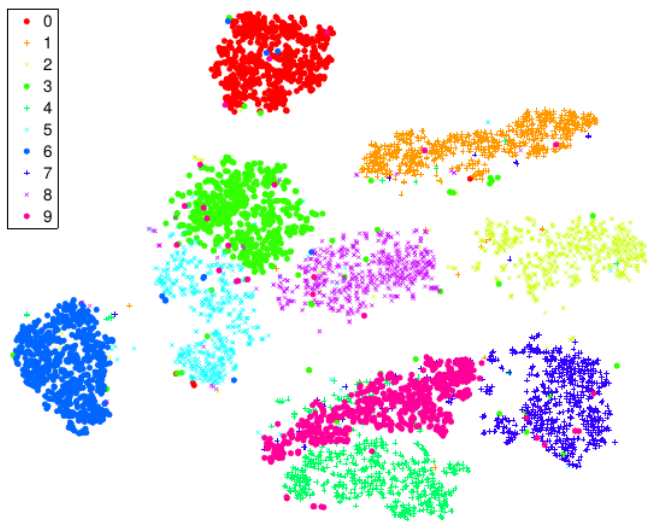
Dimensionality Reduction : Example (MNIST numbers)

SCIENCES SORBONNE UNIVERSITÉ

X1	Y1	X2	Y2	X3	Y3	X4	Y4	X5	Y5	X6	Y6	X7	Y7	X8	Y8
47	100	27	81	57	37	26	0	0	23	56	53	100	90	40	98
0	89	27	100	42	75	29	45	15	15	37	0	69	2	100	6
0	57	31	68	72	90	100	100	76	75	50	51	28	25	16	0
0	100	7	92	5	68	19	45	86	34	100	45	74	23	67	0
0	67	49	83	100	100	81	80	60	60	40	40	33	20	47	0
100	100	88	99	49	74	17	47	0	16	37	0	73	16	20	20



(a) ACP (linear)



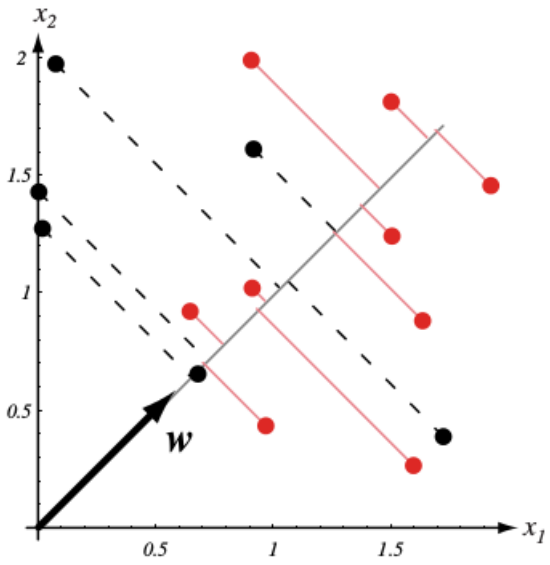
(b) -SNE (non-linear)

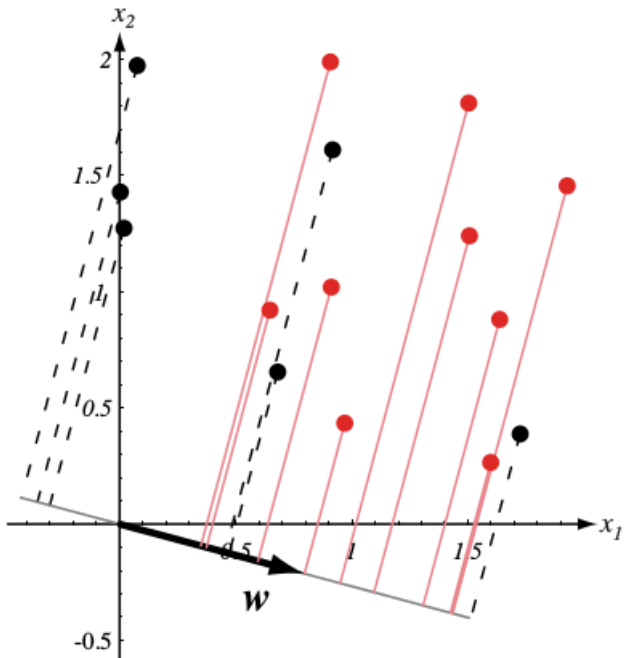
Principal Component Analysis

Principal Component Analysis

Principle :

- Find a new basis with very few dimensions
- Each new dimension is a linear combination of the original dimensions
- Each new dimension has to describe the most information as possible ⇒ Maximization of the variance





PCA : Formalization

Notations

Let's consider a data structure $X \in \mathcal{R}^{n \times p}$ with :

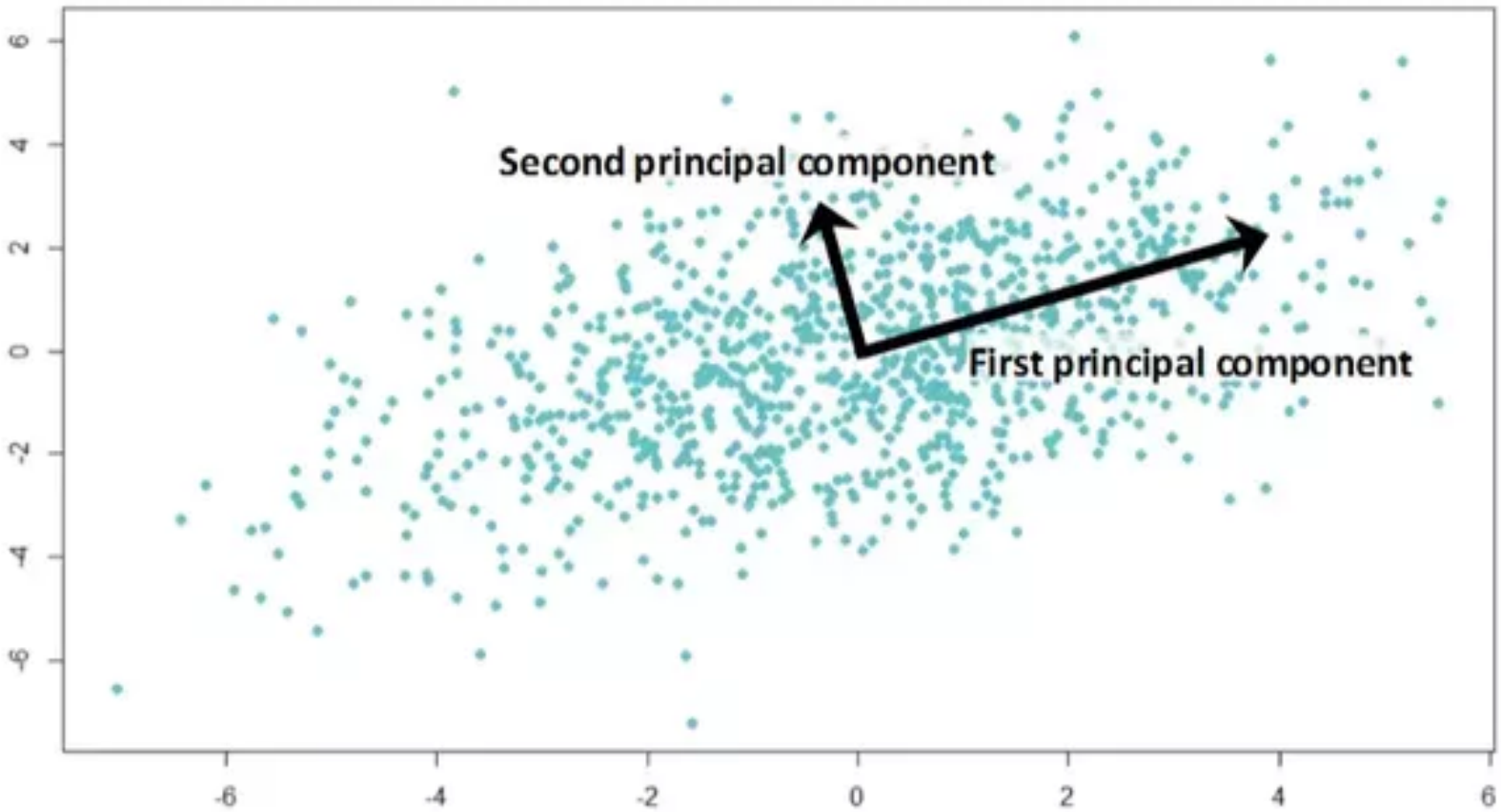
- individuals : $x_1, x_2, ..., x_n$
- variables : $v_1, v_2, ..., v_p$
- Continuous variables and they need to be centered
- if heterogeneous variables, might be reduced (normed ACP) / not mandatory

The objective is to find the vector subspace E_d of rank $d < p$ such that the information loss of the projection of x_i is minimal :

$$\hat{S}_d^* = argmin_{\hat{S}_d} \sum_{i=1}^{n} ||x_i - S_d(\hat{x}_i)||^2$$

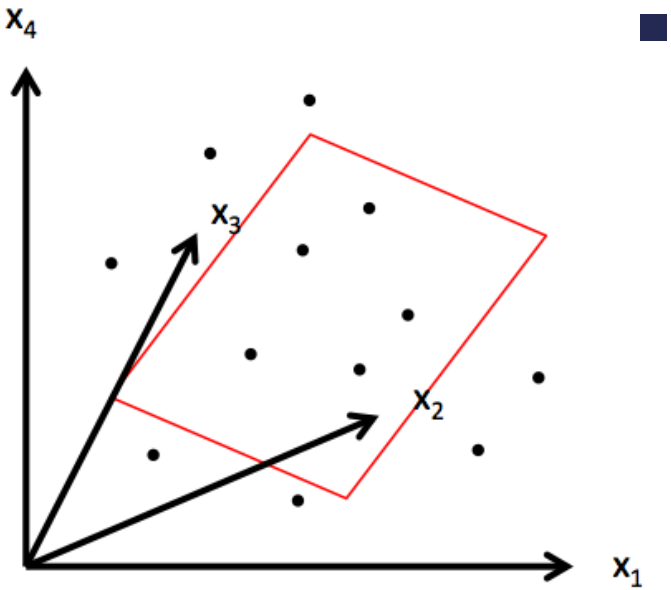
where $S_d(x_i)$ is the orthogonal projection on E_d

PCA : illustration



The first axis allow to minimize projection errors.
The second axis is perpendicular.

PCA : Change of basis



- Linear projection of x_i sur $f_i \in \mathcal{R}^d$
  - $f_i = M^T x_i$ où $M \in \mathcal{R}^{p \times d}$
  - If $d = p$, no dimension reduction, no information loss ($MM^T = Id$) : $f_i = M^T x_i \rightarrow Mf_i = MM^T x_i \rightarrow x_i = Mf_i$
  - If $d < p$, dimension reduction, reconstruction by approximation : $\hat{x}_i = Mf_i$ ou $\hat{x}_i = MM^T x_i$

Objective

Find M which minimizes the squared error
$MSE(M) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - \hat{x}_i||^2$

PCA : Optimization problem

$$
\begin{aligned}
MSE(M) &= \frac{1}{n}\sum_{i=1}^{n}||x_i - \hat{x}_i||^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - MM^T x_i)(x_i - MM^T x_{i,}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^T x_i - 2x_i^T MM^T x_i + x_i^T MM^T MM^T x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_{i,}^T x_i - \frac{1}{n}\sum_{i=1}^{n}x_i^T MM^T x_i
\end{aligned}
$$

We obtain :

$$
\begin{aligned}
\text{argmin } MSE(M) &= \text{argmax}\sum_{i=1}^{n}x_i^T MM^T x_i = \text{argmax } tr(XMM^T X^T) \\
&= \text{argmax } tr(\tilde{X}^T \tilde{X})
\end{aligned}
$$

where $\tilde{X} = XM$, corresponding to the projection of individuals in the subspace.

PCA : Optimization problem

$$
\begin{aligned}
MSE(M) &= \frac{1}{n}\sum_{i=1}^{n}||x_i - \hat{x}_i||^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - MM^T x_i)(x_i - MM^T x_{i,}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^T x_i - 2x_i^T MM^T x_i + x_i^T MM^T MM^T x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_{i,}^T x_i - \frac{1}{n}\sum_{i=1}^{n} x_i^T MM^T x_i
\end{aligned}
$$

We obtain :

$$
\begin{aligned}
\text{argmin } MSE(M) &= \text{argmax} \sum_{i=1}^{n} x_i^T MM^T x_i = \text{argmax } tr(XMM^T X^T) \\
&= \text{argmax } tr(\tilde{X}^T \tilde{X})
\end{aligned}
$$

where $\tilde{X} = XM$, corresponding to the projection of individuals in the subspace.

PCA : Optimization problem

$$
\begin{aligned}
MSE(M) &= \frac{1}{n}\sum_{i=1}^{n}||x_i - \hat{x}_i||^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - MM^T x_i)(x_i - MM^T x_{i,}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^T x_i - 2x_i^T MM^T x_i + x_i^T MM^T MM^T x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_{i,}^T x_i - \frac{1}{n}\sum_{i=1}^{n}x_i^T MM^T x_i
\end{aligned}
$$

We obtain :

$$
\begin{aligned}
\text{argmin } MSE(M) &= \text{argmax}\sum_{i=1}^{n}x_i^T MM^T x_i = \text{argmax } tr(XMM^T X^T) \\
&= \text{argmax } tr(\tilde{X}^T \tilde{X})
\end{aligned}
$$

where $\tilde{X} = XM$, corresponding to the projection of individuals in the subspace.

PCA : Optimization problem

$$
\begin{aligned}
MSE(M) &= \frac{1}{n}\sum_{i=1}^{n}||x_i - \hat{x}_i||^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - MM^T x_i)(x_i - MM^T x_{i,}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^T x_i - 2x_i^T MM^T x_i + x_i^T MM^T MM^T x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_{i,}^T x_i - \frac{1}{n}\sum_{i=1}^{n}x_i^T MM^T x_i
\end{aligned}
$$

We obtain :

$$
\begin{aligned}
argmin\ MSE(M) &= argmax \sum_{i=1}^{n} x_i^T MM^T x_i = argmax\ tr(XMM^T X^T) \\
&= argmax\ tr(\tilde{X}^T \tilde{X})
\end{aligned}
$$

where $\tilde{X} = XM$, corresponding to the projection of individuals in the subspace.

PCA : Optimization problem

$$argmin\ MSE(M) \quad = \quad argmax\ tr(\tilde{X}^T \tilde{X})$$

What is the link with "relationships" between variables ?

Covariance matrix : $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$
On centered data : $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i)(x_i)^T = \frac{1}{n}X^T X$
$\frac{1}{n-1}\tilde{X}^T\tilde{X}$ is an estimate of the covariance matrix between projected data
$\rightarrow tr(\tilde{X}^T\tilde{X})$ is the sum of variances (estimates)

Consequence :

Minimizing MSE(M) $\leftrightarrow$ maximizing the variance of data regarding the projection M.

PCA : Optimization problem

$$argmin\ MSE(M) \quad = \quad argmax\ tr(\tilde{X}^T \tilde{X})$$

What is the link with "relationships" between variables ?

Covariance matrix : $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$
On centered data : $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i)(x_i)^T = \frac{1}{n}X^T X$
$\frac{1}{n-1}\tilde{X}^T\tilde{X}$ is an estimate of the covariance matrix between projected data
$\rightarrow tr(\tilde{X}^T\tilde{X})$ is the sum of variances (estimates)

Consequence :

Minimizing MSE(M) $\leftrightarrow$ maximizing the variance of data regarding the projection M.

PCA : How to find the first axis ?

- Intuition : identify the first **factorial axis** m_1 such that the variance Xm_1 is maximal. The vector $c_1 = Xm_1$ is called a **principal component**.

Let's $M = m_1$, be the first factorial axis

$$\begin{aligned} MSE(M) &= tr(Xm_1 m_1^T X^T) = tr(m_1^T X^T X m_1) = m_1^T X^T X m_1 \\ &= m_1^T \Sigma m_1 \ with \ \Sigma = \frac{1}{n} X^T X \end{aligned}$$

Optimization problem under constraints :

$$argmin_{m_1} MSE(m_1) = -m_1^T \Sigma m_1 \ avec \ m_1^T m_1 = 1$$

PCA : How to find the first axis ?

- Intuition : identify the first **factorial axis** m_1 such that the variance Xm_1 is maximal. The vector $c_1 = Xm_1$ is called a **principal component**.

Let's $M = m_1$, be the first factorial axis

$$\begin{aligned} MSE(M) &= tr(Xm_1 m_1^T X^T) = tr(m_1^T X^T X m_1) = m_1^T X^T X m_1 \\ &= m_1^T \Sigma m_1 \ with \ \Sigma = \frac{1}{n} X^T X \end{aligned}$$

Optimization problem under constraints :

$$argmin_{m_1} MSE(m_1) = -m_1^T \Sigma m_1 \ avec \ m_1^T m_1 = 1$$

PCA : How ot find the first axis ?

$$argmin_{m_1} MSE(m_1) = -m_1^T \Sigma m_1 \ avec \ m_1^T m_1 = 1$$

- Solving with the langragian :

$$\begin{aligned}
\mathcal{L}(m_1, \lambda_1) &= -m_1^T \Sigma m_1 + \lambda_1 (m_1^T m_1 - 1) \\
\nabla_{m_1} \mathcal{L}(m_1, \lambda_1) &= -2\Sigma m_1 + 2\lambda_1 m_1 = 0 \quad \rightarrow \quad \Sigma m_1 = \lambda_1 m_1 \\
&\rightarrow \quad m_1^T \Sigma m_1 = \lambda_1 \\
\nabla_{\lambda_1} \mathcal{L}(m_1, \lambda_1) &= m_1^T m_1 - 1 = 0 \rightarrow ||m_1|| = 1
\end{aligned}$$

Reminder - Eigenvalues and eigenvectors :

An eigenvector X associated with an eigenvalue λ must satisfy the relation $AX = \lambda X$.

PCA : How ot find the first axis ?

$$argmin_{m_1} MSE(m_1) = -m_1^T \Sigma m_1 \ avec \ m_1^T m_1 = 1$$

- Solving with the langragian :

$$\begin{aligned}
\mathcal{L}(m_1, \lambda_1) &= -m_1^T \Sigma m_1 + \lambda_1 (m_1^T m_1 - 1) \\
\nabla_{m_1} \mathcal{L}(m_1, \lambda_1) &= -2\Sigma m_1 + 2\lambda_1 m_1 = 0 \quad \rightarrow \quad \Sigma m_1 = \lambda_1 m_1 \\
&\rightarrow \quad m_1^T \Sigma m_1 = \lambda_1 \\
\nabla_{\lambda_1} \mathcal{L}(m_1, \lambda_1) &= m_1^T m_1 - 1 = 0 \rightarrow ||m_1|| = 1
\end{aligned}$$

Reminder - Eigenvalues and eigenvectors :

An eigenvector X associated with an eigenvalue λ must satisfy the relation $AX = \lambda X$.

PCA : How ot find the first axis ?

$$argmin_{m_1} MSE(m_1) = -m_1^T \Sigma m_1 \ avec \ m_1^T m_1 = 1$$

- Solving with the langragian :

$$\begin{aligned}
\mathcal{L}(m_1, \lambda_1) &= -m_1^T \Sigma m_1 + \lambda_1 (m_1^T m_1 - 1) \\
\nabla_{m_1} \mathcal{L}(m_1, \lambda_1) &= -2\Sigma m_1 + 2\lambda_1 m_1 = 0 \quad \rightarrow \quad \Sigma m_1 = \lambda_1 m_1 \\
&\rightarrow \quad m_1^T \Sigma m_1 = \lambda_1 \\
\nabla_{\lambda_1} \mathcal{L}(m_1, \lambda_1) &= m_1^T m_1 - 1 = 0 \rightarrow ||m_1|| = 1
\end{aligned}$$

Reminder - Eigenvalues and eigenvectors :

An eigenvector X associated with an eigenvalue λ must satisfy the relation $AX = \lambda X$.

PCA : How ot find the first axis ?

Conclusion

- m_1 and λ_1 are respectively eigenvectors and eigenvalues
- $MSE(m_1)$ might be written as : $MSE(m_1) = -\lambda_1$. Consequently, we seek to maximize the eigenvalue
- The first factorial axis is derived from the eigenvector m_1 associated with the largest eigenvalue λ_1 of the covariance matrix Σ.

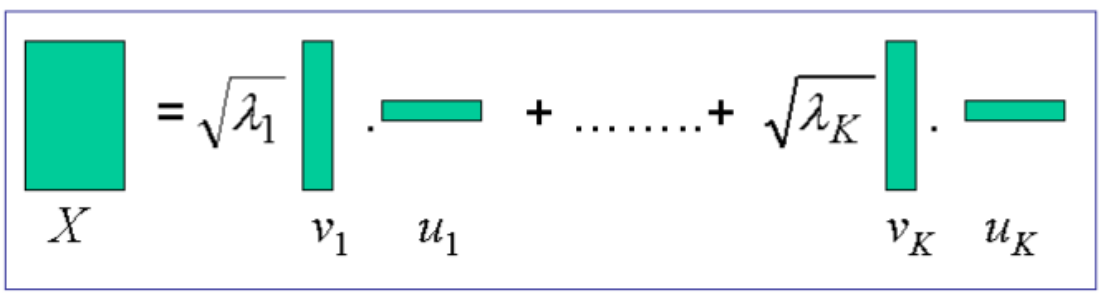
PCA : How ot find the second axis m_2 ?

$$min\ MSE(m_2) = -m_2^T \Sigma m_2$$
$$such\ that m_2^T m_2 = 1 \quad and \quad m_2^T m_1 = 0$$

This amounts to finding the second eigenvalue λ_2 and its associated eigenvector m_2.
$\rightarrow$ Same principle for the third axis, the fourth one, and so on...

Reconstitution of the X matrix from the factorial axes

The change of basis can also be seen as a singular value decomposition of the matrix X :



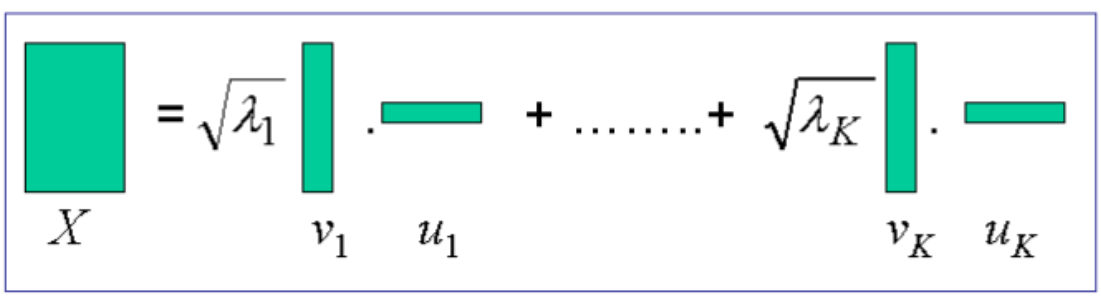
PCA : How ot find the second axis m_2 ?

$$min\ MSE(m_2) = -m_2^T \Sigma m_2$$
$$such\ that m_2^T m_2 = 1 \quad and \quad m_2^T m_1 = 0$$

This amounts to finding the second eigenvalue λ_2 and its associated eigenvector m_2.
$\rightarrow$ Same principle for the third axis, the fourth one, and so on...

Reconstitution of the X matrix from the factorial axes

The change of basis can also be seen as a singular value decomposition of the matrix X :



PCA : In practice...

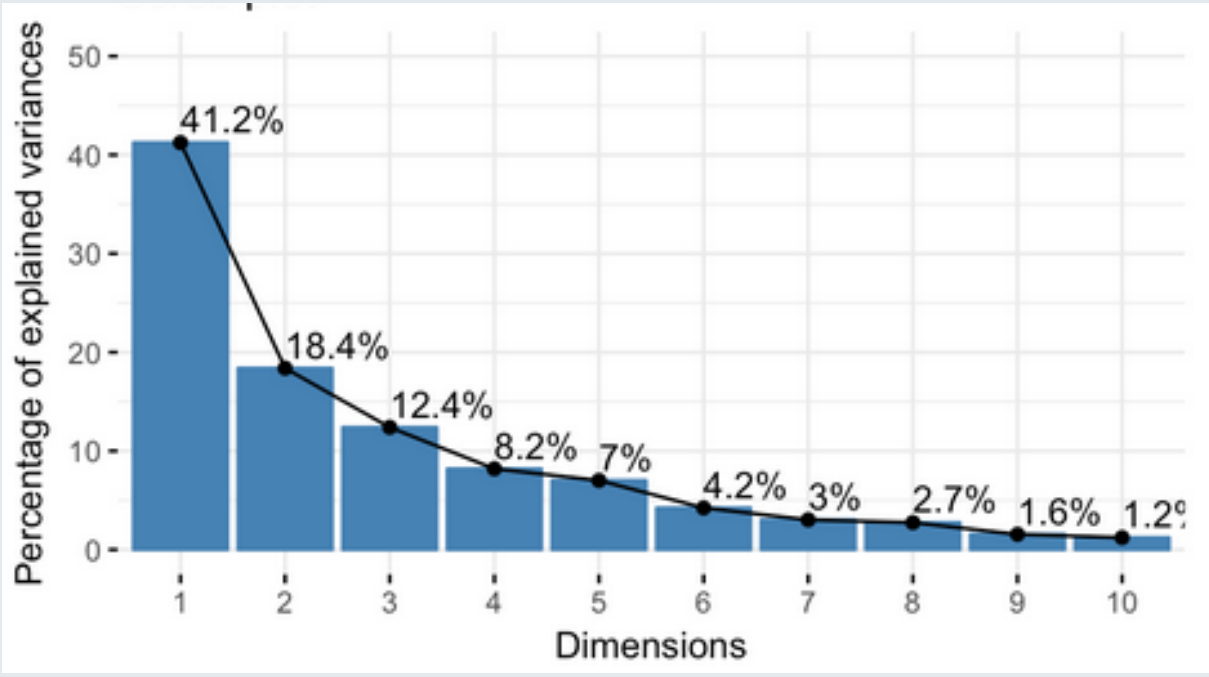
From centered (and possibly reduced) X data

- Estimate the covariance matrix $\Sigma = \frac{1}{n} X^T X$.
- Identify the eigenvalues of Σ
- Order the k eigenvalues in ascending order to form the new M basis
- Project the points to obtain the components : $C = XM$

Principal Component Analysis

For multiple dimensions

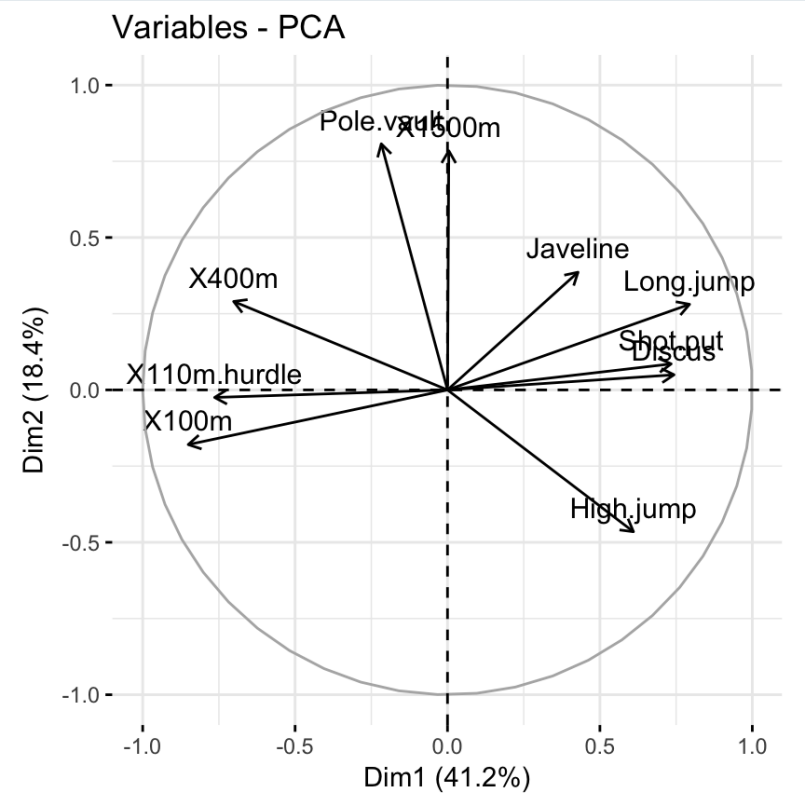
- Optimal choice given by the largest Eigenvalue of the covariance matrix $(X - \mu)(X - \mu')$
- Projected inertia given by the sum of those eigenvalues.
- Often fast decay of the eigenvalues



Principal Component Analysis : Interpretation

Correlation Circle

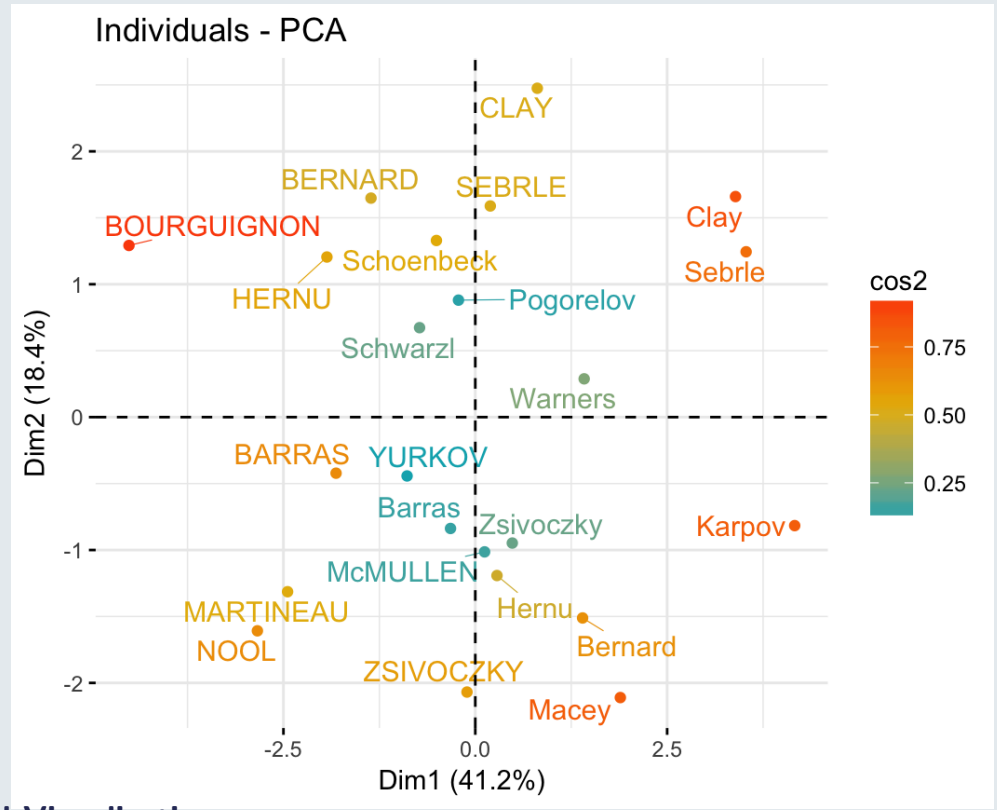
- Positively correlated variables are grouped together.
- Negatively correlated variables are positioned on opposite sides
- Variables that are away from the origin are well represented



Principal Component Analysis : Interpretation

Graph of individuals

- Close individuals tend to be similar
- But you have to look at the other dimensions !
- Those close to the origin are not well represented by the axes.



Analyse en Composantes Principales (ACP) – *Exemple illustratif*

SCIENCES SORBONNE UNIVERSITÉ

- Données

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

$\rightarrow$ Nécessité de centrer les données, mais pas de réduire (variables homogènes : notes)

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

SCIENCES SORBONNE UNIVERSITÉ

- Etape 1 : Matrice Variances-covariances dans l'espace des variables $\Sigma = \frac{1}{n} X^T X$

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Remarques

- Si les données sont centrées réduites, la variance de chaque variable est égale à 1.
- On peut aussi calculer la matrice variances-covariances dans l'espace des individus : $\Sigma = \frac{1}{n} X X^T$

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

- Etape 2 : Estimation des valeurs propres

Facteur	λ	inertie	cumul
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1
4	0.01	0.00	1.00

L'inertie mesure le pourcentage de dispersion des points autour de l'axe factoriel. $inertie_k = \frac{\lambda_k}{\sum_{l=1}^{K} \lambda_l}$

Remarques

- Les valeurs propres de $\frac{1}{n}X^TX$ et de $\frac{1}{n}XX^T$ sont égales $\rightarrow$ Rechercher la meilleure représentation des individus équivaut à rechercher la meilleure représentation des variables
- Critère de choix des axes principaux : inertie cumulée $> 80\%$, $\lambda_k > 1$ (règle de Kaiser), coude de la courbe (éboulis), ...

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

- Etape 2 : Estimation des valeurs propres

Facteur	λ	inertie	cumul
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1
4	0.01	0.00	1.00

L'inertie mesure le pourcentage de dispersion des points autour de l'axe factoriel. $inertie_k = \frac{\lambda_k}{\sum_{l=1}^{K} \lambda_l}$

Remarques

- Les valeurs propres de $\frac{1}{n}X^TX$ et de $\frac{1}{n}XX^T$ sont égales $\rightarrow$ Rechercher la meilleure représentation des individus équivaut à rechercher la meilleure représentation des variables
- Critère de choix des axes principaux : inertie cumulée $> 80\%$, $\lambda_k > 1$ (règle de Kaiser), coude de la courbe (éboulis), ...

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

SCIENCES SORBONNE UNIVERSITÉ

- Etape 2 : Estimation des valeurs propres

Facteur	λ	inertie	cumul
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1
4	0.01	0.00	1.00

L'inertie mesure le pourcentage de dispersion des points autour de l'axe factoriel. $inertie_k = \frac{\lambda_k}{\sum_{l=1}^{K} \lambda_l}$

Remarques

- Les valeurs propres de $\frac{1}{n}X^TX$ et de $\frac{1}{n}XX^T$ sont égales $\rightarrow$ Rechercher la meilleure représentation des individus équivaut à rechercher la meilleure représentation des variables
- Critère de choix des axes principaux : inertie cumulée $> 80\%$, $\lambda_k > 1$ (règle de Kaiser), coude de la courbe (éboulis), ...

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

SCIENCES SORBONNE UNIVERSITÉ

- Etape 2 : Estimation des valeurs propres

Facteur	λ	inertie	cumul
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1
4	0.01	0.00	1.00

L'inertie mesure le pourcentage de dispersion des points autour de l'axe factoriel. $inertie_k = \frac{\lambda_k}{\sum_{l=1}^{K} \lambda_l}$

Remarques

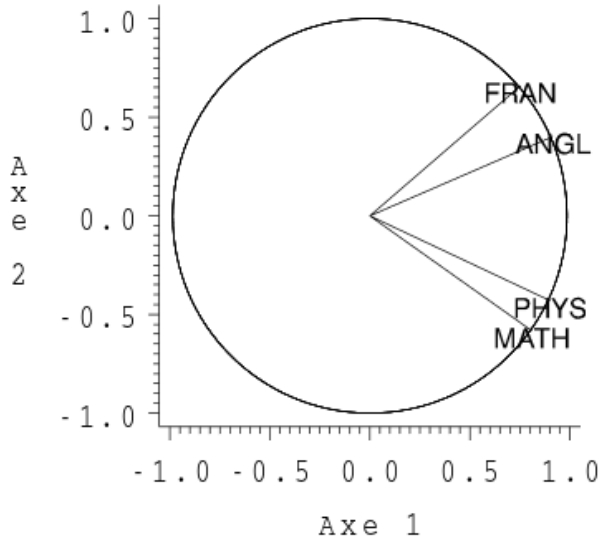
- Les valeurs propres de $\frac{1}{n}X^TX$ et de $\frac{1}{n}XX^T$ sont égales $\rightarrow$ Rechercher la meilleure représentation des individus équivaut à rechercher la meilleure représentation des variables
- Critère de choix des axes principaux : inertie cumulée $> 80\%$, $\lambda_k > 1$ (règle de Kaiser), coude de la courbe (éboulis), ...

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

- Etape 3 : Projection des individus et variables
  - Corrélation des variables avec les axes factoriels

Corrélations variables-facteurs

FACTEURS -->	F1	F2	F3	F4
MATH	0.81	-0.58	0.01	-0.02
PHYS	0.90	-0.43	-0.03	0.02
FRAN	0.75	0.66	-0.02	-0.01
ANGL	0.91	0.40	0.05	0.01

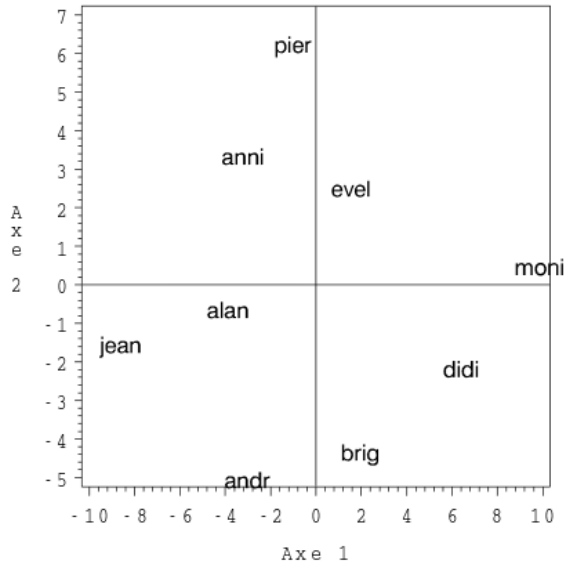


Analyse en Composantes Principales (ACP) – *Exemple illustratif*

- Etape 3 : Projection des individus et variables
  - Projection des individus $C = XM$

Coordonnées des individus ; contributions ; cosinus carrés

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	0.11	-8.61	-1.41	20.99	29.19	1.83	0.97	0.03
alan	0.11	-3.88	-0.50	4.22	5.92	0.23	0.98	0.02
anni	0.11	-3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	0.11	6.41	-2.05	12.48	16.15	3.87	0.91	0.09
andr	0.11	-3.03	-4.92	9.22	3.62	22.37	0.28	0.72
pier	0.11	-1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	0.11	1.95	-4.20	5.93	1.50	16.29	0.18	0.82
evel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73



Remarques

- On peut mesure la contribution d'un point à l'inertie d'un nuage :

$$contrib_i = \frac{w_i \sum_{k=1}^{K} (c_i^k)^2}{\sum_{k=1}^{K} \lambda_k} \quad (1)$$

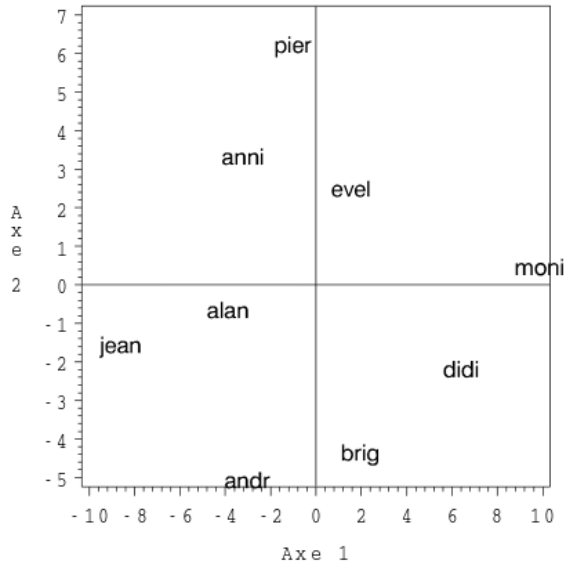
Analyse en Composantes Principales (ACP) – *Exemple illustratif*

SCIENCES SORBONNE UNIVERSITÉ

- Etape 3 : Projection des individus et variables
  - Projection des individus $C = XM$

Coordonnées des individus ; contributions ; cosinus carrés

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	0.11	-8.61	-1.41	20.99	29.19	1.83	0.97	0.03
alan	0.11	-3.88	-0.50	4.22	5.92	0.23	0.98	0.02
anni	0.11	-3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	0.11	6.41	-2.05	12.48	16.15	3.87	0.91	0.09
andr	0.11	-3.03	-4.92	9.22	3.62	22.37	0.28	0.72
pier	0.11	-1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	0.11	1.95	-4.20	5.93	1.50	16.29	0.18	0.82
evel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73



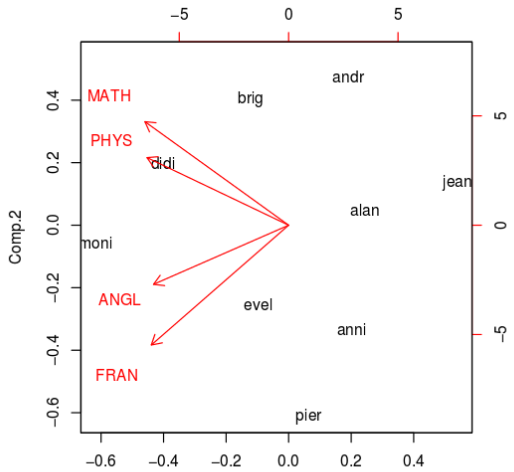
Remarques

- On peut mesure la contribution d'un point à l'inertie d'un nuage :

$$contrib_i = \frac{w_i \sum_{k=1}^{K} (c_i^k)^2}{\sum_{k=1}^{K} \lambda_k} \quad (1)$$

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

- Etape 3 : Projection des individus et variables (biplot)
  - Projection des individus et variables

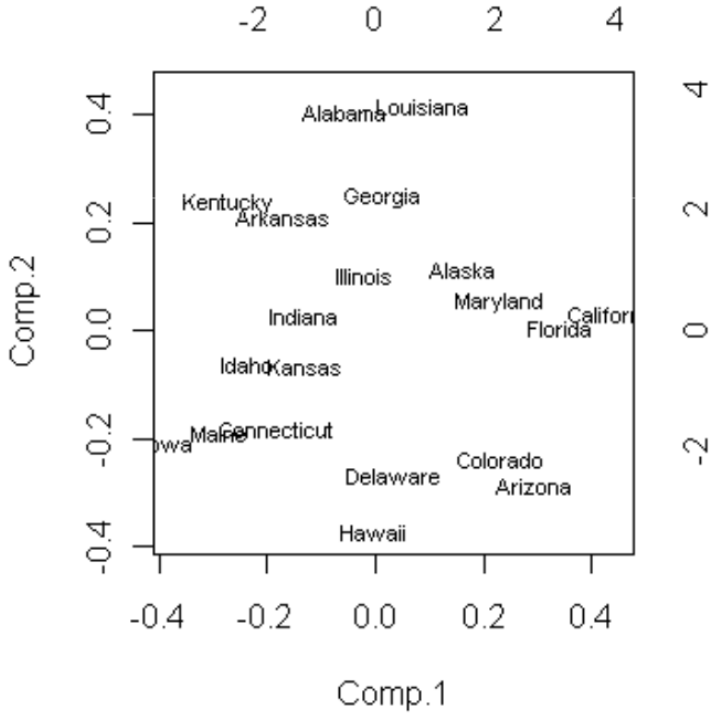


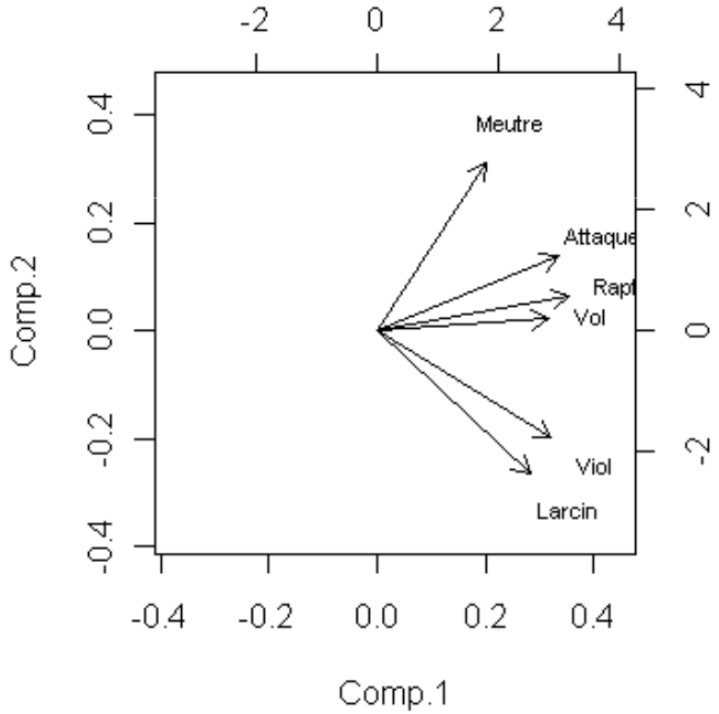
Interprétation

- Deux individus proches se ressemblent
- Deux variables très corrélées positivement sont du même côté sur un axe
- Un individu sera proche des variables pour lesquelles il a de fortes valeurs (et inversement)
- Plus les valeurs d'un individu pour une variable sont fortes, plus il sera éloigné de l'origine de l'axe factoriel.

Analyse en Composantes Principales (ACP) – *Exemple illustratif*

- Etudes des crimes aux USA





Les différentes analyses factorielles

ACP : données quanti, continues, a priori corrélées entre elles
AFC : tableau de contingence (croisement de variables quali)
ACM : données quali (extension à plusieurs variables)
AFCM : données quanti et quali
AFM : variables structurées en groupe
AFMH : variables structurées en hiérarchie

Analyse Factorielle des Correspondances (AFC)

Analyse Factorielle des Correspondances (AFC)

SCIENCES SORBONNE UNIVERSITÉ

- Données
  - Deux variables qualitatives (tableau de contingence)

Yeux \ Cheveux	Brun	Châtain	Roux	Blond	Total
Marron	68	119	26	7	220
Noisette	15	54	14	10	93
Vert	5	29	14	16	64
Bleu	20	84	17	94	215
Total	108	286	71	127	592

On note x_{ij} les éléments du tableau de contingence, $x_{i.}$ le total d'une ligne i et $x_{.j}$ le total d'une colonne j.

- Profils-lignes $x'_{ij} = \frac{x_{ij}}{x_{i.}}$ et profils-colonnes $x''_{ij} = \frac{x_{ij}}{x_{.j}}$

	Brun	Châtain	Roux	Blond	Total
Marron	0,31	0,54	0,12	0,3	1
Noisette	0,16	0,58	0,15	0,11	1
Vert	0,8	0,45	0,22	0,25	1
Bleu	0,9	0,39	0,8	0,44	1
Profil moyen	0,18	0,48	0,12	0,22	1

	Brun	Châtain	Roux	Blond	Profil moyen
Marron	0,63	0,42	0,37	0,6	0,37
Noisette	0,14	0,19	0,2	0,8	0,16
Vert	0,5	0,1	0,2	0,13	0,11
Bleu	0,19	0,29	0,24	0,74	0,36
Total	1	1	1	1	1

Analyse Factorielle des Correspondances (AFC)

SCIENCES SORBONNE UNIVERSITÉ

- Données
  - Deux variables qualitatives (tableau de contingence)

Yeux \ Cheveux	Brun	Châtain	Roux	Blond	Total
Marron	68	119	26	7	220
Noisette	15	54	14	10	93
Vert	5	29	14	16	64
Bleu	20	84	17	94	215
Total	108	286	71	127	592

On note x_{ij} les éléments du tableau de contingence, $x_{i.}$ le total d'une ligne i et $x_{.j}$ le total d'une colonne j.

- Profils-lignes $x'_{ij} = \frac{x_{ij}}{x_{i.}}$ et profils-colonnes $x''_{ij} = \frac{x_{ij}}{x_{.j}}$

	Brun	Châtain	Roux	Blond	Total
Marron	0,31	0,54	0,12	0,3	1
Noisette	0,16	0,58	0,15	0,11	1
Vert	0,8	0,45	0,22	0,25	1
Bleu	0,9	0,39	0,8	0,44	1
Profil moyen	0,18	0,48	0,12	0,22	1

	Brun	Châtain	Roux	Blond	Profil moyen
Marron	0,63	0,42	0,37	0,6	0,37
Noisette	0,14	0,19	0,2	0,8	0,16
Vert	0,5	0,1	0,2	0,13	0,11
Bleu	0,19	0,29	0,24	0,74	0,36
Total	1	1	1	1	1

Analyse Factorielle des Correspondances (AFC)

SCIENCES SORBONNE UNIVERSITÉ

- Données
  - Deux variables qualitatives (tableau de contingence)

Yeux \ Cheveux	Brun	Châtain	Roux	Blond	Total
Marron	68	119	26	7	220
Noisette	15	54	14	10	93
Vert	5	29	14	16	64
Bleu	20	84	17	94	215
Total	108	286	71	127	592

On note x_{ij} les éléments du tableau de contingence, $x_{i.}$ le total d'une ligne i et $x_{.j}$ le total d'une colonne j.

- Profils-lignes $x'_{ij} = \frac{x_{ij}}{x_{i.}}$ et profils-colonnes $x''_{ij} = \frac{x_{ij}}{x_{.j}}$

	Brun	Châtain	Roux	Blond	Total
Marron	0,31	0,54	0,12	0,3	1
Noisette	0,16	0,58	0,15	0,11	1
Vert	0,8	0,45	0,22	0,25	1
Bleu	0,9	0,39	0,8	0,44	1
Profil moyen	0,18	0,48	0,12	0,22	1

	Brun	Châtain	Roux	Blond	Profil moyen
Marron	0,63	0,42	0,37	0,6	0,37
Noisette	0,14	0,19	0,2	0,8	0,16
Vert	0,5	0,1	0,2	0,13	0,11
Bleu	0,19	0,29	0,24	0,74	0,36
Total	1	1	1	1	1

Analyse Factorielle des Correspondances (AFC)

SCIENCES SORBONNE UNIVERSITÉ

- Données
  - Deux variables qualitatives (tableau de contingence)

Yeux \ Cheveux	Brun	Châtain	Roux	Blond	Total
Marron	68	119	26	7	220
Noisette	15	54	14	10	93
Vert	5	29	14	16	64
Bleu	20	84	17	94	215
Total	108	286	71	127	592

On note x_{ij} les éléments du tableau de contingence, $x_{i.}$ le total d'une ligne i et $x_{.j}$ le total d'une colonne j.

- Profils-lignes $x'_{ij} = \frac{x_{ij}}{x_{i.}}$ et profils-colonnes $x''_{ij} = \frac{x_{ij}}{x_{.j}}$

	Brun	Châtain	Roux	Blond	Total
Marron	0,31	0,54	0,12	0,3	1
Noisette	0,16	0,58	0,15	0,11	1
Vert	0,8	0,45	0,22	0,25	1
Bleu	0,9	0,39	0,8	0,44	1
Profil moyen	0,18	0,48	0,12	0,22	1

	Brun	Châtain	Roux	Blond	Profil moyen
Marron	0,63	0,42	0,37	0,6	0,37
Noisette	0,14	0,19	0,2	0,8	0,16
Vert	0,5	0,1	0,2	0,13	0,11
Bleu	0,19	0,29	0,24	0,74	0,36
Total	1	1	1	1	1

source : http://www.irisa.fr/dream/Seminaire/Tahiti/Tahiti04/transparents/emmanuel/emmanuel.pdf

Analyse Factorielle des Correspondances (AFC)

- Objectif
  - Analyser la liaison entre deux variables : la liaison entre deux variables est grande si les profils-lignes ou colonnes sont différents.
    - Quelles sont les lignes qui se ressemblent ? sont différentes ?
    - Existe-t-il des groupes homogènes entre les lignes ? entre les colonnes ?

Principe général

Une AFC est l'équivalent d'une ACP sur les profils-lignes ou profils colonnes :

- Lignes et colonnes ont les mêmes rôles
- Analyse de la distance entre profils
- Inertie du nuage de points exprime l'indépendance entre les deux variables

Analyse Factorielle des Correspondances (AFC)

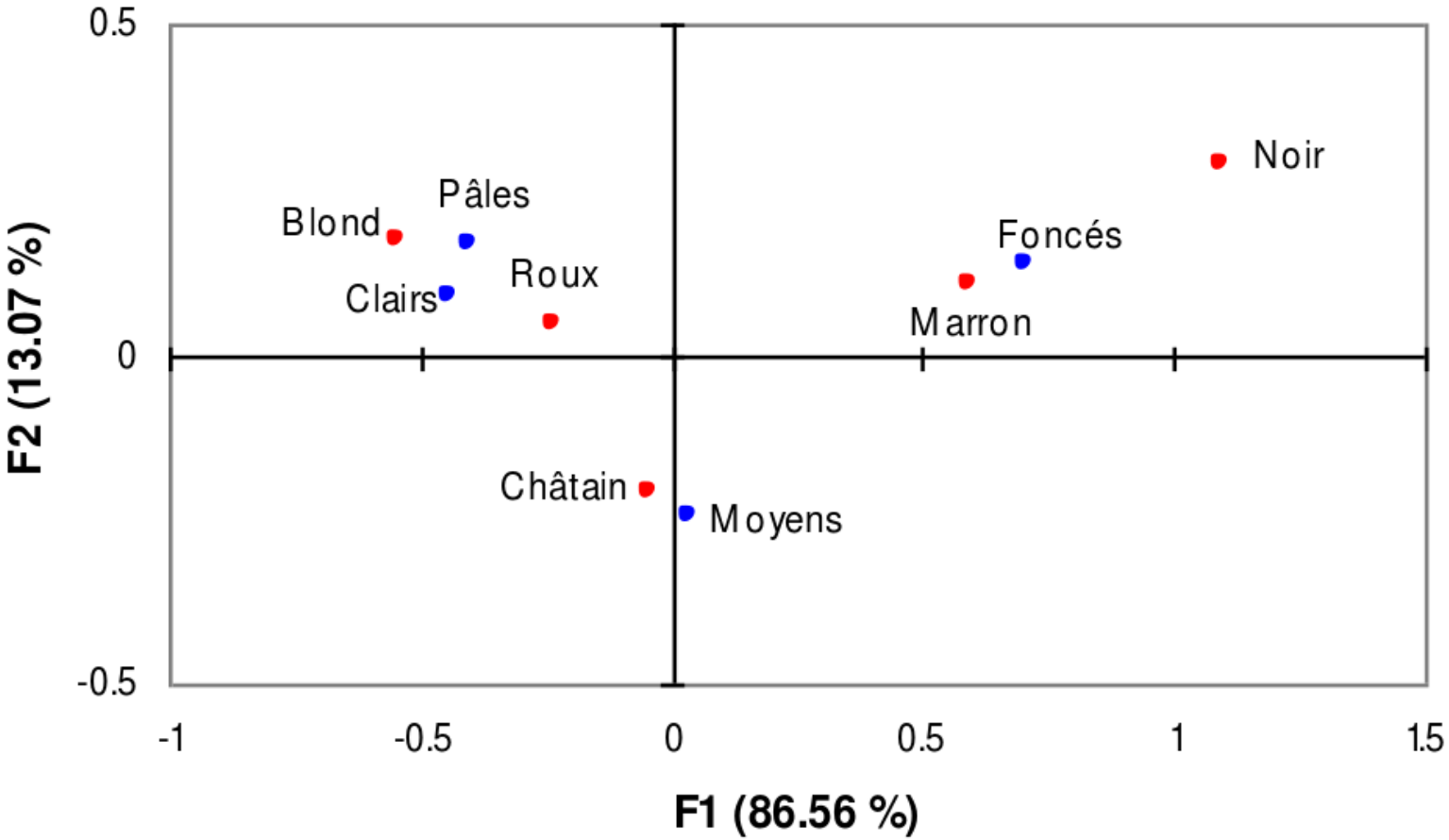
- Objectif
  - Analyser la liaison entre deux variables : la liaison entre deux variables est grande si les profils-lignes ou colonnes sont différents.
    - Quelles sont les lignes qui se ressemblent ? sont différentes ?
    - Existe-t-il des groupes homogènes entre les lignes ? entre les colonnes ?

Principe général

Une AFC est l'équivalent d'une ACP sur les profils-lignes ou profils colonnes :

- Lignes et colonnes ont les mêmes rôles
- Analyse de la distance entre profils
- Inertie du nuage de points exprime l'indépendance entre les deux variables

Analyse Factorielle des Correspondances (AFC)



Analyse des Correspondances Multiples (ACM)

Analyse des Correspondances Multiples (ACM)

- Données
  - p variables qualitatives (par exemple QCM)

individu	bac	âge	durée
1	C	>19	3
2	D	<18	2
...			

  - Transformé en tableau de Burt ("Grand tableau de contingence")

	bacC	bacD	< 18	18ans	19ans	> 19	2ans	3ans	4ans
bacC	583	0	108	323	114	38	324	192	67
bacD	0	214	25	97	68	24	76	82	56
< 18	108	25	133	0	0	0	84	35	14
18ans	323	97	0	420	0	0	224	137	59
19ans	114	68	0	0	182	0	73	75	34
> 19	38	24	0	0	0	62	19	27	16
2ans	324	76	84	224	73	19	400	0	0
3ans	192	82	35	137	75	27	0	274	0
4ans	67	56	14	59	34	16	0	0	123

Principe général

Une ACM est l'équivalent d'une AFC sur un tableau de Burt

Analyse des Correspondances Multiples (ACM)

Non-linear projections and distance preservation

Non-linear projections

- Projecting axes can distort data too much
- Data might be on variety
- Linear projections are not adapted

Non-linear projections

- Multidimensional positioning (Multidimensional Scaling, MDS) : preserves distances
- Isomap : preserves geodesic distances
- Locally linear representations (LLE - embedding) : preserves linear relations
- Laplacian eigenmaps (graph theory)
- Kohonen Maps (self-organized maps) : projection on a defined structure
- Auto-encoder (neural network) : information compression
- t-SNE (representation following a student distribution) : aligning distribution probabilities
- ....

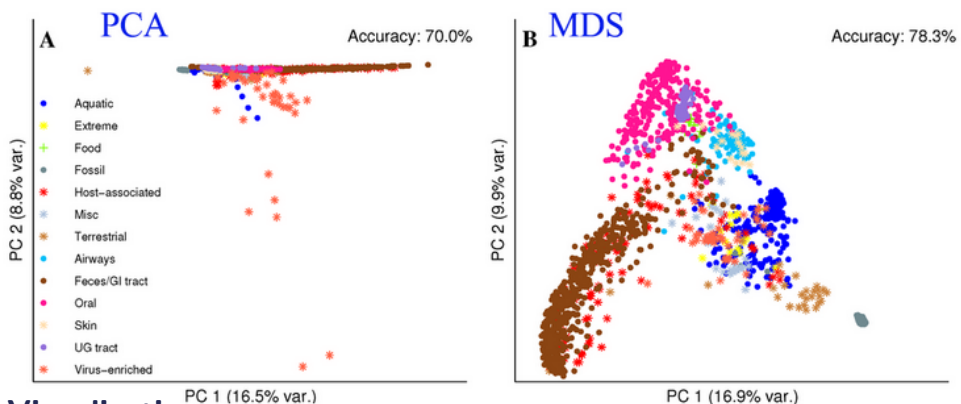
Multi-dimensional Scaling (MDS)

Goal :

- Find projection of the data in order to preserve pairwise distance of the points

$$S(\tilde{x_1}, \tilde{x_2}, ..., \tilde{x_N}) = \sum_{i \neq j} (d(x_i, x_j) - ||\tilde{x_i} - \tilde{x_j}||)^2 \quad (2)$$

- Does not compute explicitly new features
- Used essentially for visualization
- Does not need explicitly the features of the examples, only the pairwise distance



Multi-dimensional Scaling (MDS)

Principle :

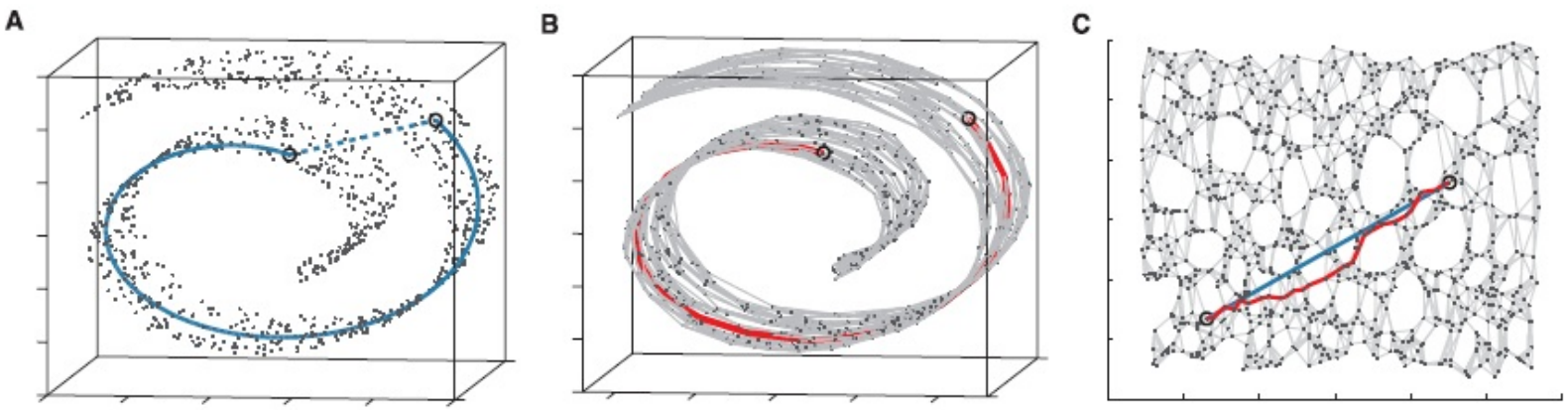
- Gram Matrix : $G = (X - m)(X - m)^T$ with $m = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$
- New coordinates : $x' = U^T(x - m)$ with U orthonormal
- Objective : Minimize the inner products between the feature vectors $\sum_{i=1}^{N}\sum_{j=1}^{N}(\langle x_i - m, x_j - m\rangle - \langle x'_i, x'_j\rangle)^2$
- Equivalent to minimize : $\sum_{i=1}^{N}\sum_{j=1}^{N}(\langle x_i - m, x_j - m\rangle - (x_i - m)^T UU^T(x_j - m))^2$ wrt $U^T(x_i - m)$
- Solution given by the eigenvector of G
- Very similar to PCA but working on XX^T, not on X^TX

Drawbacks :

- Takes into account all the dimensions (noise, correlated, ...)
- Work only if $N << d$ and if d is small ...

Isomap

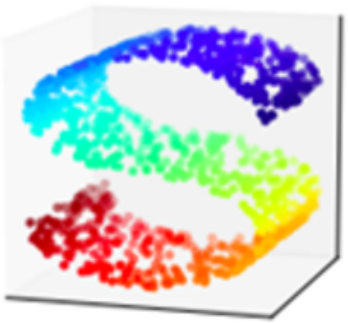
- Euclidean distance is not always the *right* distance
- Try to preserve the geodesic distance of the data
- By following a path through the data

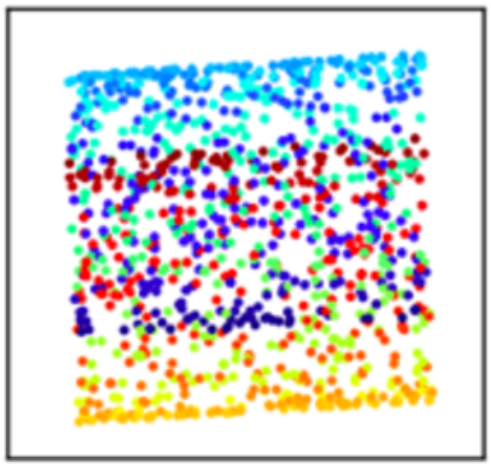


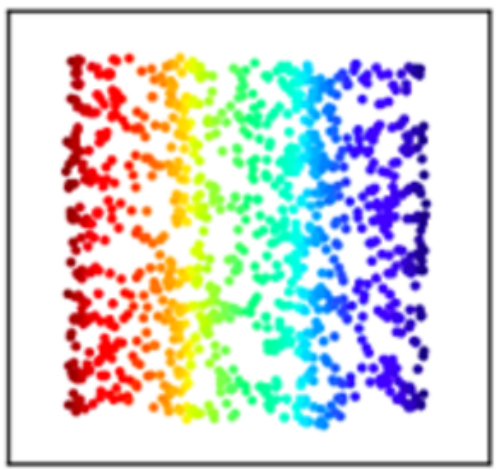
Isomap

Principle :

- Compute the neighborhood graph (using K-NN for instance)
- Compute the distance between pair of points in the graph (shortest weighted path between the points - geodesic distance)
- Apply a MDS on the result







t-SNE

Objective

- Find a mapping which preserve the neighborhood of each point
- No explicit mapping, just the new coordinates
- No assumption on the original space, only the pairwise distance
- Comparison of distribution probabilities of neighborhood defined
  - in the original space
  - in the projected space

t-SNE

Principle

- Define $P_{j|i} = \frac{e^{-\|x_i - x_j\|^2/2\sigma_i^2}}{\sum_k e^{-\|x_i - x_j\|^2/2\sigma^2}}$, the probability that x_i chooses x_j as neighbor
- Find new coordinates x_i' such that $Q_{j|i}$ the probability distribution in the new space is close to $P_{j|i}$.
- Use a Kullback-Leibler distance to match the distributions and a gradient descent algorithm.

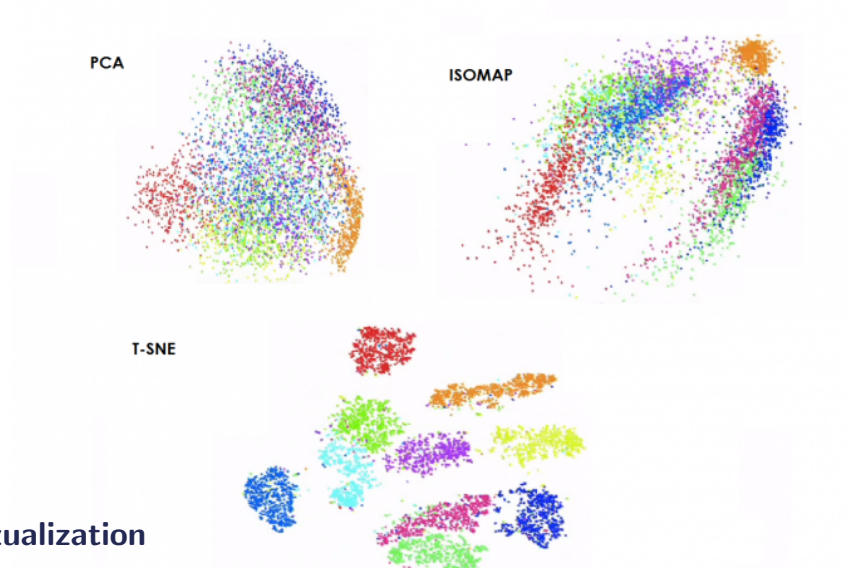
t-SNE

Interpretation

- Local neighborhood is conserved : close points are close in the original space
- But distances are not preserved !

$\Rightarrow$ outside a small region, distances are meaningless !

- Be careful : by playing with the hyperparameter, it is very easy to construct clusters, but often meaningless !



Conclusion

- A tool for understanding data
  - Identify classes with multiple modes
  - Abbreviated points
  - Anticipating difficulties (or facilities)
- A tool for analyzing model errors... Then improve the models
  - Presenting results
  - Understanding errors

WARING

In 2D, a lot of information is lost, TSNE (among others) is a stochastic algorithm, etc...
$\rightarrow$ Be aware of what you see and check it carefully before drawing conclusions !