BI = Business Intelligence
Master Data-Science
Cours 7 - Visualisation

Ludovic DENOYER - ludovic.denoyer@lip6.fr
Laure SOULIER -laure.soulier@lip6.fr

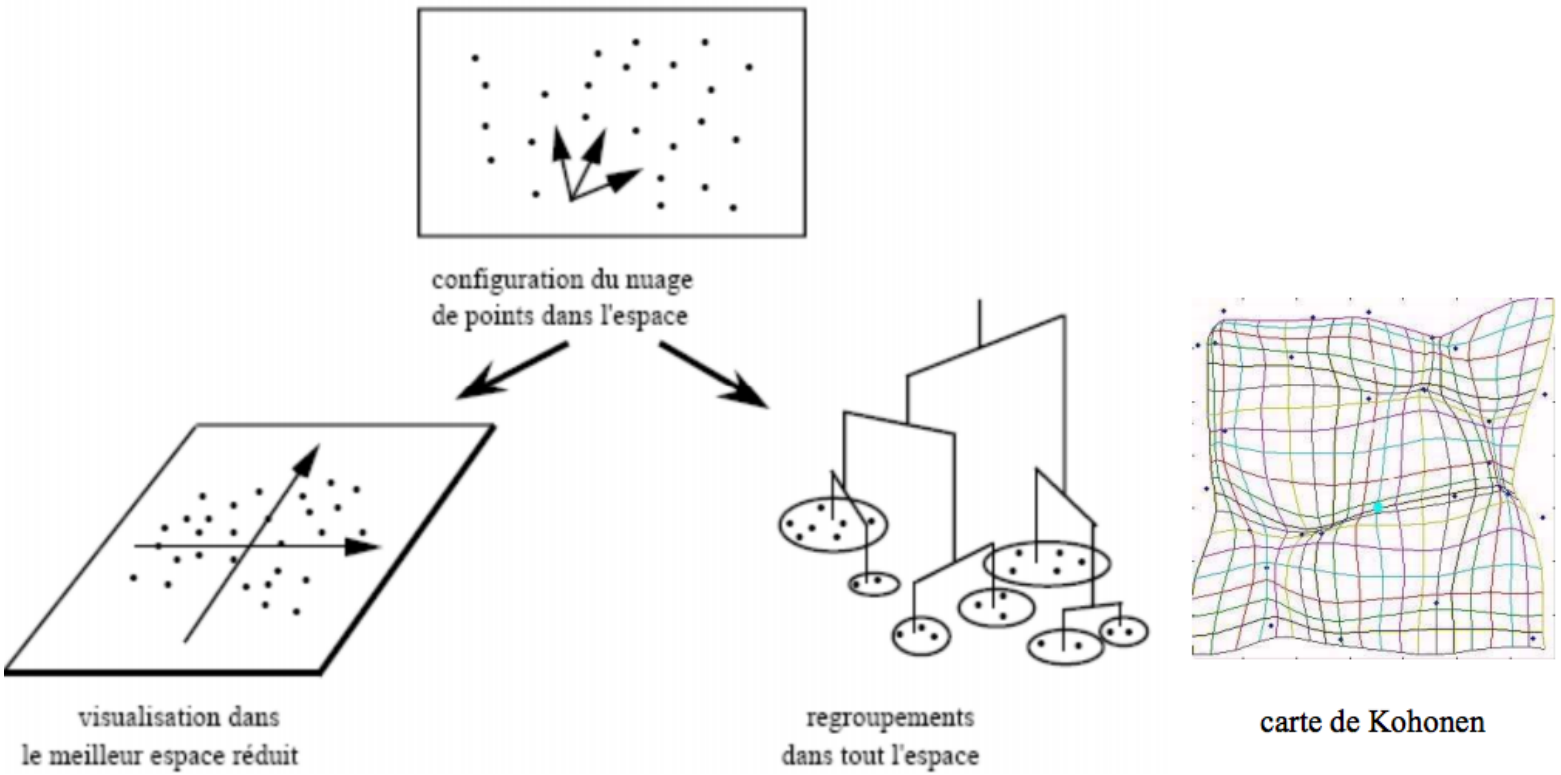
UPMC

28 février 2017

Contexte et principe général

Contexte - Rappel

Analyse descriptive des données : identifier/synthétiser les informations présentes mais cachées dans un gros volume de données



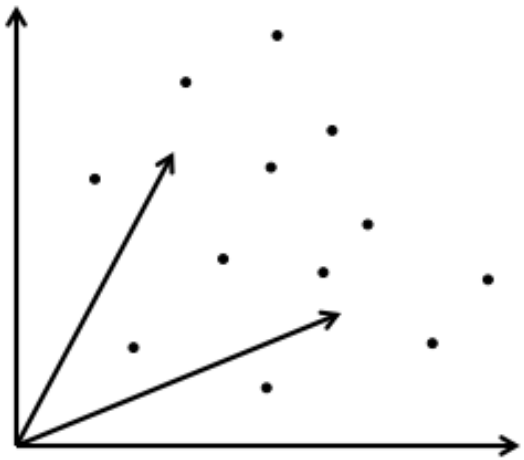
Source :

Lebart-Morineau-Piron, Statistique exploratoire multidimensionnelle

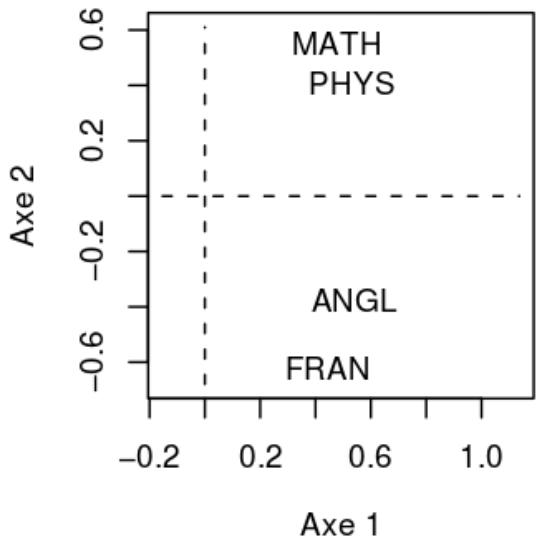
Contexte

- Données : grand nombre de variables et d'individus
- Objectif : synthétiser les données par **réduction de dimension** pour
  - Identifier les variables les plus informatives
  - Identifier les relations entre variables (notion de corrélation)
  - Identifier les relations entre individus (notion de distance)

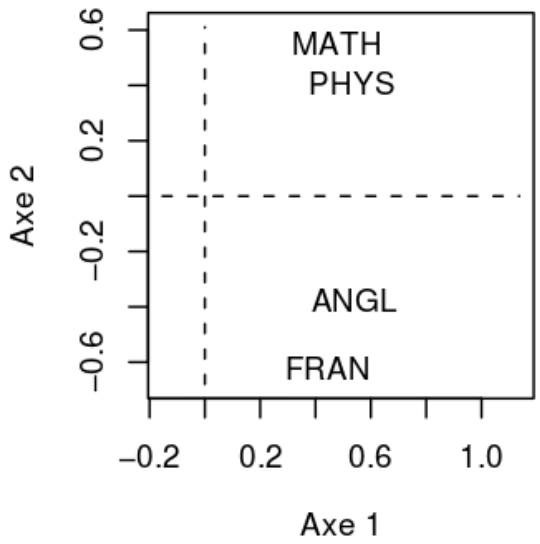
	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

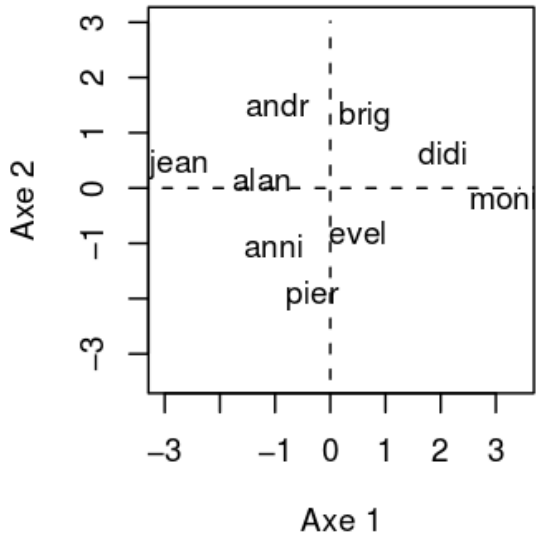


Contexte

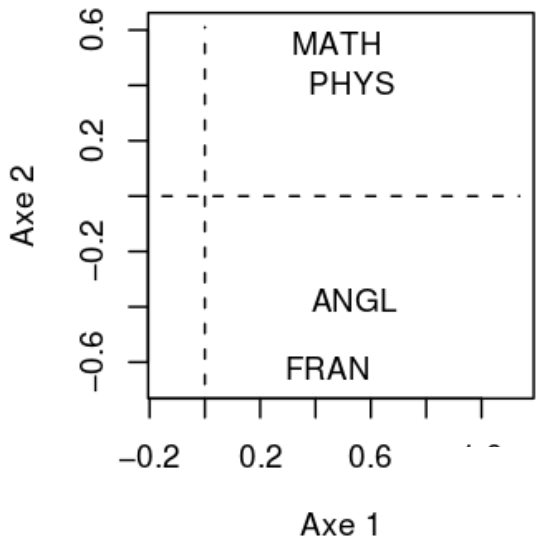


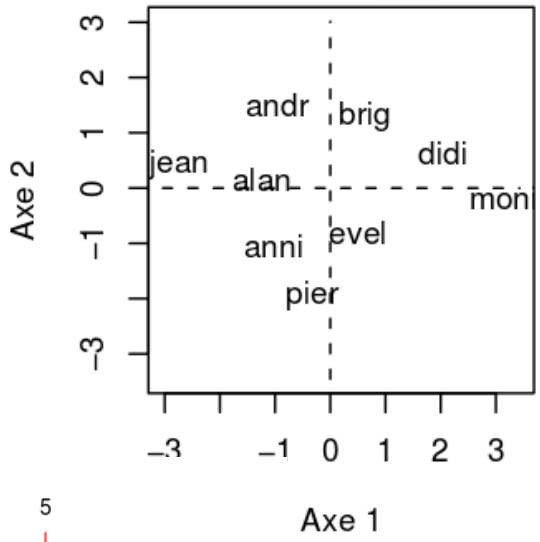
Contexte

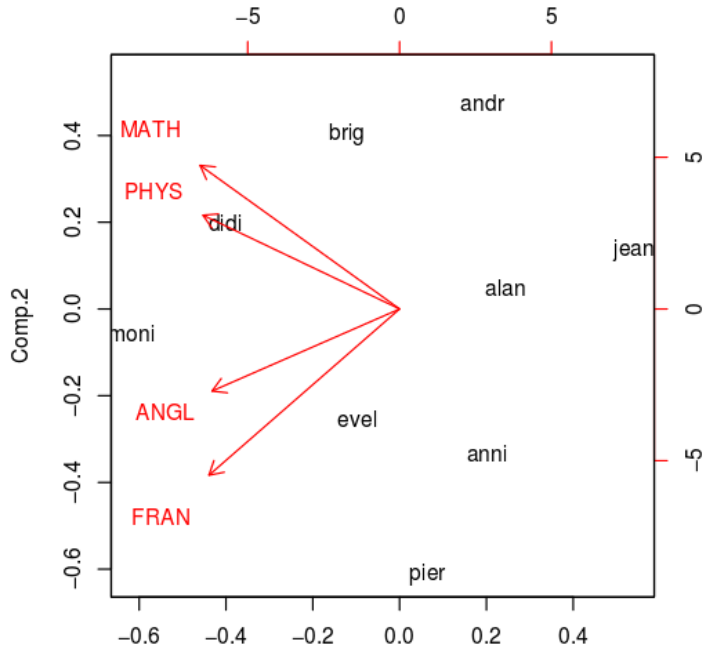

MATH
PHYS
ANGL
FRAN
Axe 2
Axe 1
0.6
0.2
−0.2
−0.6
−0.2
0.2
0.6
1.0


andr
brig
jean
alan
didi
mon
anni
evel
pier
Axe 2
Axe 1
3
2
1
0
−1
−3
−3
−1
0
1
2
3

Contexte

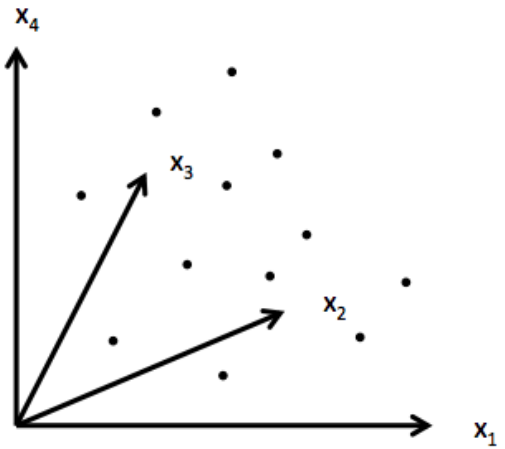

MATH
PHYS
ANGL
FRAN
Axe 2
Axe 1
0.6
0.2
−0.2
−0.6
−0.2
0.2
0.6

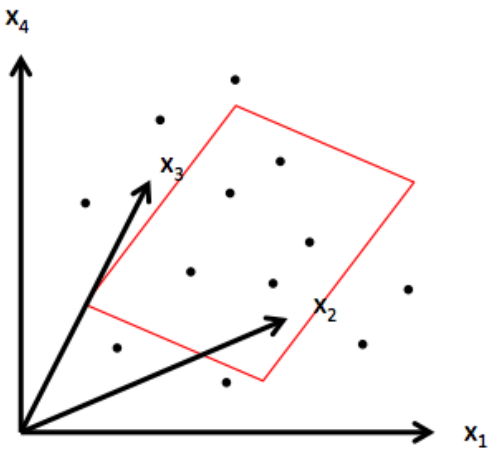

andr
brig
didi
jean
alan
mon
anni
evel
pier
Axe 2
Axe 1
3
2
1
0
−1
−3
−3
−1
0
1
2
3

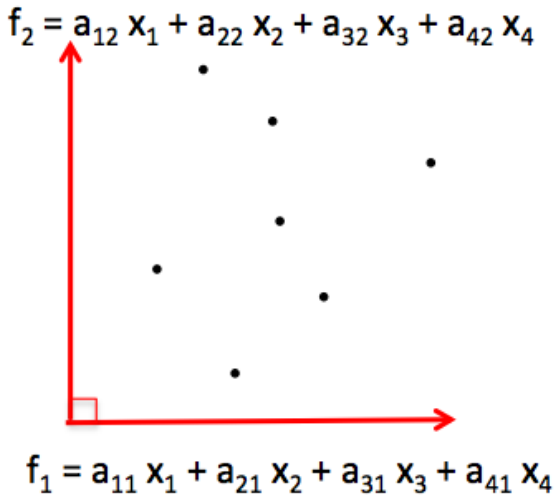

MATH
PHYS
ANGL
FRAN
andr
brig
didi
jean
alan
moni
evel
anni
pier
Comp.2
0.4
0.2
0.0
−0.2
−0.4
−0.6
−5
0
5
−0.6
−0.4
−0.2
0.0
0.2
0.4

Principe général

- Notations
  - $I_1, ..., I_i, ..., I_n$ individus et $V_1, ..., V_j, ..., V_p$ variables
  - $X \in \mathcal{R}^{n \times p}$: tableau de données
- Intuition : vers un changement de base
  - L'idée est de trouver le sous-espace F_S de rang S (S¡p) tel que :
    - minimiser la perte d'information de la projection des x_{ij} sur F_S : min : $argmin_{F_S} \sum_{i=1}^{n} ||x_i - \bar{x}_i||^2$
    - chaque nouvel axe est une combinaison linéaires des axes originaux $f_k = \sum_{j=1}^{p} \alpha_j V_j$
    - les nouveaux axes soient orthogonaux (axes non corrélés)







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 en Composantes Principales (ACP)

Analyse en Composantes Principales (ACP)

- Données
  - Variables continues et centrées
  - Variables réduites dans le cas de variables hétérogènes (ACP normée)

	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

$$X = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \dots & x_{i,j} & \dots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \dots \\ x_i^T \\ \dots \\ x_n^T \end{pmatrix}$$

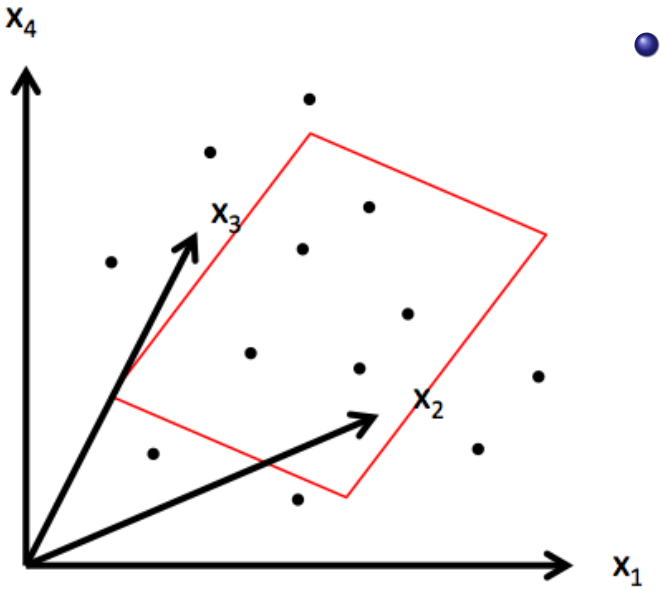
$\rightarrow$ Nécessité de centrer les données, mais pas de réduire (variables homogènes : notes)

Objectifs

Effectuer un changement de base qui prend en compte les relations entre les variables et/ou les relations entre les individus.

Analyse en Composantes Principales (ACP)

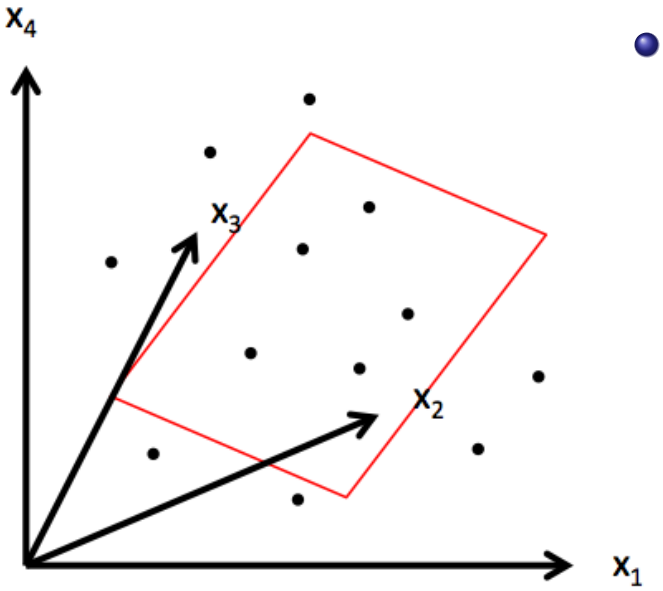
- Changement de base



- Notion de projection linéaire de x_i, sur $f_i \in \mathcal{R}^d$
  - $f_i = M^T x_i$ où $M \in \mathcal{R}^{p \times d}$ et $M^T M = 1$

Analyse en Composantes Principales (ACP)

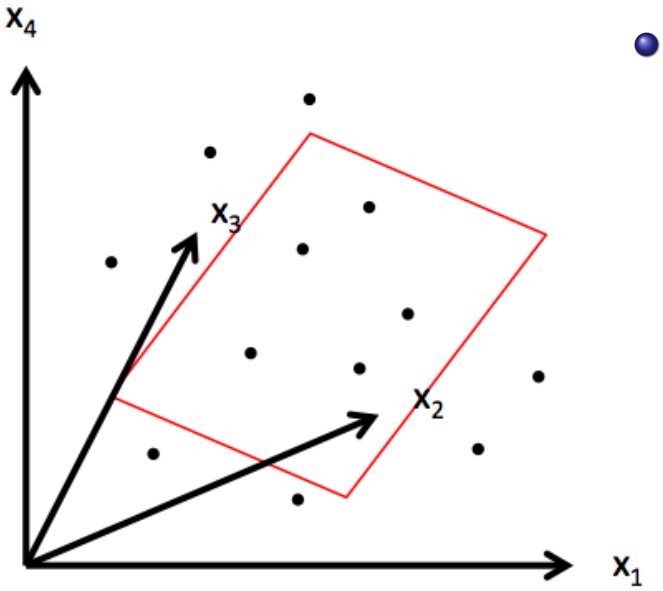
- Changement de base



- Notion de projection linéaire de x_i, sur $f_i \in \mathcal{R}^d$
  - $f_i = M^T x_i$ où $M \in \mathcal{R}^{p \times d}$ et $M^T M = 1$
  - si $d = p$, pas de réduction de dimension, pas de perte d'information :
    $f_i = M^T x_i \rightarrow Mf_i = MM^T x_i \rightarrow x_i = Mf_i$

Analyse en Composantes Principales (ACP)

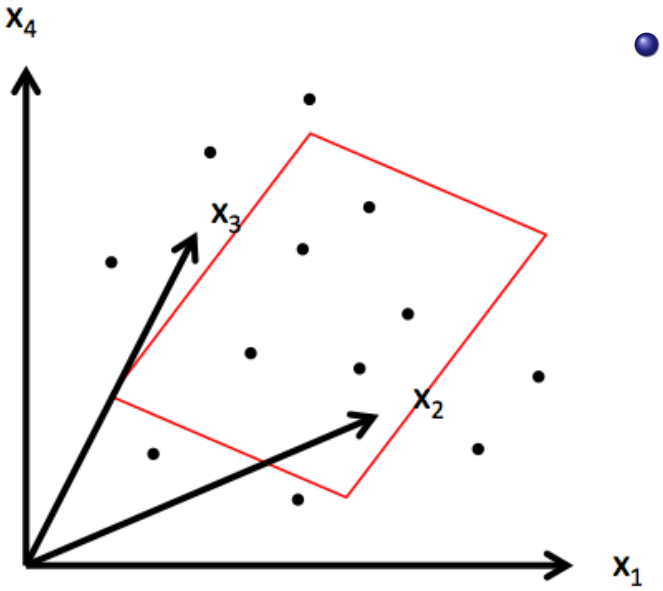
- Changement de base



- Notion de projection linéaire de x_i, sur $f_i \in \mathcal{R}^d$
  - $f_i = M^T x_i$ où $M \in \mathcal{R}^{p \times d}$ et $M^T M = 1$
  - si $d = p$, pas de réduction de dimension, pas de perte d'information :
    $f_i = M^T x_i \rightarrow Mf_i = MM^T x_i \rightarrow x_i = Mf_i$
  - si $d < p$, réduction de dimension, reconstruction par approximation :
    $\hat{x}_i = Mf_i$ ou $\hat{x}_i = MM^T x_i$

Analyse en Composantes Principales (ACP)

- Changement de base



- Notion de projection linéaire de x_i, sur $f_i \in \mathcal{R}^d$
  - $f_i = M^T x_i$ où $M \in \mathcal{R}^{p \times d}$ et $M^T M = 1$
  - si $d = p$, pas de réduction de dimension, pas de perte d'information :
    $f_i = M^T x_i \rightarrow Mf_i = MM^T x_i \rightarrow x_i = Mf_i$
  - si $d < p$, réduction de dimension, reconstruction par approximation :
    $\hat{x}_i = Mf_i$ ou $\hat{x}_i = MM^T x_i$

Objectif

Trouver M qui minimise l'erreur quadratique
$MSE(M) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - \hat{x}_i||^2$

Analyse en Composantes Principales (ACP)

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

Analyse en Composantes Principales (ACP)

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

Analyse en Composantes Principales (ACP)

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

Analyse en Composantes Principales (ACP)

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

Analyse en Composantes Principales (ACP)

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

Quel est le lien avec les "relations" entre les variables ?

Matrice de covariance $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$

Sur données centrées $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i)(x_i)^T = \frac{1}{n}X^T X$

Analyse en Composantes Principales (ACP)

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

Quel est le lien avec les "relations" entre les variables ?

Matrice de covariance $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$

Sur données centrées $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i)(x_i)^T = \frac{1}{n}X^T X$

Conséquence :

Minimiser MSE(M) $\leftrightarrow$ maximiser la variance de des données par rapport à la projection M.

Analyse en Composantes Principales (ACP)

- Intuition : identifier le premier **axe factoriel** f_1 tel que la variance Xf_1 soit maximale. On appelle le vecteur $c_1 = Xf_1$ une **composante principale**.

Analyse en Composantes Principales (ACP)

- Intuition : identifier le premier **axe factoriel** f_1 tel que la variance Xf_1 soit maximale. On appelle le vecteur $c_1 = Xf_1$ une **composante principale**.

Soit $M = f_1$, le premier axe factoriel

$$\begin{aligned} MSE(M) &= \frac{1}{n}\sum_{i=1}^{n} x_{i,}^T x_i - \frac{1}{n}\sum_{i=1}^{n} f_1^T x_i x_i^T f_1 \\ &\propto -f_1^T(\frac{1}{n}\sum_{i=1}^{n}(x_i)(x_i)^T)f_1 = -f_1^T \Sigma f_1 \end{aligned}$$

Analyse en Composantes Principales (ACP)

- Intuition : identifier le premier **axe factoriel** f_1 tel que la variance Xf_1 soit maximale. On appelle le vecteur $c_1 = Xf_1$ une **composante principale**.

Soit $M = f_1$, le premier axe factoriel

$$\begin{aligned} MSE(M) &= \frac{1}{n}\sum_{i=1}^{n} x_{i,}^T x_i - \frac{1}{n}\sum_{i=1}^{n} f_1^T x_i x_i^T f_1 \\ &\propto -f_1^T(\frac{1}{n}\sum_{i=1}^{n}(x_i)(x_i)^T)f_1 = -f_1^T \Sigma f_1 \end{aligned}$$

Problème d'optimisation sous contrainte :

$$min_{f_1} MSE(f_1) = -f_1^T \Sigma f_1 \ avec \ f_1^T f_1 = 1$$

Analyse en Composantes Principales (ACP)

Problème d'optimisation sous contrainte :

$$min_{f_1} MSE(f_1) = -f_1^T \Sigma f_1 \ avec \ f_1^T f_1 = 1$$

Analyse en Composantes Principales (ACP)

Problème d'optimisation sous contrainte :

$$min_{f_1} MSE(f_1) = -{f_1}^T \Sigma f_1 \ avec \ {f_1}^T f_1 = 1$$

- Résolution par le langragien :

$$\begin{aligned} \mathcal{L}(f_1, \lambda_1) &= -{f_1}^T \Sigma f_1 + \lambda_1({f_1}^T f_1 - 1) \\ \nabla_{f_1} \mathcal{L}(f_1, \lambda_1) &= -2\Sigma f_1 + 2\lambda_1 f_1 \quad \rightarrow \quad \Sigma f_1 = \lambda_1 f_1 \rightarrow {f_1}^T \Sigma f_1 = \lambda_1 \\ \nabla_{\lambda_1} \mathcal{L}(f_1, \lambda_1) &= {f_1}^T f_1 - 1 \end{aligned}$$

Analyse en Composantes Principales (ACP)

Problème d'optimisation sous contrainte :

$$min_{f_1} MSE(f_1) = -{f_1}^T \Sigma f_1 \ avec \ {f_1}^T f_1 = 1$$

- Résolution par le langragien :

$$\begin{aligned} \mathcal{L}(f_1, \lambda_1) &= -{f_1}^T \Sigma f_1 + \lambda_1({f_1}^T f_1 - 1) \\ \nabla_{f_1}\mathcal{L}(f_1, \lambda_1) &= -2\Sigma f_1 + 2\lambda_1 f_1 \quad \rightarrow \quad \Sigma f_1 = \lambda_1 f_1 \rightarrow {f_1}^T \Sigma f_1 = \lambda_1 \\ \nabla_{\lambda_1}\mathcal{L}(f_1, \lambda_1) &= {f_1}^T f_1 - 1 \end{aligned}$$

Rappel - Valeurs propres et vecteurs propres : Un vecteur propre X associé à une valeur propre λ doit vérifier la relation $AX = \lambda X$

Analyse en Composantes Principales (ACP)

Problème d'optimisation sous contrainte :

$$min_{f_1} MSE(f_1) = -f_1^T \Sigma f_1 \text{ avec } f_1^T f_1 = 1$$

- Résolution par le langragien :

$$\begin{aligned} \mathcal{L}(f_1, \lambda_1) &= -f_1^T \Sigma f_1 + \lambda_1 (f_1^T f_1 - 1) \\ \nabla_{f_1} \mathcal{L}(f_1, \lambda_1) &= -2\Sigma f_1 + 2\lambda_1 f_1 \quad \rightarrow \quad \Sigma f_1 = \lambda_1 f_1 \rightarrow f_1^T \Sigma f_1 = \lambda_1 \\ \nabla_{\lambda_1} \mathcal{L}(f_1, \lambda_1) &= f_1^T f_1 - 1 \end{aligned}$$

  Rappel - Valeurs propres et vecteurs propres : Un vecteur propre X associé à une valeur propre λ doit vérifier la relation $AX = \lambda X$

- Conclusion
  - f_1 et λ_1 sont respectivement des vecteurs propres et valeurs propres
  - $MSE(f_1)$ peut aussi s'écrire ainsi : $MSE(f_1) = -\lambda_1$. Par conséquent, on cherche à maximiser la valeur propre
  - Le premier axe factoriel est issu du vecteur propre f_1 associé à la plus grande valeur propre λ_1 de la matrice de covariance Σ

Analyse en Composantes Principales (ACP)

- Identification du deuxième axe factoriel f_2

$$min\ MSE(f_2) = -f_2^T \Sigma f_2$$
$$tel\ que f_2^T f_2 = 1 \quad et \quad f_2^T f_1 = 0$$

Ce qui revient à trouver la deuxième valeur propre λ_2 et son vecteur propre f_2 associé.

Analyse en Composantes Principales (ACP)

- Identification du deuxième axe factoriel f_2

$$min\ MSE(f_2) = -f_2^T \Sigma f_2$$
$$tel\ que f_2^T f_2 = 1 \quad et \quad f_2^T f_1 = 0$$

  Ce qui revient à trouver la deuxième valeur propre λ_2 et son vecteur propre f_2 associé.
- Identification du troisième axe factoriel f_3... Même principe... etc...

Analyse en Composantes Principales (ACP)

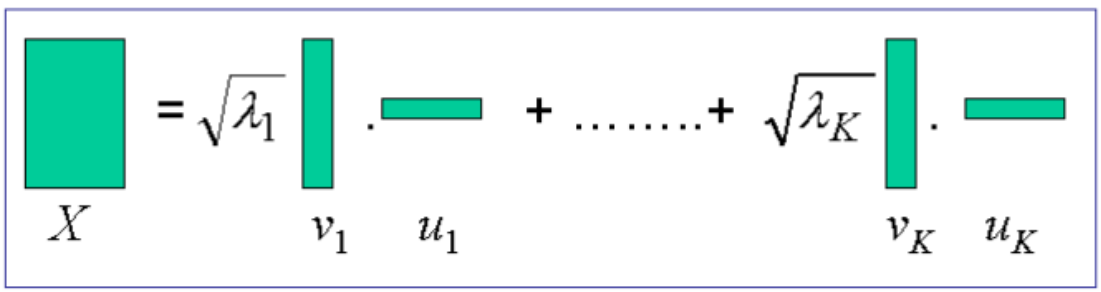
- Identification du deuxième axe factoriel f_2

$$min\ MSE(f_2) = -f_2^T \Sigma f_2$$
$$tel\ que f_2^T f_2 = 1 \quad et \quad f_2^T f_1 = 0$$

  Ce qui revient à trouver la deuxième valeur propre λ_2 et son vecteur propre f_2 associé.
- Identification du troisième axe factoriel f_3... Même principe... etc...

Reconstitution de la matrice X à partir des axes factoriels

On peut voir aussi le changement de base comme une décomposition en valeurs singulières de la matrice X :



Analyse en Composantes Principales (ACP)

Dans la pratique...

A partir de données X centrées (et éventuellement réduites)

- Estimer la matrice de covariance $\Sigma = \frac{1}{n} X^T X$
- Identifier les valeurs propres de Σ
- Ordonner les k valeurs propres par ordre croissant afin de former la nouvelle base M
- Projeter les points pour obtenir les composantes : $C = XM$

Analyse en Composantes Principales (ACP)

Exemple illustratif

- 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

- 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

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

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

Notion d'inertie 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

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

Notion d'inertie 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 $\frac{1}{n} X^T X$ et de $\frac{1}{n} X X^T$ sont égales $\rightarrow$ Rechercher la meilleure représentation des individus équivaut à rechercher la meilleure représentation des variables

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

Notion d'inertie 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), ...

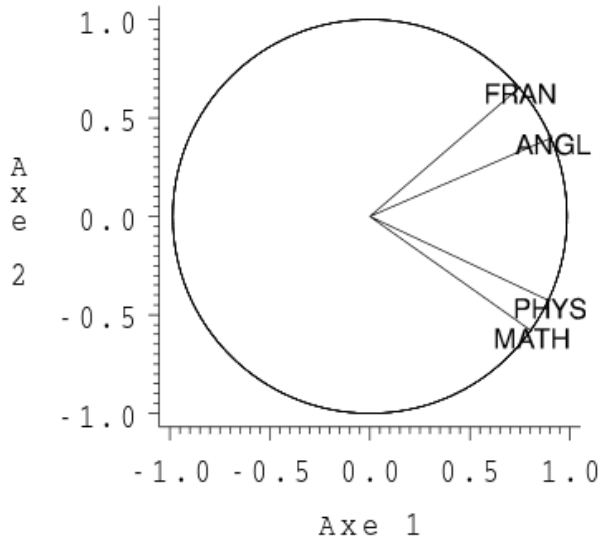
Analyse en Composantes Principales (ACP)

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



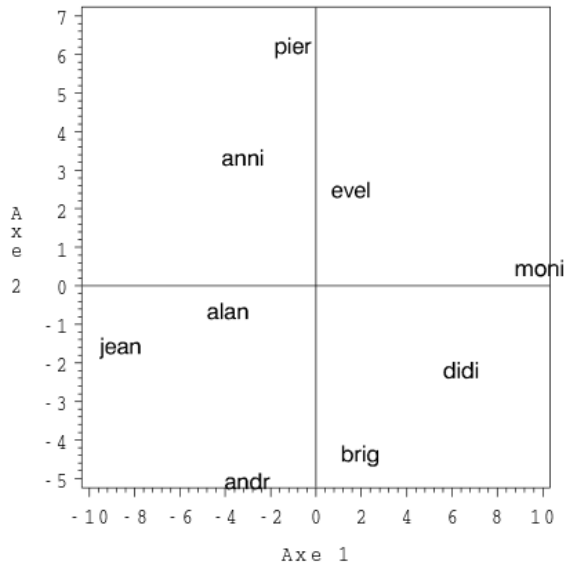
Analyse en Composantes Principales (ACP)

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



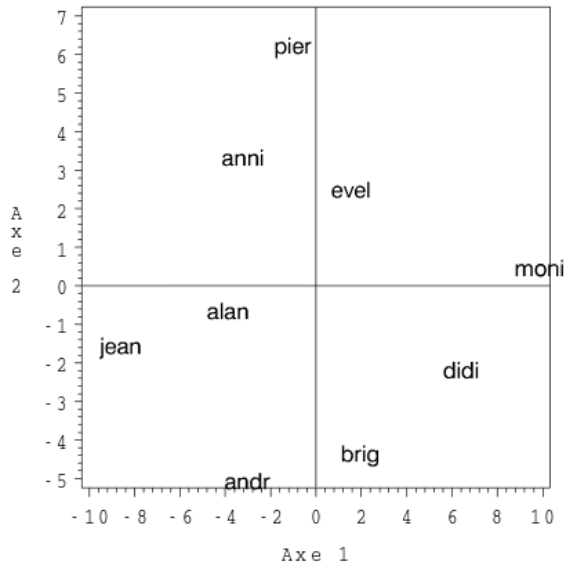
Analyse en Composantes Principales (ACP)

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

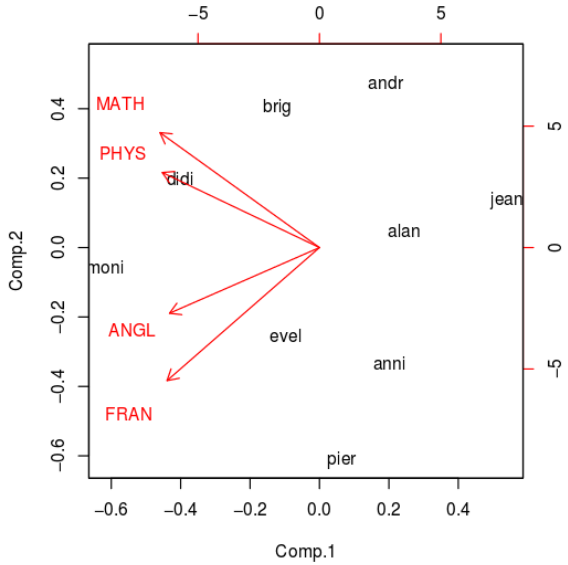
- On peut mesure la contribution d'un point à l'inertie d'un nuage :

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

Analyse en Composantes Principales (ACP)

Exemple illustratif

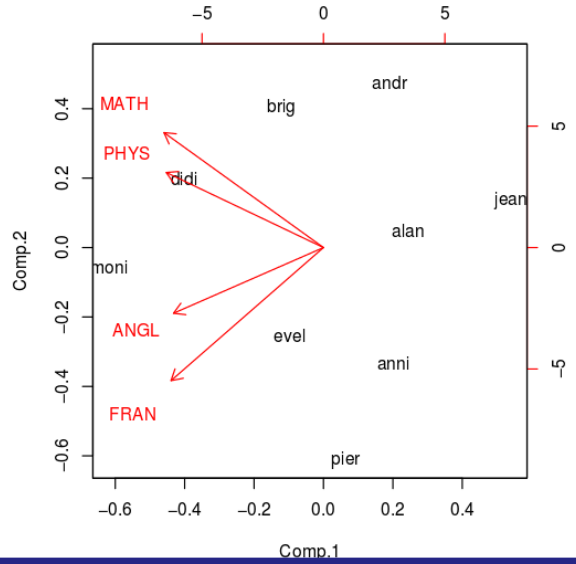
- Etape 3 : Projection des individus et variables (biplot)
  - Projection des individus et variables



Analyse en Composantes Principales (ACP)

Exemple illustratif

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

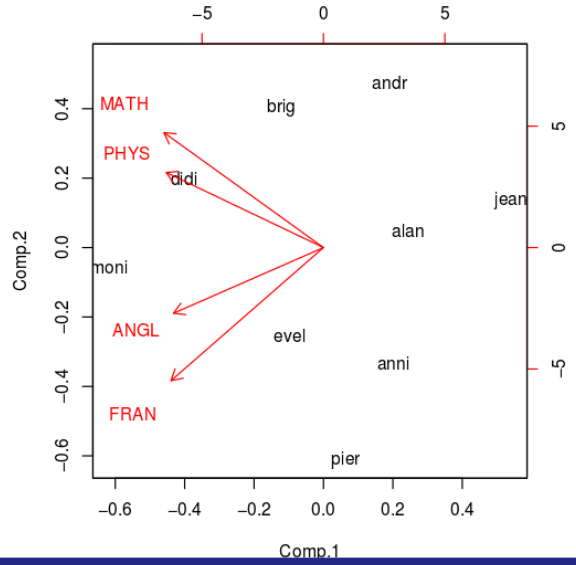


Interprétation

Analyse en Composantes Principales (ACP)

Exemple illustratif

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



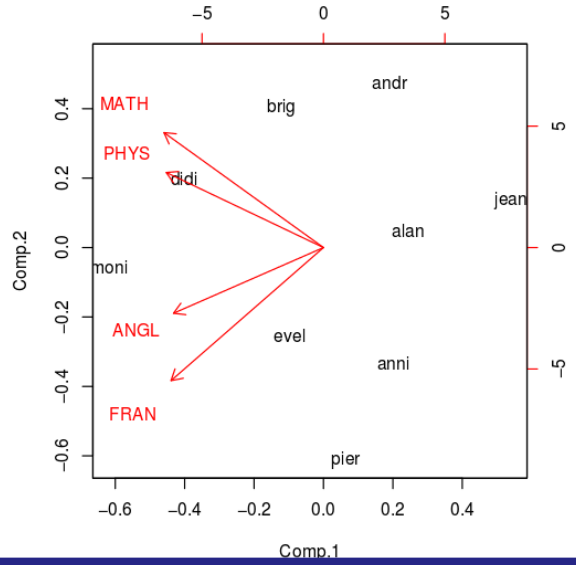
Interprétation

- Deux individus proches se ressemblent

Analyse en Composantes Principales (ACP)

Exemple illustratif

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



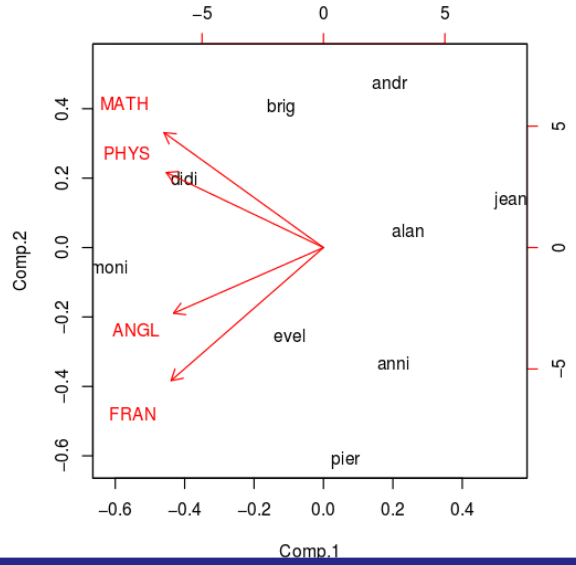
Interprétation

- Deux individus proches se ressemblent
- Deux variables très corrélées positivement sont du même côté sur un axe

Analyse en Composantes Principales (ACP)

Exemple illustratif

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



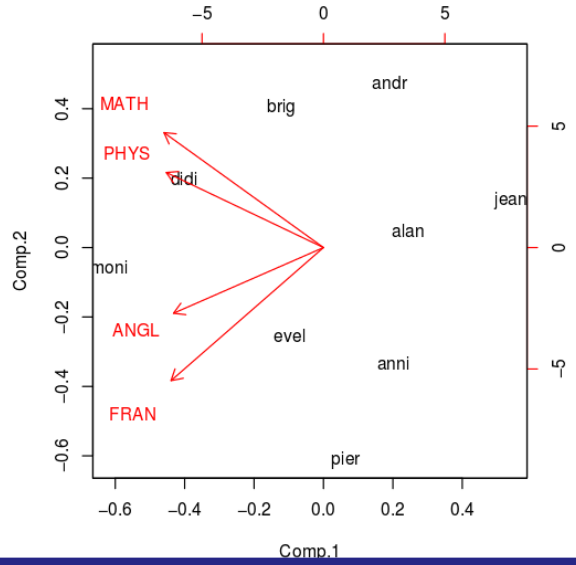
Interprétation

- 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)

Analyse en Composantes Principales (ACP)

Exemple illustratif

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



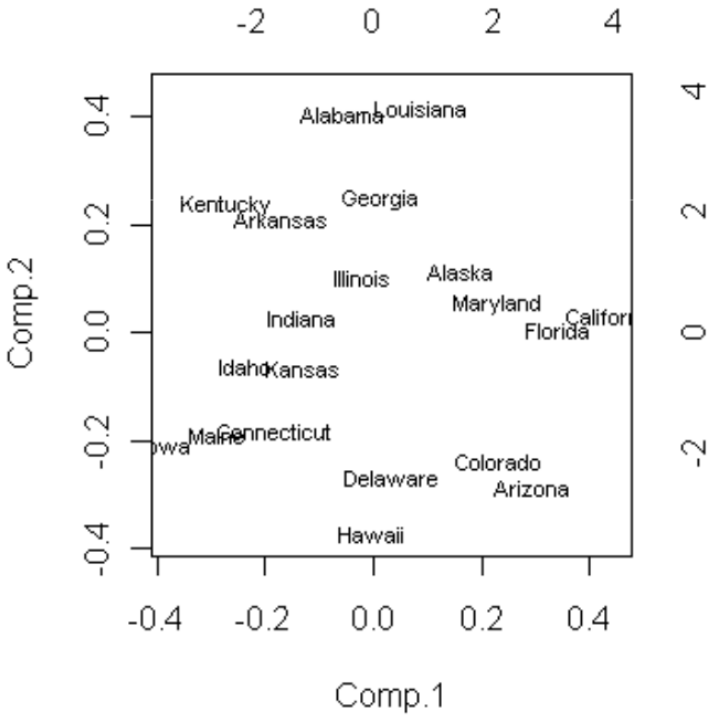
Interprétation

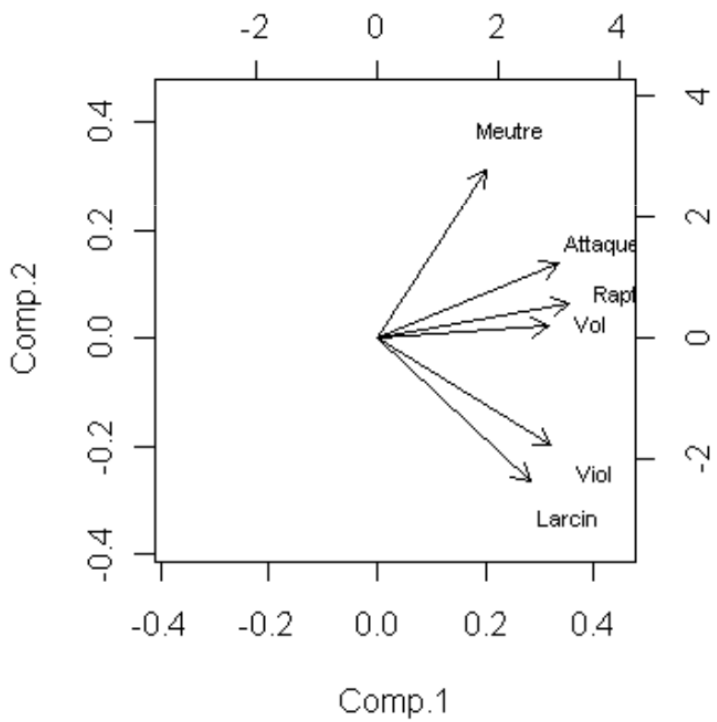
- Deux individus proches se ressemblent
- Deux variables très corrélées positivement sont du même côté sur un axe
- Un individu sera proche des variables pour lesquelles il a de fortes valeurs (et inversement)
- Plus les valeurs d'un individu pour une variable sont fortes, plus il sera éloigné de l'origine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

Analyse en Composantes Principales (ACP)

Exemple illustratif

- Etudes des crimes aux USA





Analyse Factorielle des Correspondances (AFC)

Analyse Factorielle des Correspondances (AFC)

- 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.

Analyse Factorielle des Correspondances (AFC)

- 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 $x''_{ij} = \frac{x_{ij}}{x_{.j}}$

Analyse Factorielle des Correspondances (AFC)

- 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

Analyse Factorielle des Correspondances (AFC)

- 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 ?

Analyse Factorielle des Correspondances (AFC)

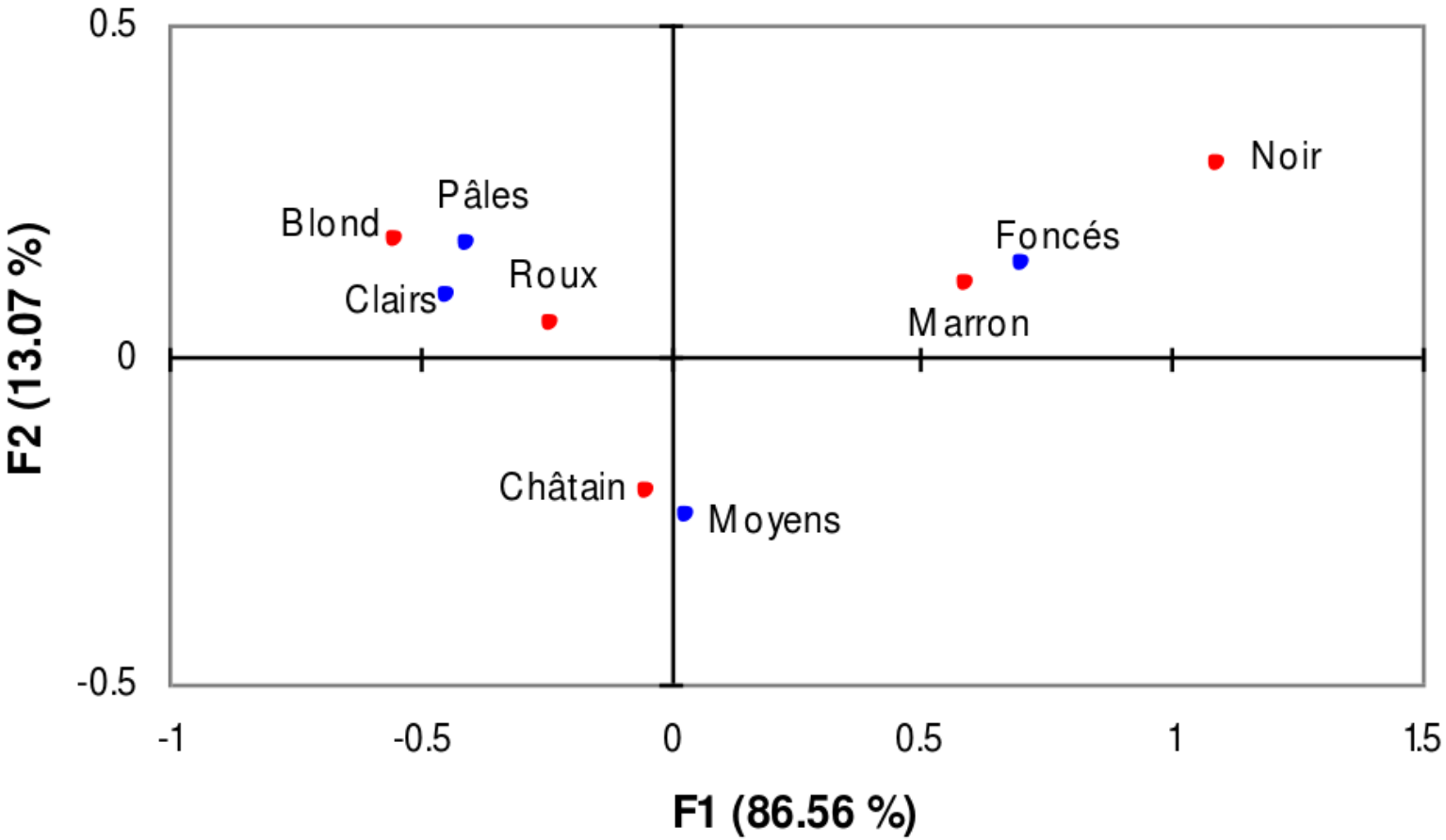
- Objectif
  - Analyser la liaison entre deux variables : la liaison entre deux variables est grande si les profils-lignes ou colonnes sont différents.
    - Quelles sont les lignes qui se ressemblent ? sont différentes ?
    - Existe-t-il des groupes homogènes entre les lignes ? entre les colonnes ?

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
...			

Analyse des Correspondances Multiples (ACM)

- Données
  - p variables qualitatives (par exemple QCM)

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

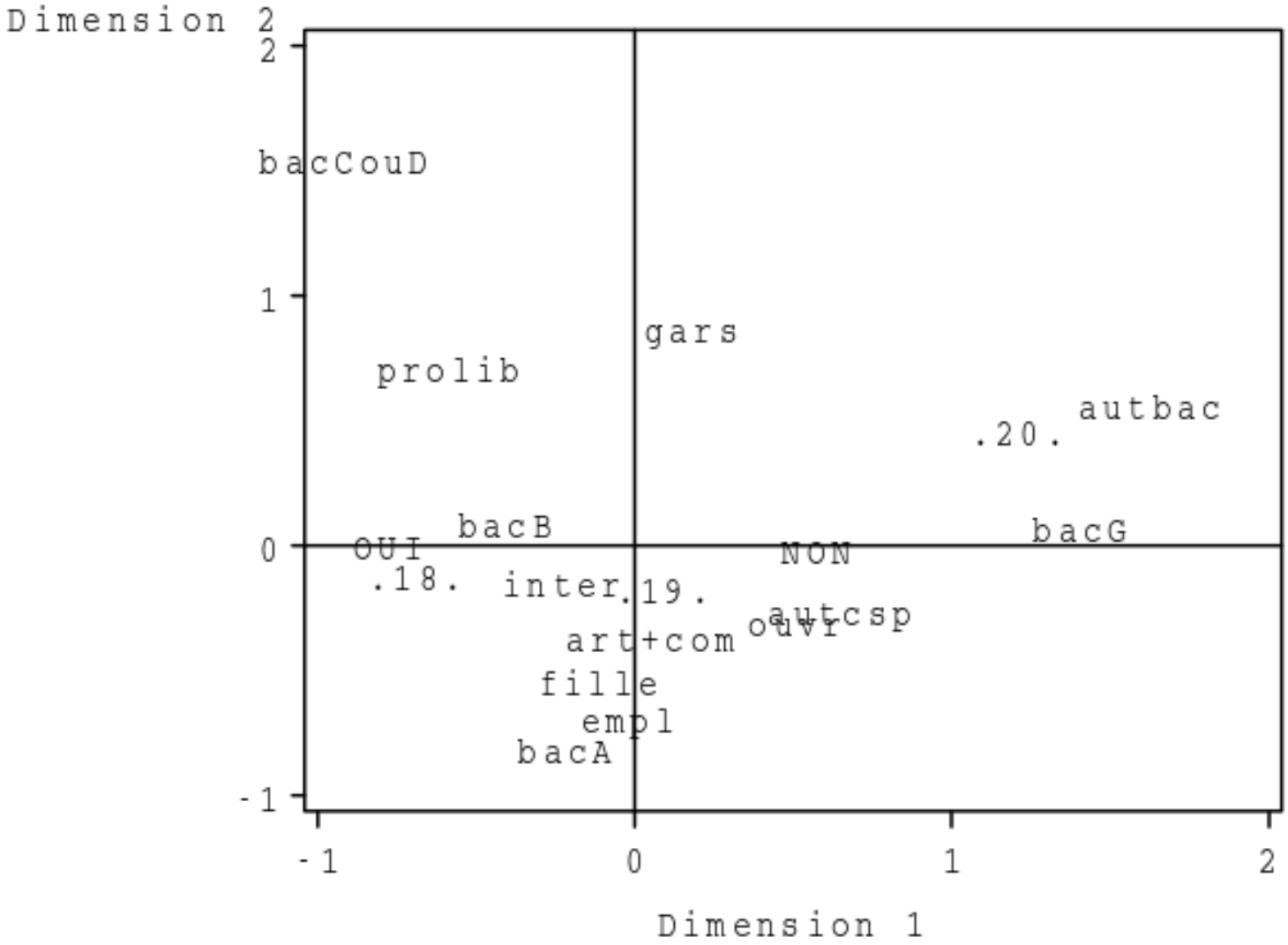
  - Transformé en tableau de Burt ("Grand tableau de contingence")

	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)



Références

- http://www.irisa.fr/dream/Seminaire/Tahiti/Tahiti04/transparents/emmanuel/emmanuel.pdf
- https://moodle.insa-rouen.fr/pluginfile.php/1337/mod_resource/content/0/Parties_1_et_3_DM/pcabeamer.pdf
- https://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf