Resumé du cours d'ACP.

Soient une population de n individus caractérisés par p variables. On va voir comment on peut projeter le nuage de n points dans $\mathbb{R}^p$ sur un sous espace vectoriel en conservant le maximum d'information.

Etape 1 : Definition des variables :

$$X = \begin{array}{c} \\ e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{array} \begin{array}{c} \begin{array}{cccccc} X^{(1)} & X^{(2)} & \cdots & X^{(j)} & \cdots & X^{(p)} \end{array} \leftarrow \text{variables} \\ \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(j)} & \cdots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(j)} & \cdots & x_2^{(p)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_i^{(1)} & x_i^{(2)} & \cdots & x_i^{(j)} & \cdots & x_i^{(p)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(i)} & \cdots & x_n^{(p)} \end{pmatrix} \end{array}$$

$\uparrow$ individus

Pour chaque variable, on peut calculer sa moyenne et sa variance

$$\forall i \in [\![1, p]\!] \qquad \overline{X^{(i)}} = \frac{1}{n} \sum_{k=1}^{n} x_k^{(i)}$$

$$S^2_{X^{(i)}} = \frac{1}{n} \sum_{k=1}^{n} \left(x_k^{(i)} - \overline{X^{(i)}} \right)^2$$

Etape 2 : Definition des variables centrées-reduites :

Cette etape est obligatoire si les $X^{(i)}$ ont des unités differentes

$$\tilde{X} = \begin{pmatrix} \widetilde{X^{(1)}} & \widetilde{X^{(2)}} & \cdots & \widetilde{X^{(p)}} & \\ \frac{x_1^{(1)} - \overline{X^{(1)}}}{S_{X^{(1)}}} & \frac{x_1^{(2)} - \overline{X^{(2)}}}{S_{X^{(2)}}} & \cdots & \frac{x_1^{(p)} - \overline{X^{(p)}}}{S_{X^{(p)}}} & e_1 \\ \frac{x_2^{(1)} - \overline{X^{(1)}}}{S_{X^{(1)}}} & \frac{x_2^{(2)} - \overline{X^{(2)}}}{S_{X^{(2)}}} & \cdots & \frac{x_2^{(p)} - \overline{X^{(p)}}}{S_{X^{(p)}}} & e_2 \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{x_n^{(1)} - \overline{X^{(1)}}}{S_{X^{(1)}}} & \frac{x_n^{(2)} - \overline{X^{(2)}}}{S_{X^{(2)}}} & \cdots & \frac{x_n^{(p)} - \overline{X^{(p)}}}{S_{X^{(p)}}} & e_p \end{pmatrix}$$

Etape 3 : Calcul de la matrice des corrélations linéaires

$$r = \begin{pmatrix} 1 & r(\tilde{X}_{(1)}, \tilde{X}_{(2)}) & \cdots & r(\tilde{X}_{(1)}, \tilde{X}_{(p)}) \\ r(\tilde{X}_{(1)}, \tilde{X}_{(2)}) & 1 & \cdots & r(\tilde{X}_{(2)}, \tilde{X}_{(p)}) \\ \vdots & & & \\ r(\tilde{X}_{(2)}, \tilde{X}_{(p)}) & \cdots & & 1 \end{pmatrix}$$

où $r(\tilde{X}_{(k)}, \tilde{X}_{(j)}) = \dfrac{C_{\tilde{X}^{(k)}, \tilde{X}^{(j)}}}{S_{\tilde{X}^{(k)}} \, S_{\tilde{X}^{(j)}}}$

r est une matrice symétrique définie positive $\Rightarrow$ diagonalisable et il existe une base orthonormée de vecteurs propres

Rappel : a. Si $r(\tilde{X}^{(k)}, \tilde{X}^{(j)}) \sim 0$ alors les deux variables sont indépendantes

b. Si $r(\tilde{X}^{(k)}, \tilde{X}^{(j)}) \sim 1$ alors les deux variables sont positivement et linéairement corrélées :

→ Si $\tilde{X}^{(k)} \nearrow$ alors $\tilde{X}^{(j)} \nearrow$ et réciproquement

c. Si $r(\tilde{X}^{(k)}, \tilde{X}^{(j)}) \sim -1$ alors les deux variables sont négativement et linéairement corrélées :

→ Si $\tilde{X}^{(k)} \nearrow$ alors $\tilde{X}^{(j)} \searrow$

→ Si $\tilde{X}^{(l)} \nearrow$ alors $\tilde{X}^{(k)} \searrow$

Etape 4 : Diagonalisation de r :

On obtient (λ_i, u_i) où l'on peut choisir $\{u_i\}$ telle que ce soit une famille orthonormée.

λ_i : valeur propre ; u_i : vecteur propre associé

Vocabulaire : les vecteurs propres u_i sont appelés axes principaux.

Etape 5 : Les composantes principales :

Les composantes principales sont les projections orthogonales de chacune des vecteurs sur les axes principaux et elles sont données par

$$C_{(i)} = \tilde{X} u_i \quad \text{pour } \forall i \in [\![1, p]\!]$$

Propriétés des composantes principales :

$$S^2_{C_i} = \lambda_i \text{ pour } \forall i \in [\![1,p]\!] \text{ où } \overline{C_i} = \frac{1}{n}\sum_{k=1}^{n} C_i^k$$

$$S^2_{C_i} = \frac{1}{n}\sum_{k=1}^{n}\left(C_i^k - \overline{C_i}\right)$$

Or $\sum_{i=1}^{p} \lambda_i = p$ (car la trace d'une matrice ne dépend pas de la base)

on a donc $\sum_{i=1}^{p} S^2_{C_i} = \sum_{i=1}^{p} \lambda_i = p = \sum_{i=1}^{p} S^2_{x^{(i)}}$

On pourra ainsi faire une représentation 2D du nuage initial si $\lambda_1 + \lambda_2 \gg \sum_{\substack{i\neq 1 \\ i\neq 2}} \lambda_i$ ou dit autrement si $\frac{\lambda_1+\lambda_2}{\sum \lambda_i} \sim 1$.

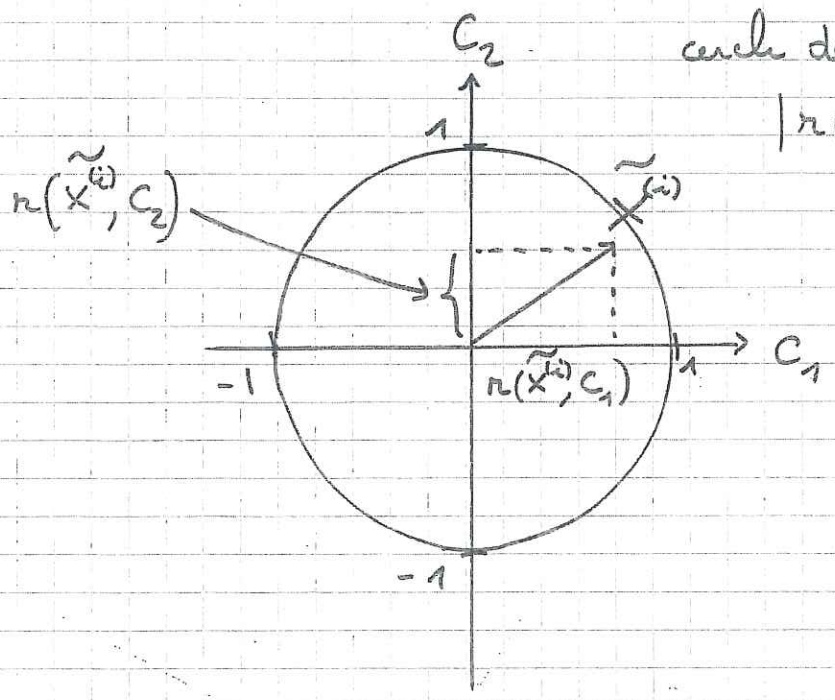
En effet $\lambda_1 + \lambda_2$ représente la variance du nuage initial portée par le plan (C_1, C_2).

Etape 6 : Représentation des variables :

$$r(\widetilde{x^{(k)}}, C_i) = \begin{pmatrix} r(\widetilde{x^{(1)}}, C_1) & r(\widetilde{x^{(1)}}, C_2) & \cdots & r(\widetilde{x^{(1)}}, C_p) \\ r(\widetilde{x^{(2)}}, C_1) & r(\widetilde{x^{(2)}}, C_2) & \cdots & r(\widetilde{x^{(2)}}, C_p) \\ \vdots & & & \\ r(\widetilde{x^{(p)}}, C_1) & r(\widetilde{x^{(p)}}, C_2) & \cdots & r(\widetilde{x^{(p)}}, C_p) \end{pmatrix}$$

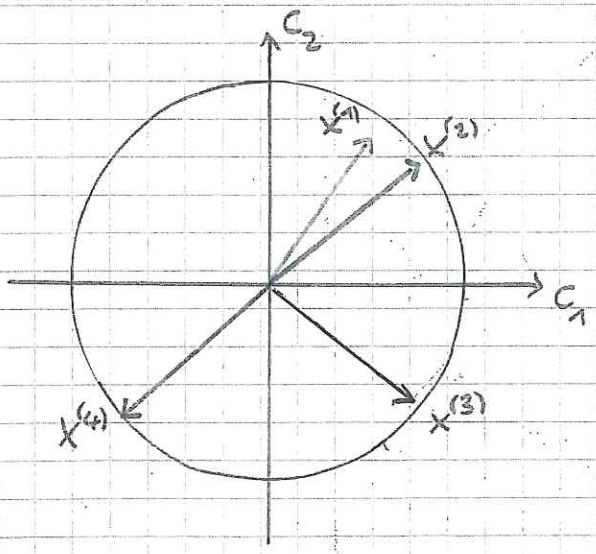
corrélation entre les anciennes variables et les nouvelles

Representation de la variable $X^{(i)}$



cercle de rayon 1 car $|r(\tilde{X}^{(k)}, C_i)| \leq 1$

Interpretation 1 :

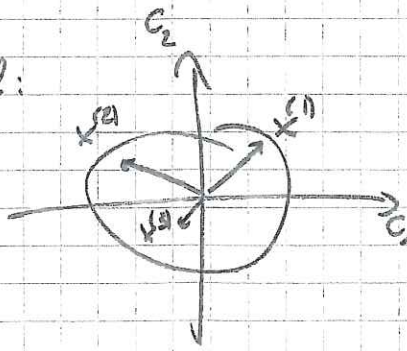


* $X^{(1)}$, $X^{(2)}$ sont très corrélées linéairement et positivement puisque l'angle est faible

* $X^{(4)}$ est très corrélé linéairement et négativement avec $X^{(2)}$ et $X^{(1)}$ car l'angle est négatif

* $X^{(3)}$ est indépendant de $X^{(1)}$, $X^{(2)}$ et $X^{(4)}$ car l'angle fait par $X^{(3)}$ avec ces 3 variables est proche de $\frac{\pi}{2}$

Interpretation 2 :



$X^{(3)}$ est mal représenté par le plan (C_1, C_2) car le point est trop loin du bord du cercle

On définit $\cos^2$ par

$$\cos^2(\tilde{X}^{(i)}, C_k) = \frac{r^2(\tilde{X}^{(i)}, C_k)}{\sum_{m=1}^{p} r^2(\tilde{X}^{(i)}, C_m)} = r^2(\tilde{X}^{(i)}, C_k)$$

Si on représente les variables dans le plan (C_1, C_2), pour que la variable $\tilde{X}^{(i)}$ soit convenablement représentée il faut que

$$\cos^2(\tilde{X}^{(i)}, C_1) + \cos^2(\tilde{X}^{(i)}, C_2) \simeq 1$$

cela est équivalent à dire que la flèche représentant la variable est proche du périmètre du cercle.

Calcul des contributions des variables aux C_i :

$$\text{Contribution de la variable } \tilde{X}_{(i)} \text{ à l'axe } C_k = \frac{\cos^2(\tilde{X}^{(i)}, C_k)}{\sum_{m=1}^{p} \cos^2(\tilde{X}^{(m)}, C_k)}$$

Etape 7 : Représentation des individus :

Si $\frac{\lambda_1 + \lambda_2}{\sum \lambda_i} \sim 1$ alors on peut représenter les individus dans le plan (C_1, C_2) où on rappelle que $C_1 = \tilde{X} u_1$ et $C_2 = \tilde{X} u_2$

On prend les notes :

$X =$

	Math	Physique	Francais	Ang
Jean	6	6	5	5.5
Alan	8	8	8	8
Anni	6	7	11	9.5
Moni	14.5	14.5	15.5	15
Didi	14	14	12	12.5
Andr	11	10	5.5	7
Pier	5.5	7	14	11.5
Brig	13	12.5	8.5	9.5
Evel	9	9.5	12.5	12

$S_{Math} = 3.374743$ $\quad \overline{Math} = 9.666667$

$S_{Physique} = 2.990726$ $\quad \overline{Physique} = 9.8333$

$S_{Francais} = 10.2222$ $\quad \overline{Francais} = 3.473$

$S_{Anglais} = 10.05556$ $\quad \overline{Anglais} = 2.813109$

On peut alors calculer $\tilde{X}$: matrice des variables reduites centrées

$$\tilde{x} = \frac{X - \overline{X}}{S_X}$$

$\frac{Math - \overline{Math}}{S_{Math}}$ $\quad \frac{Physique - \overline{Physique}}{S_{Physique}}$ ...

cette formule signifie :

Jean

Alan

Anni

Moni

On a alors :

$\tilde{X} =$

	$\widetilde{\text{Math}}$	$\widetilde{\text{Physique}}$	$\widetilde{\text{Francais}}$	$\widetilde{\text{Anglais}}$
Jean	-1.0866	-1.2517	-1.5037	-1.6194
Alan	-0.4939	-0.6130	-0.6399	-0.7307
Anni	-1.0866	-0.9474	0.2240	-0.1975
Moni	1.4322	1.5604	1.5197	1.7576
Didi	1.2840	1.3932	0.5119	0.8689
Anah	0.3951	0.0557	-1.3597	-1.0862
Pier	-1.2347	-0.9474	1.0878	0.5135
Brig	0.9877	0.8917	-0.4959	-0.1975
Evel	-0.1976	-0.1114	0.6559	0.6912

On calcule maintenant la matrice de corrélation linéaire

	Math	Physique	Français	Anglais
Math	1	0.9825357	0.2267319	0.508144
Physique	0.9825357	1	0.3966932	0.6515305
Français	0.2267319	0.3966932	1	0.9512058
Anglais	0.508144	0.6515305	0.9512058	1

$= r(X_i, X_j)$

$$r(\text{Math}, \text{Anglais}) = \frac{C_{\text{Math Anglais}}}{S_{\text{Math}}\, S_{\text{Anglais}}}$$

où $C_{\text{Math Anglais}}$ est la covariance

On diagonalise la matrice $r(x_i, x_j)$:

$$\begin{cases} \lambda_1 = 2.875687 \\ \lambda_2 = 1.119687 \\ \lambda_3 = 0.003577 \\ \lambda_4 = 0.001048 \end{cases}$$

avec $\sum \lambda_i = 4$ car $\text{Tr}(r(X_i, X_j)) = 4$ et la trace ne dépend pas de la base.

On a alors

$\lambda_1 = 2.875687$	$\frac{\lambda_1}{\sum \lambda_i} = 71.8922$	71.8922
$\lambda_2 = 1.119687$	$\frac{\lambda_2}{\sum \lambda_i} = 27.9921$	99.88435
$\lambda_3 = 0.003577$	$\frac{\lambda_3}{\sum \lambda_i} = 0.0894$	99.97379
$\lambda_4 = 0.001048$	$\frac{\lambda_4}{\sum \lambda_i} = 0.0262$	100

↑ pourcentage de variances portés par la valeur propre

↑ variance cumulée

Vecteurs Propres :

$$\begin{cases} u_1 = (0.4784540, 0.5319172, 0.4439304, 0.5395106) \\ u_2 = (-0.5519489, -0.4068018, 0.6212336, 0.3793856) \\ u_3 = (-0.2025732, 0.4412026, 0.5324079, -0.6934308) \\ u_4 = (-0.6522256, 0.5974251, -0.3654264, 0.2900837) \end{cases}$$

On peut maintenant calculer les composantes principales c'est à dire les projections du nuage de points sur les axes principaux u_i

$$C_i = \tilde{X} u_i$$

Exemple :

$$C_1 = \begin{pmatrix} -1.0866 & -1.2817 & -1.5037 & -1.6194 \\ -0.4939 & -0.6130 & -0.6399 & -0.7307 \\ -1.0866 & -0.9474 & 0.2240 & -0.1975 \\ 1.4322 & 1.5604 & 1.5197 & 1.7576 \\ 1.2840 & 1.3932 & 0.5119 & 0.8689 \\ 0.3951 & 0.0557 & -1.3597 & -1.0862 \\ -1.2347 & -0.9474 & 1.0878 & 0.5135 \\ 0.9877 & 0.8917 & -0.4959 & -0.1975 \\ -0.1976 & -0.1114 & 0.6559 & 0.6912 \end{pmatrix} \begin{pmatrix} 0.4784540 \\ 0.5319172 \\ 0.4439304 \\ 0.5395106 \end{pmatrix} \simeq \begin{pmatrix} -2.7429 \\ -1.2407 \\ -1.0309 \\ 3.1381 \\ 2.0515 \\ -0.9709 \\ -0.3347 \\ 0.6202 \\ 0.5103 \end{pmatrix}$$

$$C_2 = \begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ & & & \end{pmatrix} \begin{pmatrix} -0.5519 \\ -0.4068 \\ 0.6212 \\ 0.3794 \end{pmatrix} \simeq \begin{pmatrix} -0.4274 \\ -0.1528 \\ 1.0493 \\ 0.1856 \\ -0.6278 \\ -1.4375 \\ 1.9374 \\ -1.2909 \\ 0.8240 \end{pmatrix}$$

Calcul des variances de C_1 et C_2

$\overline{C_1} = 0$ $\quad S^2_{C_1} \simeq 2.8757$

$\overline{C_2} = 0$ $\quad S^2_{C_2} \simeq 1.1197$

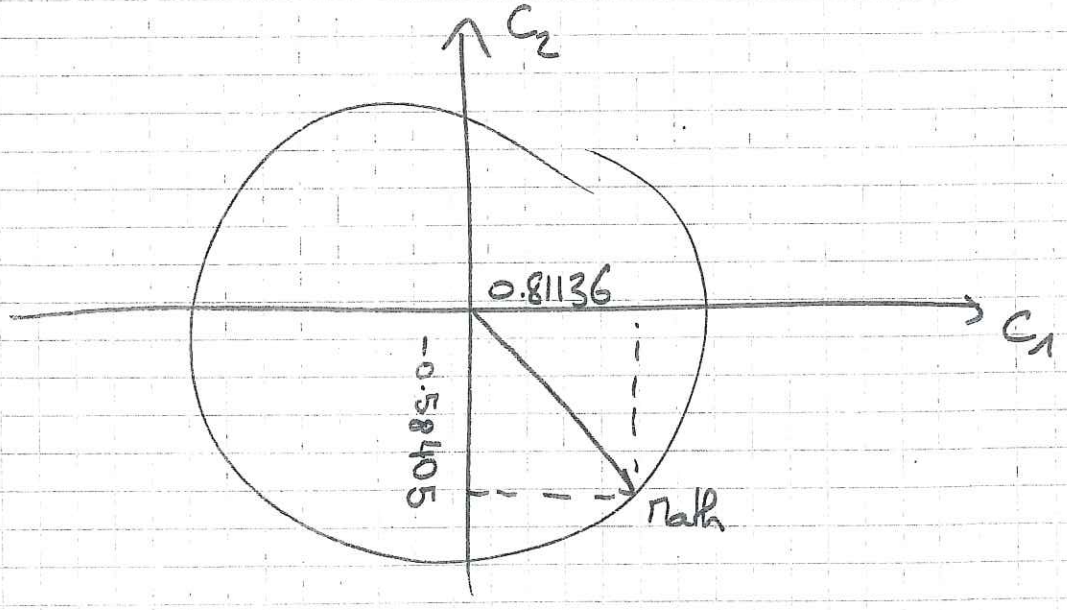
} On retrouve bien $\boxed{S^2_{C_i} = \lambda_i}$

* Description des variables :

On peut maintenant calculer les coordonnées des variables dans le cercle des corrélations. Pour cela, on calcule les coefficients de corrélation linéaires entre les anciennes coordonnées et les nouvelles coordonnées C_i

$r(X_i, C_k)$ = (ième variable originale, composante principale k)

	C_1	C_2	C_3	C_4	
	0.81136	-0.58405	-0.01211	0.02117	Math
	0.90202	-0.43046	0.02639	-0.01934	Physique
	0.75281	0.65736	0.03184	0.01183	français
	0.91489	0.40145	-0.04148	-0.00939	Anglais



→ Calcul de $\cos^2$:

$$\cos^2(X_i, C_k) = \frac{r^2(X_i, C_k)}{\sum_{m=1}^{4} r^2(X_i, C_m)} = r^2(X_i, C_k)$$

X_i : Ancienne variable ; C_k : nouvelle variable

On obtient donc la matrice

	C_1	C_2	C_3	C_4
Math	0.65829	0.34111	0.00014	0.00045
Physique	0.813635	0.18529	0.00069	0.00037
Français	0.56672	0.43212	0.00101	0.0014
Anglais	0.83703	0.16116	0.00172	0.00009

Si cette somme est proche de 1 alors la variable est bien représenté : ceci correspond au fait de se trouver proche d'un bord du cercle unité des corrélations.

→ Calcul des contributions :

On peut calculer quelle est la contribution de chaque variable aux axes principaux.

$$\text{contribution de la variable } X_i \text{ à l'axe } C_k = \frac{\cos^2(X_i, C_k)}{\sum_{m=1}^{p} \cos^2(X_m, C_k)}$$

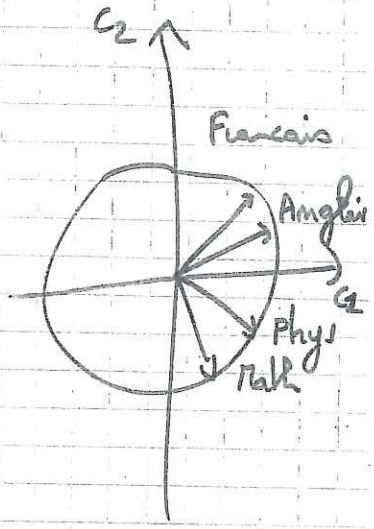
X_i : Ancienne variable

On obtient alors

	C_1	C_2	C_3	C_4
Math	22.89182%	30.46476%	4.1035%	42.5398%
Physique	28.29359%	16.54877%	19.465969%	35.691671%
Français	19.70742%	38.59312%	28.345812%	13.353649%
Anglais	29.10717%	14.39335%	48.08463%	8.414854%

l'axe C_1 est construit à 22.89% par la math, 28.3% par la physique, 19.7% par le français, 29.1% par l'anglais

l'axe C_2 est construit à 30.5% par la math, 6.5% par la physique, 38.6% par le français, 14.4% par l'anglais



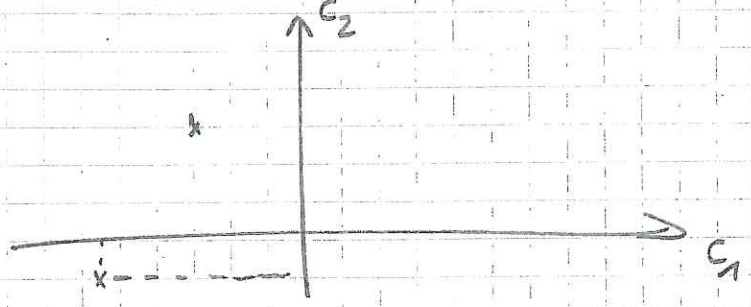
* Interpretation du diagramme des variables :

Soit θ l'angle entre 2 variables :

* si $\theta \simeq 0$ alors les deux variables sont positivement et linéairement corrélées
* si $\theta \simeq \pi$ alors les deux variables sont negativement et linéairement corrélées
* si $\theta \simeq \frac{\pi}{2}$ alors les deux variables sont independantes

ici :
* Physique et Math très corrélées ($\theta \simeq 0$)
* Français et Anglais très corrélés ($\theta \simeq 0$)
* Français et Math independant ($\theta \simeq \frac{\pi}{2}$)

→ Representation des individus :



* Les coordonnées des points ont déjà été calculées

$$C_i = \tilde{X} u_i$$

* $\cos^2$. Pour chaque individu, on calcule $\cos^2(X_i, C_k)$

$$\cos^2(X_i, C_k) = \frac{C_k^2}{\sum C_j^2} \quad \text{pour un individu donné.}$$

Exemple : Pour Jean : on calcule $\tilde{X}_{Jean} u_i = \begin{pmatrix} -2.7428277 \\ -0.4273962 \\ -0.023029 \\ -0.022619 \end{pmatrix} = C_i$

$$\cos^2(X_{Jean}, C_1) = \frac{(-2.7428277)^2}{\sum C_i^2} = \frac{7.523104}{7.706813} = 0.9761628$$

Si on représente les individus dans le plan (C_1, C_2)

il faut pour qu'un individu soit bien représenté il faut

$$\cos^2(X_{individu}, C_1) + \cos^2(X_{individu}, C_2) \simeq 1$$