Analyse en Composantes Principales (ACP)

Il s'agit d'une méthode d'analyse et de description multidimensionnelle d'un tableau de données quantitatives.

L'ACP permet (entre autre) des représentations graphiques des similitudes entre les lignes (des distances euclidiennes entre les individus) et des liaisons entre les colonnes (les corrélations entre les variables)

Il s'agit également d'une méthode de réduction de dimension (construction d'un "petit" nombre de variables synthétiques "résumant" au mieux toutes les variables initiales)

1) Notations

On considère (n) individus $\{1,\dots,i,\dots,n\}$ décrits sur (p) variables $\{1,\dots,j,\dots,p\}$. On notera w_i, $i=1,\dots,n$ le poids d'un individu i (généralement $\frac{1}{n}$).

1.1) Trois matrices de données

Données brutes : $X_{n\times p}$ Données centrées : $Y_{n\times p}$ Données centrées-réduites : $Z_{n\times p}$

$$X = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array}\begin{array}{c} \begin{array}{cccc} 1 & \cdots & j & \cdots & p \end{array} \\ \begin{pmatrix} & & & \\ & & x_{ij}\in\mathbb{R} & \\ & & & \end{pmatrix} \end{array}$$

$$Y = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array}\begin{array}{c} \begin{array}{cccc} 1 & \cdots & j & \cdots & p \end{array} \\ \begin{pmatrix} & & & \\ & & y_{ij} & \\ & & \uparrow & \\ & & y_i & \end{pmatrix} \end{array}$$

$$y_{ij} = x_{ij} - \bar{x}_j$$

$$Z = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array}\begin{array}{c} \begin{array}{cccc} 1 & \cdots & j & \cdots & p \end{array} \\ \begin{pmatrix} & & & \\ & & z_{ij} & \\ & & \uparrow & \\ & & z_i & \end{pmatrix} \end{array}$$

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} = \frac{y_j}{s_j}$$

$\Rightarrow$ 3 matrices différentes pour décrire les mêmes individus $\Rightarrow$ 3 nuages de points différents

Dans chaque matrice :
- un individu i est décrit par un vecteur de $\mathbb{R}^p$ noté x_i, y_i, z_i
- une variable j est décrite par un vecteur de $\mathbb{R}^n$ noté x^j, y^j, z^j

⚠ On parle toujours de vecteurs colonnes (même si ils décrivent les lignes des tableaux X, Z ou Y).

1.2) Deux métriques

On munit l'espace $\begin{cases} \mathbb{R}^p \text{ d'une métrique } M, \text{ de dimension } p \times p \\ \mathbb{R}^n \text{ d'une métrique } N, \text{ de dimension } n \times n \end{cases}$

L'ACP va consister à analyser les n points-individus (les lignes) et les p points-variables (les colonnes) de la matrice des données centrées-réduites, Z, avec les métriques :

$$M = \underbrace{\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}}_{p} = I_p \quad \text{sur } \mathbb{R}^p$$

$$N = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix} \text{ sur } \mathbb{R}^n \text{ avec en général } w_i = \frac{1}{n} \ \forall i = 1-n,$$

$$\text{donc } N = \frac{1}{n}\underbrace{\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}}_{n} = \frac{1}{n} I_n .$$

On dit qu'on fait l'analyse du triplet $(Z, I_p, N) \Leftrightarrow \begin{cases} \text{ACP normée} \\ \text{ACP sur matrice des corrélations} \end{cases}$

Remarque : Faire l'analyse du triplet (Y, M, N) avec $M = I_p$

$\Leftrightarrow \begin{cases} \text{ACP non normée} \\ \text{ACP sur matrice des covariances} \end{cases}$

Exemple :

	Intensité bulles	Saveur salée	Appréciation globale
St Yorre	3.9	6.4	2.9
Vichy	1.4	6.0	2.8
Quezac	5.1	4.7	3.5
Salvetat	2.9	4.1	3.4
Perrier	8.2	4.9	2.8
moyenne	4,28	5.22	3.08
écart-type	2.29	0.85	0.3

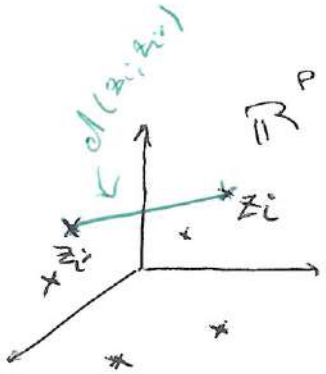
$$\Rightarrow Z = \begin{pmatrix} -0.17 & 1.38 & -0.58 \\ -1.26 & 0.91 & -0.91 \\ 0.35 & -0.61 & 1.37 \\ -0.6 & -1.31 & 1.04 \\ 1.69 & -0.37 & -0.91 \end{pmatrix} \qquad N = \begin{pmatrix} 1/5 & & 0 \\ & \ddots & \\ 0 & & 1/5 \end{pmatrix} = \frac{1}{5} I_5$$

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I_3$$

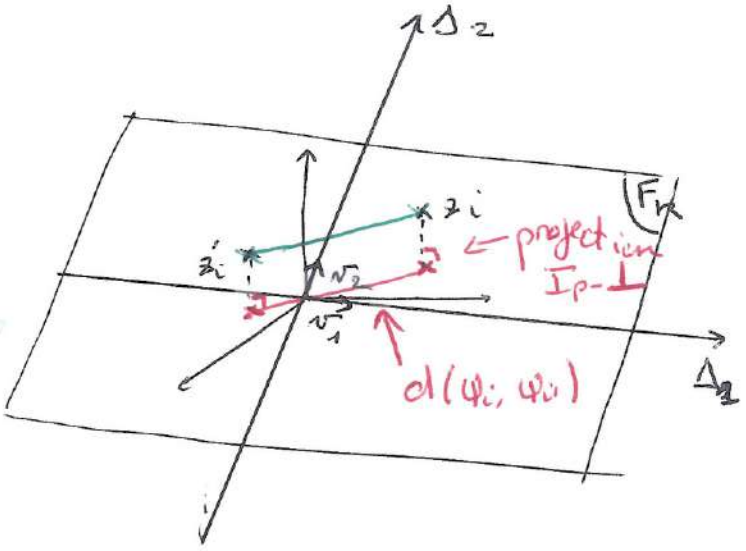
2) Analyse du nuage des n points-individus de $\mathbb{R}^p$

2.1) Le principe

On va chercher un sous-espace vectoriel F_k de $\mathbb{R}^p$, de dimension $k \leq r$ (avec $r = \text{rang}(Z)$), c'est-à-dire k axes $\Delta_1, \dots, \Delta_k$ (I_p-orthogonaux) tels que le nuage des n points individus I_p-projetés sur F_k, "déforme le moins possible" les distances euclidiennes entre les individus.

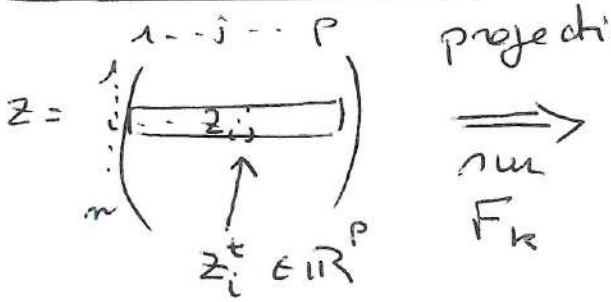


I_p-projection sur F_k ($k=2$) $\Longrightarrow$

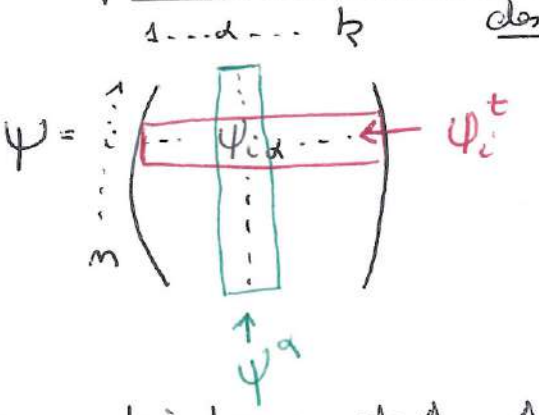


Nuage de n points-individus centrés-réduits

Matrice centrée-réduite



Matrice des coordonnées factorielles des individus



- On appelle $v_1 \dots v_a \dots v_k$ les vecteurs directeurs de $\Delta_1 \dots \Delta_k$ les <u>axes principaux</u>

- On appelle $\psi_{n \times k}$ la matrice des coordonnées des n <u>individus projetés</u> sur ces axes =

  { matrice des <u>coordonnées factorielles</u> des individus
  matrice des <u>scores</u> des individus sur les k premières <u>composantes principales</u>

- On appelle $\psi^a \in \mathbb{R}^n$ la "a-ième composante principale" c'est à dire le vecteur des scores des n individus sur cette composante

- On a $\psi_i \in \mathbb{R}^k$ le vecteur des coordonnées factorielles de l'individu i sur les k premiers axes = vecteur des scores de l'individu i sur les k premières composantes principales.

Que veut dire "déformer le moins" et quel critère va-t-on optimiser?

On a la relation suivante :

$$\underbrace{\sum_i \sum_{i'} w_i w_{i'} d^2(z_i, z_{i'})}_{\text{Somme pondérée des carrés des distances entre paires d'individus}} = 2 \underbrace{\sum_i w_i d^2(z_i, \bar{z})}_{\text{Inertie} = I(z) = p \text{ (déjà demo)}}$$

Somme des carrés des distances entre les individus et le centre de gravité = <u>mesure de dispersion du nuage</u>.

On a toujours que la distance entre i et i' dans $\mathbb{R}^p$ (en vert) plus grande que la distance "en projection" c'est à dire entre i et i' dans $\mathbb{R}^k$ (en rouge) :

$$d^2(z_i, z_{i'}) \geq d^2(\psi_i, \psi_{i'})$$

$\Rightarrow$ $I(z) = p$ est ⊕ grande que $I(\psi)$

$\Rightarrow$ on va chercher les axes qui maximisent $I(\psi)$

Donc "déformer" le moins possible le nuage des n points ind. dans $\mathbb{R}^p$ (centrés-réduits) $\Leftrightarrow$ maximiser $I(\psi)$

$$\text{Or } I(\psi) = \sum_{i=1}^{n} \omega_i \sum_{d=1}^{k} \left(\psi_{id} - \bar{\psi}^d\right)^2 = \sum_{d=1}^{k} \underbrace{\sum_{i=1}^{n} \omega_i \left(\psi_{id} - \bar{\psi}^d\right)^2}_{\text{var}(\psi^d)}$$

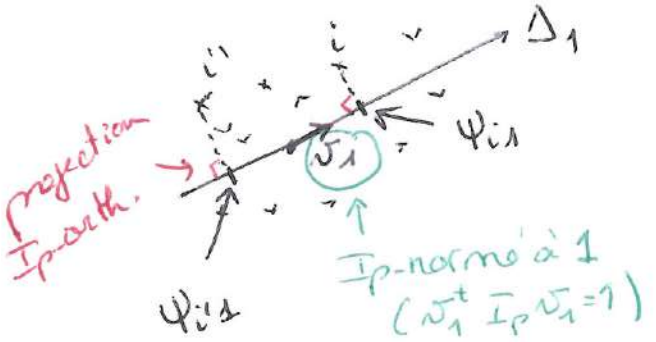
$$= \sum_{d=1}^{k} \text{var}(\psi^d)$$

$\Rightarrow$ On va chercher à maximiser $\text{var}(\psi^d)$ pour $d = 1 - k$.

Remarque : si $k = r$, $I(\psi) = I(z) = p$

2.2) Calcul des coordonnées factorielles des individus (scores)

• Coordonnées factorielles sur le premier axe.



$$\Rightarrow \quad \psi^1 = \begin{pmatrix} \psi_{11} \\ \vdots \\ \psi_{i1} \\ \vdots \\ \psi_{n1} \end{pmatrix} \leftarrow 1^{\text{ère}} \text{ colonne de la matrice } \psi.$$

Par définition :

$$\psi^1 = Z I_p v_1 = \underline{Z v_1}$$

Remarque : $\psi^1 = v_{11} \underbrace{\| z^1 \|}_{1^{\text{ère}} \text{ colonne de } Z} + \dots + v_{1p} \underbrace{\| z^p \|}_{p^{\text{ème}} \text{ colonne de } Z}$ = combinaison linéaire des colonnes de Z = variable "synthétique" résumant les colonnes de Z

Le problème d'optimisation: Trouver le vecteur $v_1 \in \mathbb{R}^p$, I_p-normé à 1, tel que $var(\psi^1)$ soit maximum.

Remarque: $\bar{\psi}^1 = 0 \Rightarrow var(\psi^1) = \|\psi^1\|_N^2$

Solution: v_1 est le vecteur propre associé à la plus grande valeur propre λ_1 de la matrice des corrélations

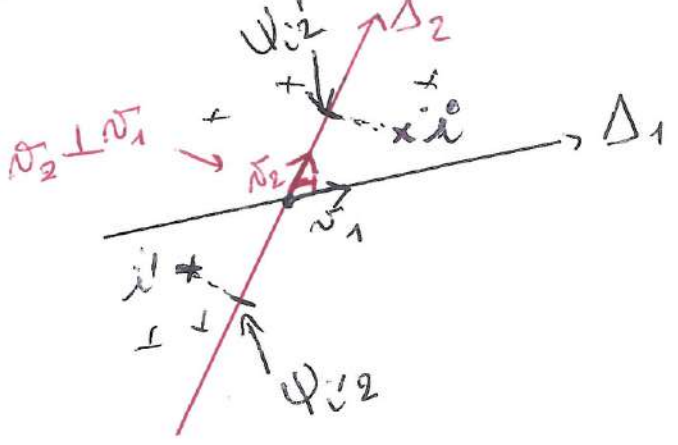
$$R = Z^t N Z$$

Exercice 1 : Démontrer ce résultat

Finalement on a:

- $\psi^1 = Z v_1$, v_1 est le vecteur propre associé à la plus grande valeur propre λ_1 de $R = Z^t N Z$
- $\bar{\psi}^1 = 0$
- $var(\psi^1) = \lambda_1$

- Coordonnées factorielles sur le second axe



$$\psi^2 = \begin{pmatrix} \psi_{12} \\ \vdots \\ \psi_{i2} \\ \vdots \\ \psi_{n2} \end{pmatrix} \leftarrow 2^{ème} \text{ colonne de } \psi.$$

Par définition : $\psi^2 = Z v_2$

Le problème d'optimisation : Trouver le vecteur $v_2 \in \mathbb{R}^p$, I_p-normé à 1 et I_p-orthogonale à v_1, tel que $var(\psi^2)$ soit maximum

Solution: v_2 est le vecteur propre associé à la grande valeur propre λ_2 de la matrice des corrélations R

Finalement on a :

$$\begin{cases} \cdot\ \psi^2 = Z v_2 \\ \cdot\ \bar{\psi}^2 = 0 \\ \cdot\ \mathrm{var}(\psi^2) = \lambda_2 \end{cases}$$

- Idem pour les axes suivants c'est à dire pour $\alpha = 3, \dots, k$ (et $k \leq r$)

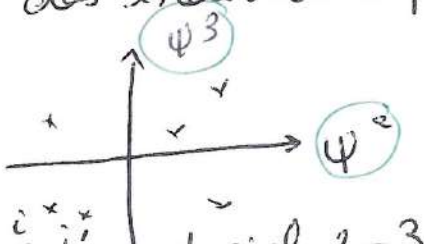
- Conclusion : Pour calculer la matrice $\psi_{n \times k}$ des coordonnées factorielles des individus (scores des individus), on effectue la décomposition spectrale de la matrice $R = Z^t N Z$ des corrélations : on cherche les k vecteurs propres $v_1, \dots, v_k$ de R (p-normés à 1 et associés aux k plus grandes valeurs propres triées par ordre décroissant $\lambda_1, \dots, \lambda_k$ et on calcule $\begin{cases} \psi^\alpha = Z v_\alpha \\ \mathrm{var}(\psi^\alpha) = \lambda_\alpha \end{cases}$

Ces calculs permettent :

→ Des représentations graphiques des individus projetés

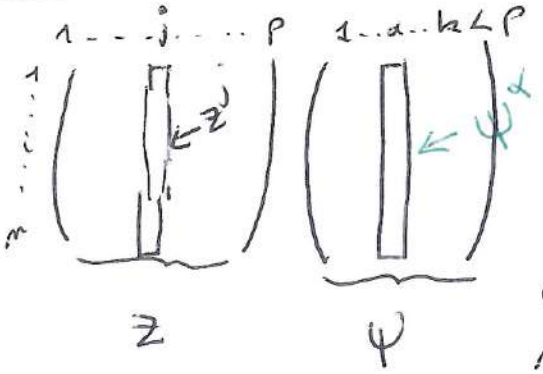
Plan factoriel 1-2

⚠ Qualité de la projection !



Plan factoriel 2-3

→ Une réduction de la dimension :



Les p variables (colonnes) $z^1 \dots z^p$ (éventuellement corrélées) sont remplacées par k variables synthétiques ψ^α, $\alpha = 1-k$, appelées composantes principales et non corrélées : $\mathrm{cor}(\psi^\alpha, \psi^{\alpha'}) = 0 \ \forall \alpha \neq \alpha'$

Exemple :

Matrice des corrélations R :

	Intensité bulles	Saveur salé	Appréc. globale
Int. bulles	1	-0.29	-0.09
Saveur salé	-0.29	1	-0.70
app. globale	-0.09	-0.70	0

Décomposition spectrale de R : $\lambda_1 = 1.73$, $\lambda_2 = 1.06$, $\lambda_3 = 0.20$

	v_1	v_2	v_3
Int. bulles	0.20	0.92	-0.32
Sav. salé	-0.71	-0.08	-0.68
Appréc. globale	0.66	-0.37	-0.64

Normés : $\|v_1\| = \|v_2\| = \|v_3\| = 1$ et $v_1 \perp v_2 \perp v_3$; 1ère composante principale

$\Rightarrow V^t V = I_3$

Matrice ψ des coordonnées factorielles (scores)

$\psi^1 = Z v_1$, $\psi^2 = Z v_2$, $\psi^3 = Z v_3$ $\Rightarrow \psi = Z V$ $\Rightarrow$

	ψ^1	ψ^2	ψ^3
St Yorre	-1.42	-0.06	-0.51
Vichy	-1.52	-0.90	0.37
Quezac	1.42	-0.13	-0.58
Salvetat	1.51	-0.89	0.42
Perrier	0.00	1.94	0.30

$= \psi$

Moyenne	0	0	0
Variance	1.73 (λ_1)	1.06 (λ_2)	0.18 (λ_3)

et $\psi^1 \perp \psi^2 \perp \psi^3$

et $p = 3 = \lambda_1 + \lambda_2 + \lambda_3$

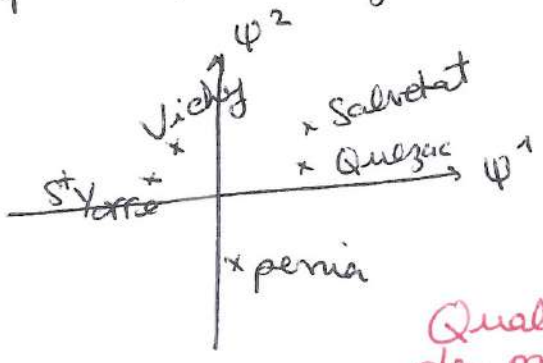
$I(Z)$ $\quad$ $I(\psi)$

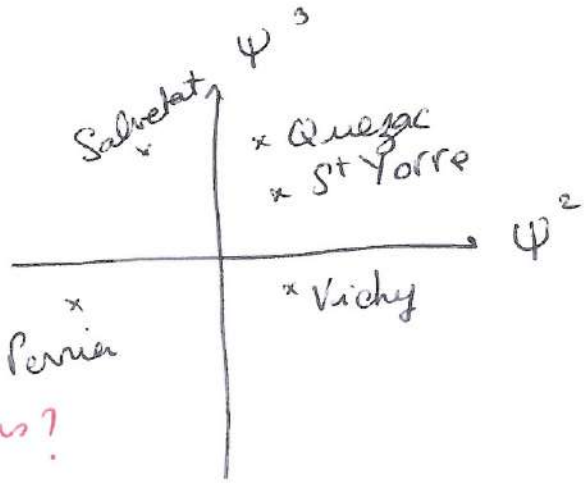
Variables synthétiques :

$\psi^1 = 0.2\, z^1 + 0.71\, z^2 + 0.66\, z^3$

z^1 : Int. bulles ; z^2 : Sav. salé ; z^3 : App globale.

Représentations graphiques :





Qualité des projections ?

2.3) Inerties associées aux axes.

2.3.1) Pourcentage de l'inertie du nuage de points initial expliqué par l'axe d :

- Ce pourcentage vaut :

$$\frac{\lambda_d}{\lambda_1 + \dots + \lambda_p} \times 100 = \frac{\lambda_d}{p}$$

λ_d ← inertie (variance) du nuage des individus projeté sur Δ_d $= I(\varphi_d)$

$\lambda_1 + \dots + \lambda_p = p = I(Z)$ = inertie du nuage initial

: mesure le "pouvoir explicatif" de l'axe d.

- Le pourcentage d'inertie expliquée moyen par axe vaut $\frac{1}{p} \times 100$ car la valeur propre moyenne vaut 1

⚠ Il faut faire attention lorsqu'on interprète le pourcentage d'inertie expliqué par un axe car cela dépend du nombre p de variables. Par exemple 10% peut être beaucoup si $p = 100$ et peu si $p = 10$. En effet $0,1 = \frac{10}{100}$ ($\rightsquigarrow \lambda_d > 1$, $p = 100$) et $0,1 = \frac{1}{10}$ ($\rightsquigarrow \lambda_d = 1$, $p = 10$)

2.3.2) Pourcentage d'inertie expliquée par les k premiers axes.

- Ce pourcentage vaut

$$\frac{\lambda_1 + \dots + \lambda_k}{p} \times 100$$

- Pour "savoir" combien d'axes retenir pour modéliser le nuage de points (ou réduire la dimension) sans "manquer" trop d'information, il existe plusieurs méthodes.

Par exemple :

→ Choisir k pour avoir 90% d'inertie expliquée par exemple

→ Choisir k tel que $\lambda_k > 1$ et $\lambda_{k+1} < 1$: règle de Kaiser

→ Utiliser un critère statistique

Exemple des eaux :

	λ_k	%ages	% ages cumulés	
Axe 1	1.73	57.7	57.7	} 2 axes retenus
Axe 2	1.06	35.4	93.3	
Axe 3	0.20	6.7	100	

3) Analyse du nuage des p points-variables de $\mathbb{R}^n$

On va "calquer" l'analyse des individus pour faire cette analyse des variables.

3.1) Le principe

On va chercher un sous-espace vectoriel G_k de $\mathbb{R}^n$, $\Delta^1 \perp \Delta^2 \neq \Delta_1$

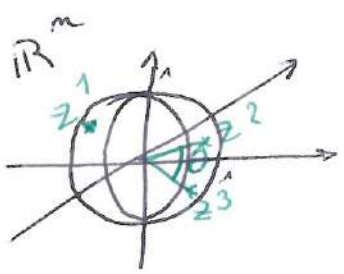
de dimension $k \leq r$, c'est à dire k axes $\Delta^1 \dots \Delta^k$

(N-orthogonaux) tels que le nuage des p points-variables N-projetés sur G_k, "déforme" le moins possible les distances entre les variables et donc leurs corrélations.

En effet, dans une matrice centrée-réduite :

* les points-variables z^j sont sur l'hypersphère unité

$\forall j,\ \|z^j\|_N = \text{var}(z^j) = 1$

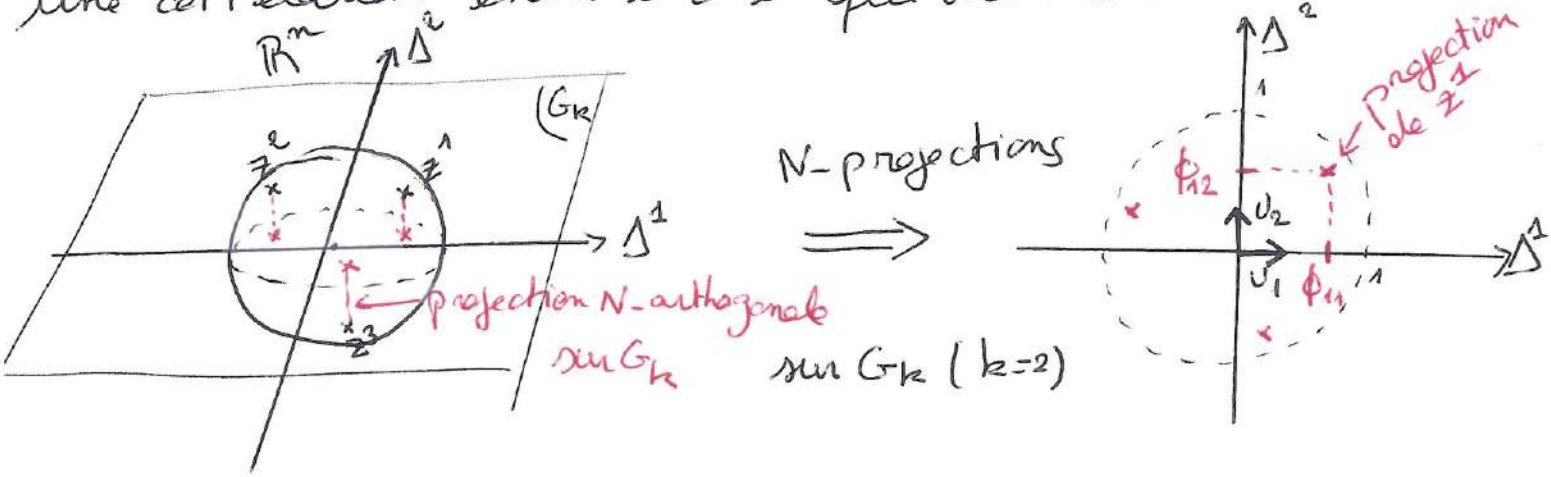


* la corrélation entre deux variables x^j et $x^{j'}$ est égale au cosinus de l'angle entre z^j et $z^{j'}$:

$r_{jj'} = \langle z^j, z^{j'} \rangle_N = \cos_N \theta$

$\Rightarrow$ Dans $\mathbb{R}^n$, si z^j et $z^{j'}$ sont proches en terme de distance (avec la métrique N), alors x^j et $x^{j'}$ sont corrélées.

Exercice 2 : Montrer que $d_N^2(z^i, z^{i'}) = 2(1 - r_{ii'})$ où $r_{ii'}$ = corrélation entre x^i et $x^{i'}$. En déduire que la distance entre deux points variables z^i et $z^{i'}$ varie entre 0 et 2 pour une corrélation entre x^i et $x^{i'}$ qui varie entre −1 et 1.



Matrice centrés-réduite :

$$Z = \begin{array}{c} \\ 1 \\ i \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & \cdots & j & \cdots & p \end{array} \\ \begin{pmatrix} & & \vdots & & \\ \cdots & & z_{ij} & & \cdots \\ & & \vdots & & \end{pmatrix} \end{array}$$

$z^j \in \mathbb{R}^n$

N-projections sur G_k ⟹

Matrice des coordonnées factorielles des variables

$$\Phi = \begin{array}{c} \\ 1 \\ j \\ p \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & \cdots & d & \cdots & k \end{array} \\ \begin{pmatrix} & & \vdots & & \\ \cdots & & \phi_{jd} & & \cdots \\ & & \vdots & & \end{pmatrix} \end{array}$$

Φ^d

- On note $u_1, \ldots u_d, \ldots u_k$ les vecteurs directeurs de axes $\Delta^1, \ldots \Delta^k$.

  On verra que $v^d = \dfrac{\psi^d}{\sqrt{\lambda_d}}$ = $d^{ème}$ composante principale standardisée

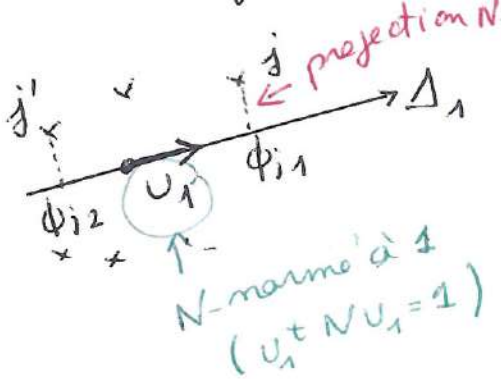
- On appelle $\Phi_{p \times k}$ la matrice des coordonnées des p points-variables projetés sur ces axes =

  } matrice des coordonnées factorielles des variables
  matrice des loadings (saturations) des variables sur les k premières composantes principales.

- On notera Φ^d la d-ème colonne de Φ et on verra que les loadings sont des corrélations : $\Phi_{jd} = \text{corr}(\psi^d, x^j)$

3.2) Calcul des coordonnées factorielles des variables (loadings)

- coordonnées factorielles sur le 1er axe :



$$\Rightarrow \phi^1 = \begin{pmatrix} \phi_{11} \\ \vdots \\ \phi_{j1} \\ \vdots \\ \phi_{p1} \end{pmatrix} \in \mathbb{R}^p$$

($i^{ère}$ colonne de Φ)

Par définition :

$$\phi^1 = Z^t N U_1$$

Le problème d'optimisation : Trouver le vecteur $U_1 \in \mathbb{R}^n$, N-normé à 1, tel que $\|\phi^1\|^2_{I_p}$ soit maximale.

Solution : U_1 est le vecteur propre associé à la plus grande valeur propre λ_1 de la matrice $Z Z^t N$.

Remarque : On verra plus tard que λ_1 est aussi la plus grande valeur propre de $R = Z^t N Z$

Exercice 3 : Demontrer le resultat "selection".

- coordonnées factorielles sur le second axe :

Par définition :

$$\phi^2 = Z^t N U_2$$

Le problème : Trouver U_2 tel que $U_2^t N U_2 = 1$ et $U_2^t N U_1 = 0$ qui maximise $\|\phi^2\|^2$

Solution : U_2 est le vecteur propre, N-normé à 1, de $Z Z^t N$ associé à la seconde plus grande valeur propre λ_2.

- Idem pour les axes suivant c'est à dire pour $\alpha = 3, \ldots, k$.

- Conclusion : Pour calculer la matrice $\phi_{p\times k}$ des coordonnées factorielles des variables (loadings des variables), on effectue la décomposition spectrale de la matrice $Z Z^t N$ (appelée matrice des produits scalaires des individus) : on cherche les k vecteurs propres $U_1, \dots, U_k$ de $Z Z^t N$, N-normés à 1, associés aux valeurs propres $\lambda_1 \dots \lambda_k$ triées par ordre décroissant et on calcule $\phi^\alpha = Z^t N U_\alpha$.

- Parallèle avec les coordonnées factorielles des individus

v_α, $\begin{cases} \text{vecteur propre de } Z^t N Z I_p \\ I_p\text{-normé à 1} \end{cases}$ et $\psi^\alpha = Z I_p v_\alpha$

U_α, $\begin{cases} \text{vecteur propre de } Z I_p Z^t N \\ N\text{-normé à 1} \end{cases}$ et $\phi^\alpha = Z^t N U_\alpha$

$\Rightarrow$ on "échange" lignes et colonnes et les métriques

- Ecriture matricielle des scores des individus et des loadings des variables

Matrice $n\times k$ des scores : $\underset{n\times k}{\Psi} = \underset{n\times p}{Z}\,\underset{p\times k}{V}$ où V = matrice dont les colonnes sont les k vect. propres de $R = Z^t N Z$

Matrice $p\times k$ des loadings : $\underset{p\times k}{\Phi} = \underset{p\times n}{Z^t}\,\underset{n\times U}{U}$ où U = matrice dont les colonnes sont les k vect. propres de $Z Z^t N$.

Exemple :

Matrice des produits scalaires des individus $Z Z^t N$:

	St Yorre	Vichy	Quezac	Salvetat	Perrier
St Yorre	0.46	0.41	-0.34	-0.47	-0.06
Vichy	0.41	0.65	-0.45	-0.28	-0.33
Quezac	-0.34	-0.45	0.48	0.41	-0.09
Salvetat	-0.47	-0.28	0.41	0.64	-0.30
Perrier	-0.06	-0.33	-0.09	-0.3	0.77

matrice symétrique.

Décomposition spectrale de $Z Z^t N$.

$$\begin{cases} \lambda_1 = 1.73 \\ \lambda_2 = 1.06 \\ \lambda_3 = 0.20 \\ \lambda_4 = 0 \\ \lambda_5 = 0 \end{cases}$$

U :

	U_1	U_2	U_3
St Yorre	0.48	-0.03	-0.5
Vichy	-0.52	-0.39	0.37
Quezac	0.48	-0.06	-0.58
Salvetat	0.52	-0.36	0.42
Perrier	0.00	0.84	0.30

$\Rightarrow$

U_1	U_2	U_3
-1.08	-0.06	-1.15
-1.16	-0.88	0.82
1.08	-0.13	-1.29
1.15	-0.81	0.95
0.00	1.88	0.67

I_5-Norme : 1 1 1 — N-norme : 1 1 1

pour N-normer à 1, on divise chaque colonne par $\sqrt{\quad}$ $\left(a = \frac{b}{(b^t N b)^{1/2}}\right)$

$$\begin{cases} U_1^t N U_2 = 0 \\ U_2^t N U_3 = 0 \\ U_1^t N U_3 = 0 \end{cases}$$

$\Downarrow$

$$U^t N U = I_3$$

Matrice ϕ des coordonnées factorielles des variables (loadings) :

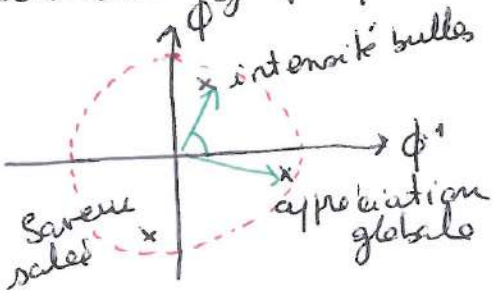
$$\begin{cases} \phi^1 = Z^t N U_1 \\ \phi^2 = Z^t N U_2 \\ \phi^3 = Z^t N U_3 \end{cases} \Rightarrow \phi = \underbrace{Z^t}_{p \times n} \underbrace{U}_{n \times 3} \Rightarrow$$

corrélation entre qual. saléé et ϕ^1

	ϕ^1	ϕ^2	ϕ^3
Intensité bulles	0.26	0.95	-0.15
Saveur salée	-0.95	-0.09	-0.29
Appréciation globale	0.88	-0.38	-0.29

$= \phi$

Représentations graphiques :



* Ajout de flèche pour aider à l'interprétation : si points bien projetés, un angle faible ($\Leftrightarrow$) forte corrélation

* Ajout d'un cercle (dit des corrélation) : si point bien projeté alors il est proche du cercle.

4) Lien entre l'analyse des individus et des variables

4.1) Décomposition en valeurs singulières de Z (DVS)

La DVS (ou SVD pour Singular Value Decomposition) de la matrice réelle Z, de rang r, est avec les métriques N sur $\mathbb{R}^n$ et I_p sur $\mathbb{R}^p$:

$$\underset{n\times p}{Z} = \underset{n\times r}{U}\ \underset{r\times r}{\Lambda}\ \underset{r\times p}{V^t}$$

où

$\Lambda = \text{diag}(\sqrt{\lambda_1}, \dots \sqrt{\lambda_r})$ avec λ_α valeur propre de (valeurs singulières)

$$\begin{cases} Z^t N Z I_p = R \\ Z I_p Z^t N = Z Z^t N \end{cases}$$

U est la matrice dont les colonnes sont les vecteurs propres de ZZ^tN (triés par ordre décroissant des valeurs propres), N-normés et N-orthogonaux : $U^t N U = I_r$

V est la matrice dont les colonnes sont les vecteurs propres de Z^tNZ (triés par ordre décroissant des valeurs propres) et $V^t V = I_r$

$\Rightarrow$ Cette décomposition en valeurs singulières donne directement les matrices Ψ et Φ des scores et loadings :

$$\begin{cases} \Psi = Z V \\ \Phi = Z^t N U \end{cases}$$

⚠ En pratique, les logiciels font les SVD avec les métriques I_n et I_p

Donc on pose :

$\tilde{U} = N^{1/2} U \Rightarrow \tilde{U}^t \tilde{U} = U N U = I_r$

$\Rightarrow Z = N^{-1/2} \tilde{U} \Lambda V^t$

$\Rightarrow \underbrace{N^{1/2} Z}_{\tilde{Z}} = \tilde{U} \Lambda V^t \Rightarrow$ SVD "classique" de $\tilde{Z}$ et $U = N^{-1/2} \tilde{U}$

4.2) Formules de passage (ou de transition)

$$\begin{cases} v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \underbrace{Z^t N u_\alpha}_{\phi^\alpha} \\ u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \underbrace{Z v_\alpha}_{\psi^\alpha} \end{cases}$$

Exercice 4 : Demontrer ce resultat.

On en déduit deux resultats importants :

* $u_\alpha = \frac{\psi^\alpha}{\sqrt{\lambda_\alpha}}$ = $\alpha^{ème}$ composante principale standardisée

car $\lambda_\alpha = var(\psi^\alpha) \Rightarrow \sqrt{\lambda_\alpha}$ est son écart-type

$\Rightarrow$ U est la matrice des scores standardisés

* $\phi^\alpha = \sqrt{\lambda_\alpha}\, v_\alpha$
  $\psi^\alpha = \sqrt{\lambda_\alpha}\, u_\alpha$
  $\Rightarrow$ $\boxed{\begin{aligned} \phi &= V\Lambda \\ \psi &= U\Lambda \end{aligned}}$

$\Rightarrow$ La SVD de Z donne directement ϕ et ψ.

Exemple :

	v_1		ϕ^1
intensité	0.20		0.26
bulles	-0.71	$\times \sqrt{1.73} =$	-0.95
saveur salé	0.66		0.88
appréciation globale			

idem pour v_2, v_3, ϕ^2, ϕ^3

	u_1		ψ^1
St Yorre	-1.08		-1.42
Vichy	-1.16		-1.52
Quezac	1.08	$\times \sqrt{1.73} =$	1.42
Salvetat	1.15		1.51
Perrier	0.00		0.00

idem pour u_2, u_3, ψ^2, ψ^3

Remarque : Si on prend k colonnes de U, Λ et V,

on a $Z = \underbrace{U^{(k)} \Lambda^{(k)} V^{(k)}}_{\hat{Z}} + E_k$

$\hat{Z}$: meilleure approximation des moindres carrés de Z par une matrice de rang $k \leq r$, au sens de la norme de Hilbert-Schmidt.

$$\|A\|_{M,N} = \sqrt{tr(AMA^tN)} \quad \text{avec} \begin{cases} M = \frac{1}{p} I_p \\ N = diag(\omega_i) \end{cases}$$

$\Rightarrow \|Z - \hat{Z}\|_{M,N}$ est minimum.

4.3) Corrélations variables - composantes principales

Les coordonnées factorielles des variables (les loadings) vérifient la propriété suivante :

$$\phi_{jl} = cor(\psi^l, x^j)$$

↑ corrélation.

Remarque : En ACP non normé,

$$\phi_{jl} = s^j cor(\psi^l, x^j)$$

Exercice 5 : Démontrer ce résultat.

Ce résultat est indispensable pour pouvoir interpréter les plans factoriels des individus en fonction de celui des variables.

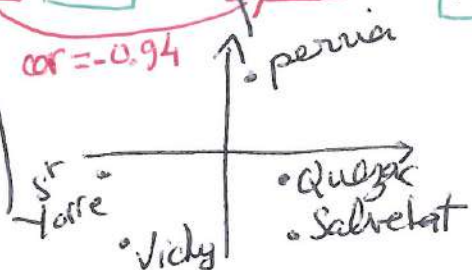
	Intensité bulles	Saveur salés	App. globale
S^t Yorre	3.9	6.4	2.9
Vichy	1.4	6.0	2.8
Quezac	5.1	4.7	3.5
Salvetat	2.9	4.1	3.4
Perrier	8.2	4.9	2.8

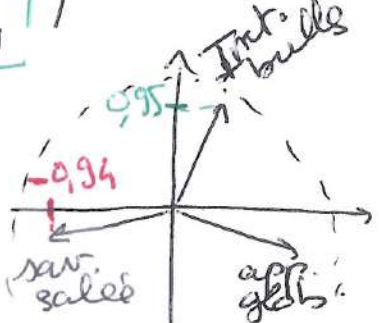
	ψ^1	ψ^2
S^t Yorre	-1.4	-0.06
Vichy	-1.52	-0.90
Quezac	1.42	-0.13
Salvetat	1.51	-0.89
Perrier	0.00	1.94

cor = -0.95 (Intensité bulles / ψ^1)

cor = -0.94 (Saveur salés / ψ^2)

	ϕ^1	ϕ^2
Intens. bulles	0.26	0.95
Saveur salés	-0.94	-0.08
App. globale	0.87	-0.38





Exercice 6 : A votre avis, pourquoi les eaux Quezac et Salvetat sont ⊕ appréciées et moins salées tandis que St Yorre est ⊕ salée ? Qu'est-ce qui caractérise l'eau de Perrier ?

5) Interpretation des resultats

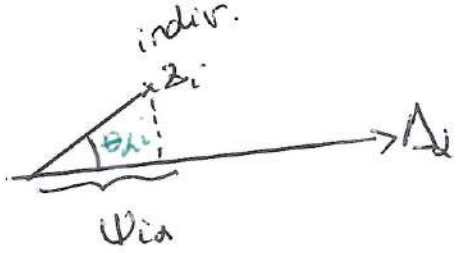
5.1) Plans factoriels des individus

a) Deux individus bien projetés et proches en projection sont effectivement proches de $\mathbb{R}^p$

⇒ Comment savoir si ils sont bien projetés ? : on calcule le cosinus carré de l'angle entre le point-individu et l'axe (ou le plan)

* Qualité de la representation de l'individu i sur l'axe Δ_d :

$$QLT_d(i) = \cos^2(\theta_{di}) = \frac{\psi_{id}^2}{\|z_i\|^2}$$



$$\Rightarrow \begin{cases} \cos^2 = 1 & \text{si } i \text{ parfaitement projeté} \\ = 0 & \text{si très mal projeté} \end{cases}$$

* Qualité de la representation de l'individu i sur le plan $(\Delta_d, \Delta_{d'})$:

$$\underbrace{QLT_{d,d'}(i)}_{\cos^2} = \frac{\psi_{id}^2 + \psi_{id'}^2}{\|z_i\|^2}$$

← distance au carré de "i projeté" à l'origine du plan

⚠ Dans SPAD, $QLT_{dd'}(i)$ définit le diamètre des points dans les graphiques des individus en choisissant en option que leur taille soit proportionnelle au $\cos^2$.

b) Les individus qui contribuent de manière excessive aux axes factoriels sont source d'instabilité (si on les retire, cela change "fond"). Ils sont donc souvent retirés de l'analyse ou mis en illustratif

=> Comment savoir si un individu contribue à un axe ? : on calcule la part (le pourcentage si on ⊗ par 100) de l'inertie de l'axe expliquée par l'individu i.

* Contribution relative d'un individu i à la construction de l'axe Δ_α :

$$\underbrace{CTR_\alpha(i)}_{\text{contribution}} = \frac{w_i \psi_{i\alpha}^2}{\lambda_\alpha}$$

Rappel : $\lambda_\alpha = \sum_{i=1}^{n} w_i (\psi_{i\alpha} - \bar{\psi}^{\alpha})^2$ ($\bar{\psi}^{\alpha} = 0$)

$= \sum_{i=1}^{n} w_i \psi_{i\alpha}^2$

=> $\sum_{i=1}^{n} CTR_\alpha(i) = 1$

=> s'exprime bien en pourcentage si ⊗ par 100.

⚠ Si tous les poids w_i sont identiques, ce qui est le cas avec $w_i = \frac{1}{n}$, les individus les plus excentrés (pour lesquels $|\psi_{i\alpha}|$ est max) sont ceux qui contribuent le plus.

Remarque : Contribution absolue = $w_i \psi_{i\alpha}^2$

* Contribution relative d'un individu i à la construction du plan $(\Delta_\alpha, \Delta_{\alpha'})$:

$$CTR_{\alpha\alpha'}(i) = w_i \frac{(\psi_{i\alpha}^2 + \psi_{i\alpha'}^2)}{\lambda_1 + \lambda_2}$$

Là encore, $\sum_i CTR_{\alpha\alpha'}(i) = 1$

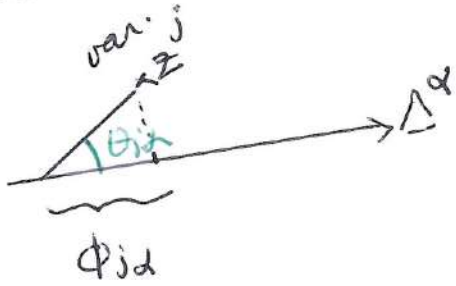
⚠ Dans SPAD, $CTR_{dd'}(i)$ définit le diamètre des points dans les graphiques des individus en choisissant en option que leur taille soit proportionnelle à leur contribution

5.2) Plans factoriels des variables.

L'angle entre deux variables bien projetées donne une idée de leurs corrélations

* Qualité de la projection de la variable j sur l'axe Δ^α : $\cos^2$ de l'angle entre le point-variable et l'axe :

$$\underbrace{QLT_\alpha(j)}_{\cos^2} = \cos(\theta_{j\alpha}) = \frac{\phi_{j\alpha}^2}{\|z^j\|^2} = 1 \text{ par construction}$$



* Qualité de la projection de la variable j sur le plan $(\Delta^\alpha, \Delta^{\alpha'})$:

$$\underbrace{QLT_{dd'}(j)}_{\cos^2} = \underbrace{\phi_{j\alpha}^2 + \phi_{j\alpha'}^2}_{\text{carré de distance entre point-projeté et origine du plan.}}$$

Donc $QLT_{dd'} = 1$ si le point-variable projeté est sur le cercle $\Rightarrow$ Qualité de projection est d'autant meilleure que la flèche est proche du cercle des corrélations

5.3) Interprétation des plans des individus à partir de celui des variables :

On utilise $\phi_{j\alpha} = corr(y^\alpha, z^j)$