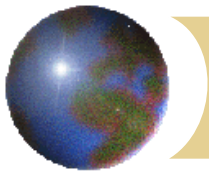


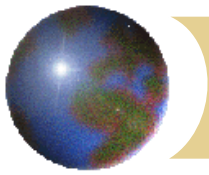
CHAPITRE 5

Analyse en Composantes Principales



Analyse de la structure de la matrice variance-covariance

- ✚ **Analyse de la variabilité et de la dispersion des données**
- ✚ Excepté si l'une des variables peut s'exprimer comme une fonction d'autres, on a besoin des p variables pour prendre en compte toute la variabilité du système
- ✚ **Objectif de l'ACP: *décrire*** à l'aide de $q < p$ composantes ***un maximum*** de cette variabilité => trouver des **descripteurs "efficaces"** des données.
- ✚ Utilité:
 - ▣ réduction de la dimension des données
 - ▣ outil de visualisation des données à 2 ou 3 dimensions (si $q = 2$ ou 3)
 - ▣ interprétation des données : liaisons inter-variables
 - ▣ Etape intermédiaire souvent utilisée avant d'autres analyses !



Analyse de la structure de la matrice variance-covariance

- **Recherche des composantes principales**

Propriétés statistiques désirées:

Nouveau système de q ($< p$) variables: $C_1, C_2, \dots, C_k, \dots, C_q$

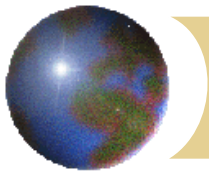
- non corrélées (représentation optimale)
- de variance maximale (respect de la dispersion des données initiales)
- d'importance décroissante ($\Rightarrow$ permet de se limiter au k meilleures)
- obtenues par combinaisons linéaires des variables d'origine centrées par rapport à leur moyenne $Y_1, \dots, Y_p$

$$C_k = a_{1k} Y_1 + a_{2k} Y_2 + \dots + a_{pk} Y_p \quad \Rightarrow \quad \text{coefficients } a_{jk} \text{ à déterminer}$$

Propriétés géométriques désirées:

Représentation approchées du nuage des n individus dans un sous-espace de dimension réduite.

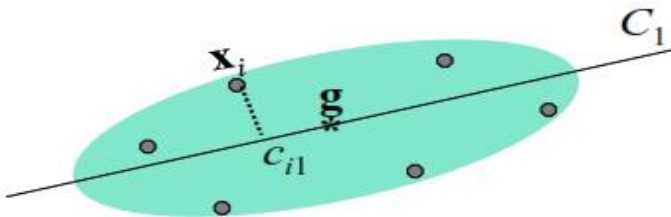
$\Rightarrow$ Espace de projection avec distorsion minimale des distances projetées.



Définition des composantes par constructions successives

Définition des composantes par constructions successives:

C_1 détermine une nouvelle direction dans le nuage de points qui suit l'axe d'allongement (étirement) maximal du nuage.



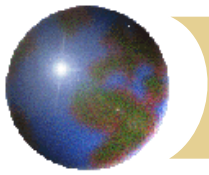
c_{i1} = coordonnée du point i sur l'axe C_1
= projection de $\mathbf{x}_i$ sur C_1

$$c_{i1} = \sum_{j=1}^p a_{1j} (x_{ij} - \bar{x}_j) = \sum_{j=1}^p a_{1j} y_{ij}$$

C_1 de variance maximale $\Rightarrow$ les projections c_{i1} sont les plus dispersées possible. Pour fixer la droite, on impose qu'elle passe par $\mathbf{g}$ (centre de gravité, sinon l'ensemble des droites parallèles conviendrait).

$\Rightarrow C_1$ = droite passant par le centre de gravité, réalisant le meilleur ajustement possible du nuage c-à-d : qui conserve au mieux la distance entre les points (après projection)

$\Rightarrow$ droite de projection assurant une distorsion minimale.

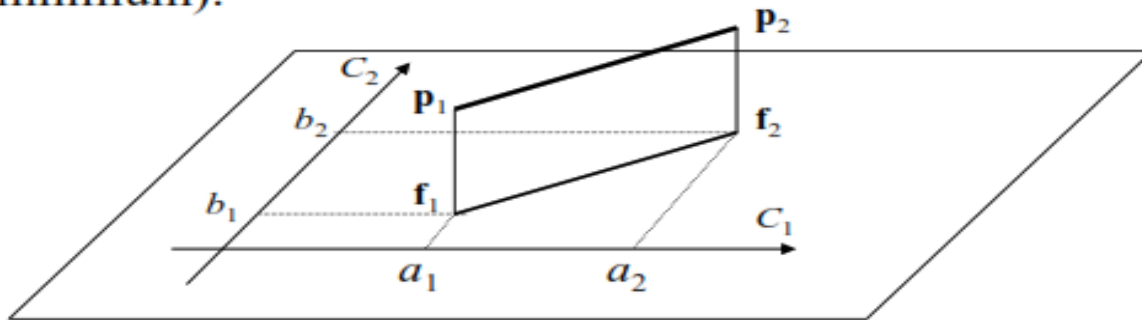


Définition des composantes par constructions successives

$C_2 = 2^{\text{ème}}$ composante, **orthogonale à C_1** et de variance maximale.

$\Rightarrow C_2$ détermine un axe perpendiculaire à C_1 (au point g) suivant la 2^{ème} direction d'allongement maximum du nuage.

$\Rightarrow C_1$ et C_2 déterminent le **plan principal** : le meilleur plan de projection (de distorsion minimum).

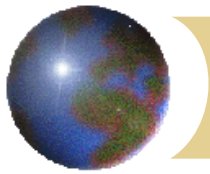


C_1 est telle que la moyenne des $d^2(a_i, a_i')$ max.

C_2 est $\perp$ à C_1 et telle que la moyenne des $d^2(b_i, b_i')$ max.

$\Rightarrow C_1$ et C_2 déterminent le plan tel que $d^2(f_i, f_i')$ soit maximum.

$\Rightarrow C_3$ est la droite $\perp$ à C_1 et C_2 (par g) telle que la variance des coord. soit maximum ...

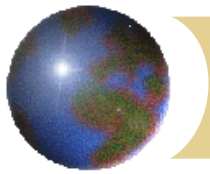


Principe et critère

Principe on cherche à projeter le nuage de points sur un espace F_k de dimension $k < p$.

Critère on veut que la moyenne des carrés des distances entre les points projetés soit maximale (elle est toujours plus petite que pour le nuage original).

Pour cela on cherche F_k , sous espace de dimension k de $\mathbb{R}^k$, tel que l'inertie du nuage projeté sur F_k soit maximale.



Valeurs propres et vecteurs propres

Définition un vecteur $\mathbf{v}$ de taille p est un *vecteur propre* d'une matrice $\mathbf{A}$ de taille $p \times p$ s'il existe $\lambda \in \mathbb{C}$ telle que

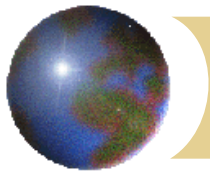
$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

λ est une *valeur propre* de $\mathbf{A}$ associée à $\mathbf{v}$.

Domaine En général, les vecteurs propres et valeurs propres sont complexes; dans tous les cas qui nous intéressent, ils seront réels.

Interprétation des vecteurs propres ce sont les directions dans lesquelles la matrice agit.

Interprétation des valeurs propres c'est le facteur multiplicatif associé à une direction donnée.



Valeurs et vecteurs propres : cas particuliers

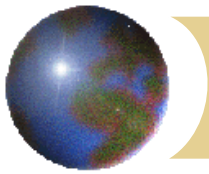
Matrice nulle sa seule valeur propre est 0, et tout vecteur est vecteur propre.

Matrice identité tout vecteur est vecteur propre de $\mathbf{I}$ avec valeur propre 1, puisque $\mathbf{I}\mathbf{v} = \mathbf{v}$.

Matrice diagonale si $\mathbf{D}_\lambda$ est une matrice diagonale avec les coefficients $\lambda_1, \dots, \lambda_p$, alors le i -ème vecteur coordonnée est vecteur propre de $\mathbf{D}_\lambda$ associé à la valeur propre λ_i .

L'action d'une matrice diagonale est de multiplier chacune des coordonnées d'un vecteur par la valeur propre correspondante.

Matrice diagonalisable c'est une matrice dont les vecteurs propres forment une *base* de l'espace vectoriel : tout vecteur peut être représenté de manière unique comme combinaison linéaire des vecteurs propres. Une matrice de taille $p \times p$ qui a p valeurs propres réelles distinctes est diagonalisable dans $\mathbb{R}$.



Quelques matrices diagonalisables

Matrice symétrique une matrice symétrique réelle ($\mathbf{A}' = \mathbf{A}$) possède une base de vecteurs propres orthogonaux et ses valeurs propres sont réelles

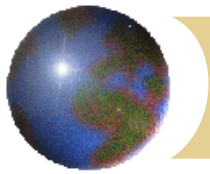
$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \text{ si } i \neq j, \quad \text{et } \lambda_i \in \mathbb{R}.$$

Matrice $\mathbf{M}$ -symétrique une matrice $\mathbf{M}$ -symétrique réelle ($\mathbf{A}'\mathbf{M} = \mathbf{M}\mathbf{A}$) possède une base de vecteurs propres $\mathbf{M}$ -orthogonaux et ses valeurs propres sont positives ou nulles

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle_{\mathbf{M}} = 0 \text{ si } i \neq j, \quad \text{et } \lambda_i \in \mathbb{R}.$$

Matrice définie positive c'est une matrice symétrique dont les valeurs propres sont strictement positives et donc

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \text{ si } i \neq j, \quad \text{et } \lambda_i > 0.$$



Analyse de de VM

Valeurs propres la matrice $\mathbf{VM}$ est $\mathbf{M}$ -symétrique : elle est donc diagonalisable et ses valeurs propres $\lambda_1, \dots, \lambda_p$ sont réelles.

Vecteurs propres il existe donc p vecteurs $\mathbf{a}_1, \dots, \mathbf{a}_p$ tels que

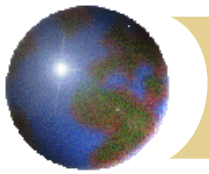
$$\mathbf{VMa}_i = \lambda_i \mathbf{a}_i, \quad \text{avec } \langle \mathbf{a}_i, \mathbf{a}_j \rangle_{\mathbf{M}} = 1 \text{ si } i = j, 0 \text{ sinon.}$$

Les $\mathbf{a}_i$ sont les *axes principaux d'inertie* de $\mathbf{VM}$. Ils sont $\mathbf{M}$ -orthonormaux.

Signe des valeurs propres les valeurs propres de $\mathbf{VM}$ sont positives et on peut les classer par ordre décroissant

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_p \geq 0.$$

Idée du lien avec l'inertie on sait que $\text{Tr}(\mathbf{VM}) = \lambda_1 + \dots + \lambda_p$. Si on ne garde que les données relatives à $\mathbf{a}_1, \dots, \mathbf{a}_k$, on gardera l'inertie $\lambda_1 + \dots + \lambda_k$, et c'est le mieux qu'on puisse faire.



Résultat principal

Théorème principal (Admis)

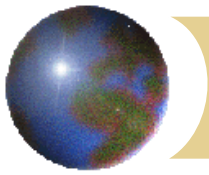
1. Si F_k est le sous-espace de dimension k portant l'inertie principale, alors

$$F_{k+1} = F_k \oplus f_{k+1},$$

où f_{k+1} est le sous espace de dimension 1 $\mathbf{M}$ -orthogonal à F_k portant l'inertie maximale : les solutions sont « emboîtées » ;

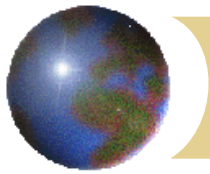
2. F_k est engendré par les k vecteurs propres de $\mathbf{VM}$ associés aux k plus grandes valeurs propres.

Interprétation du théorème l'ACP sur $k + 1$ variables est obtenue par ajout d'une variable d'inertie maximale à l'ACP sur k variables. Il n'est pas nécessaire de refaire tout le calcul.



Détermination des composantes principales

- ✚ **Détermination de la 1^{ère} composante principale:** La solution est donc de projeter les données sur le vecteur propre de la matrice centrée des variances –covariances **S** ayant la valeur propre λ la plus élevée
- ✚ **Détermination de la 2^{ème} composante principale:** Comme les vecteurs propres de **S** sont orthonormés, la solution est de choisir le deuxième vecteur propre de **S** (associé à la deuxième valeur propre maximale)
- ✚ **Et ainsi de suite pour les composantes successives ...**



Interprétation des composantes principales

- $k^{\text{ème}}$ composante principale C_k définie par la combinaison linéaire:

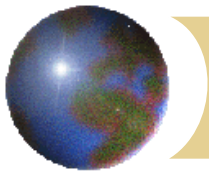
$$C_k = v_{1k} Y_1 + v_{2k} Y_2 + \dots + v_{pk} Y_p \quad (Y_j = \text{variables centrées})$$

où $(v_{1k}, v_{2k}, \dots, v_{pk})^T = \mathbf{v}_k$ le $k^{\text{ème}}$ vecteur propre de $\mathbf{S}$

avec $\lambda_1 \geq \lambda_2 \geq \dots \lambda_k \geq \dots \lambda_p$

et v_{jk} = contribution de la variable Y_j (X_j) à la composante C_k

- Tout vecteur de données initial $\mathbf{x}_i$ devient $\mathbf{c}_i$ où $c_{ik} = \sum_{j=1}^p (x_{ij} - \bar{x}_j) v_{jk}$
- Moyenne de $C_k = 0$
- Variance de $C_k = \lambda_k$, $\text{cov}(C_k, C_{k'}) = 0$ (pour tout $k \neq k'$),
=> matrice var.-covar. des C_k = matrice diagonale $\mathbf{\Lambda}$ des valeurs propres de $\mathbf{S}$
- $\text{Cov}(C_k, X_j) = \lambda_k v_{jk} \Rightarrow \text{Cor}(C_k, X_j) = \frac{\lambda_k v_{jk}}{\sqrt{\lambda_k s_{jj}}} = \sqrt{\frac{\lambda_k}{s_{jj}}} v_{jk}$



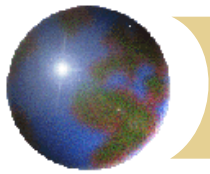
Interprétation des valeurs propres

- La somme des valeurs propres correspond à la variance totale des données:

$$\text{tr}(\mathbf{S}) = \sum_{j=1}^p s_j^2 = \sum_{j=1}^p \lambda_j$$

- Chaque valeur propre mesure la part de variance expliquée par l'axe factoriel (composante principale) correspondant(e) = variance des projections sur l'axe

- % de variance expliquée par chaque composante: $\frac{\lambda_k}{\sum_{j=1}^p \lambda_j}$



Résultats

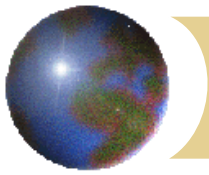
Par transformation linéaire, l'ACP remplace les p variables de départ (corrélées et de variances $\neq$) en q nouvelles composantes ($q \leq p$) C_k

- orthogonales 2 à 2 (décorrélées)
- de variances maximales et décroissantes: $\text{var}(C_1) \geq \text{var}(C_2) \dots \geq \text{var}(C_q)$
- nombre maximum de composantes principales: q (nbre de $\lambda_k \neq 0$) $\leq p$
avec $q < p$ dès que l'une des variables d'origine est une combinaison linéaire d'autres.

$\Rightarrow$ mise en évidence de relations linéaires dans les données

$\Rightarrow$ les données occupent, en réalité, un sous-espace de dimensions réduites ($q < p$)

$\Rightarrow$ Le nombre maximum de composantes principales = dimension intrinsèque des données



Choix des r premières composantes principales

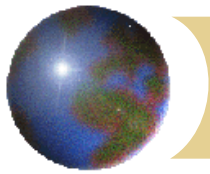
$r \ll p \longrightarrow$ réduction de la dimension

objectif : garder un maximum d'information des données initiales.

Mesure de cette information : le % de variance expliquée:

$$\frac{\sum_{k=1}^r \lambda_k}{\sum_{k=1}^p \lambda_k} = \frac{\sum_{k=1}^r V(C_k)}{\text{Variance totale}}$$

Si les variables originales sont fortement corrélées entre elles, un nombre réduit de composantes permet d'expliquer 80% à 90% de variance !



Interprétation géométrique

Projection des données dans un sous-espace de dimension r , centré sur $\mathbf{g}$ et défini par les r premiers axes principaux d'allongement du nuage.

La projection de $\mathbf{x}_i$ dans ce nouvel espace à r dimensions est un vecteur $\mathbf{c}_i$:

$$\mathbf{c}_i = \sum_{k=1}^r c_{ik} \mathbf{v}_k$$

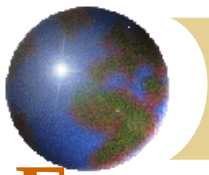
où $\mathbf{v}_k = (v_{k1}, \dots, v_{kp})^T$ est le $k^{\text{ème}}$ vecteur propre de $\mathbf{S}$ ⎧ matrice var.-covar. ⎫
des données initiales

$$\text{et } c_{ik} = \sum_{j=1}^p v_{kj} (x_{ij} - \bar{x}_j) = \sum_{j=1}^p v_{kj} y_{ij}$$

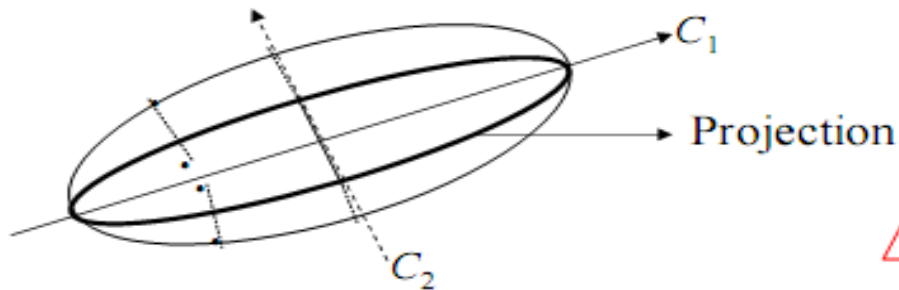
Les projections $\mathbf{c}_i$ sont les plus dispersées possible et minimisent les distorsions au sens des moindres carrés:

$$J_r = \sum_{i=1}^n \|\mathbf{c}_i - \mathbf{y}_i\|^2 \quad (\mathbf{y}_i = \mathbf{x}_i - \mathbf{g})$$

$\Rightarrow$ Les $\mathbf{c}_i$ sont des approximations des $\mathbf{y}_i$ minimisant l'erreur des moindres carrés.



Exemple : données initiales à 3 dimensions distribuées dans un «ballon de rugby»

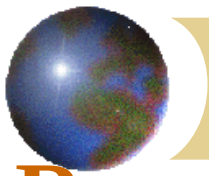


! proximité sur le plan C_1, C_2 $\neq$ proximité dans l'espace initial

Plus le nuage est aplati sur $C_1, C_2 \Rightarrow$ moins de variance sur la 3^{ème} dimension.
 $\Rightarrow$ % de variance expliquée par C_1, C_2 élevé

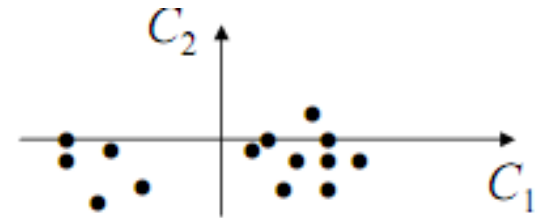
En général :

- % de variance expliquée par $C_1, C_2, \dots, C_r$ = mesure d'aplatissement du nuage sur le sous-espace des composantes (à r dim.). Plus ce % est grand, meilleure est la représentation des données dans le sous-espace !
- Les composantes principales sont entièrement déterminées par la matrice S variance-covariance (vecteurs propres).
 $\Rightarrow$ modification de $S \longrightarrow$ modification des composantes



Représentation des individus dans le plan principal

=> peut faire apparaître des groupes d'individus présentant des similitudes.



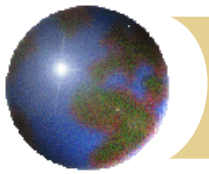
⚠ proximités abusives dues aux projections

=> la représentation n'est valable que si le % de variance expliquée par C_1 et C_2 est suffisamment élevé (nuage assez aplati sur le plan)

=> vérifier si les proximités se maintiennent dans d'autres plans de projection: $C_1 - C_3$, $C_2 - C_3$, ...

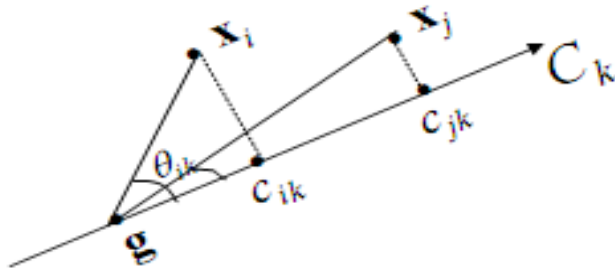
=> individus les mieux représentés = points proches du plan de projection (distorsion peu importante)

=> évaluation de la qualité de représentation des différents points



Mesure de la qualité de représentation d'un point

Pour chaque composante:



$$\cos^2 \theta_{ik} = \frac{c_{ik}^2}{d^2(\mathbf{x}_i, \mathbf{g})} = \frac{\left[\sum_{j=1}^p (x_{ij} - \bar{x}_j) v_{jk} \right]^2}{\left[\sum_{k=1}^p \sum_{j=1}^p (x_{ij} - \bar{x}_j) v_{jk} \right]^2}$$

$$\text{car } \sum_k \cos^2 \theta_{ik} = 1$$

bonne représentation si $\theta_{ik} \approx 0^\circ$ ou 180° ($\cos^2 \theta_{ik} \approx 1$)

mauvaise représentation si $\theta_{ik} \approx 90^\circ$ ou 270° ($\cos^2 \theta_{ik} \approx 0$)

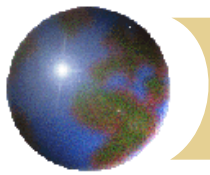
$\Rightarrow \mathbf{x}_j$ mieux représenté que $\mathbf{x}_i$ sur C_k

Qualité de la représentation *sur le plan principal*: $\cos^2 \theta_{i1} + \cos^2 \theta_{i2} (\leq 1)$

Contribution de chaque point à l'inertie (variance) de C_k :

$$\text{Var}(C_k) = \lambda_k = \sum_{i=1}^n c_{ik}^2 \Rightarrow \text{contribution de } \mathbf{x}_i = \frac{c_{ik}^2}{\lambda_k}$$

$\Rightarrow$ identification des éléments responsables de la position de l'axe



Interprétation des composantes principales

Base: corrélations avec les variables initiales

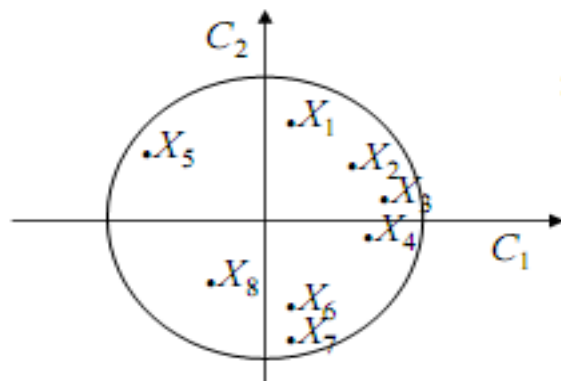
=> repérer les variables très corrélées
($r \approx 1$ ou $r \approx -1$)

	C_1	C_2	C_3	...
X_1	r_{11}	r_{12}	r_{13}	...
X_2	r_{21}	r_{22}	r_{23}	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...
X_p	r_{p1}	r_{p2}	r_{p3}	...

Interprétation des 2 premières composantes C_1 , C_2 : cercle des corrélations :

C_1 et C_2 étant non-corrélées, on a $Cor^2(C_1, X_j) + Cor^2(C_2, X_j) \leq 1$

=> chaque variable représentée par les coordonnées : $(Cor(C_1, X_j), Cor(C_2, X_j))$ est dans un cercle de rayon 1



=> groupes de variables liées ou opposées



si proches de la circonférence, bien représentées par les 2 composantes !