# Variables quantitatives : analyse en composantes principales

*Jean-Marc Lasgouttes*

https://who.rocq.inria.fr/Jean-Marc.Lasgouttes/ana-donnees/

### Préambule : 3 approches des données

**Décrire** les données de 3 manières complémentaires
- *Variables* : chaque colonne représente une variable qui se prête à des calculs statistiques
- *Matrice* : le tableau complet de données est une matrice de nombres réels
- *Nuage de points* : chaque ligne du tableau représente les coordonnées d'un point dans un espace dont la dimension est le nombre de variables

**Combiner** ces trois approches pour définir l'ACP en termes de
- moyenne, variance, corrélation
- valeurs propres, vecteurs propres
- distances, angles, projection

**Conséquences** sur le cours
- les trois premières parties sont des préliminaires qui durent la moitié du cours !
- il faut faire attention pour comprendre le rôle des différentes approches

# Partie I. Données : vision statistique

### Les données quantitatives

**Définition** On appelle « variable » un vecteur $\mathbf{x}$ de taille $n$. Chaque coordonnée $x_i$ correspond à un individu. On s'intéresse ici à des valeurs numériques.

**Poids** Chaque individu peut avoir un poids $p_i$, tel que $p_1 + \cdots + p_n = 1$, notamment quand les individus n'ont pas la même importance (échantillons redressés, données regroupées,...). On a souvent $p = 1/n$.

**Résumés** on dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, étendue, 1er quartile (25% inférieurs), 4ème quartile (25% supérieurs), ... Ces indicateurs mesurent principalement la tendance centrale et la dispersion.

On utilisera principalement la moyenne, la variance et l'écart type.

### Moyenne arithmétique

**Définition** On note

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left[x_1 + x_2 + \cdots + x_n\right],$$

ou pour des données pondérés

$$\bar{x} = \sum_{i=1}^{n} p_i x_i = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n.$$

**Propriétés** la moyenne arithmétique est une mesure de *tendance centrale* qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée à cause de ses bonnes propriétés mathématiques.

### Variance et écart-type

**Définition** la *variance* de $\mathbf{x}$ est définie par

$$\operatorname{var}(\mathbf{x}) = \sigma_{\mathbf{x}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ ou } \operatorname{var}(\mathbf{x}) = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$

L'*écart-type* $\sigma_{\mathbf{x}}$ est la racine carrée de la variance.

**Propriétés** La variance satisfait la formule suivante

$$\operatorname{var}(\mathbf{x}) = \sum_{i=1}^{n} p_i x_i^2 - (\bar{x})^2$$

La variance est « la moyenne des carrés moins le carré de la moyenne ». L'écart-type, qui a la même unité que $\mathbf{x}$, est une mesure de *dispersion*.

**Attention !** les calculatrices utilisent l'estimateur sans biais de la variance dans lequel le $1/n$ est remplacé par $1/(n-1)$.

### Mesure de liaison entre deux variables

**Définitions** la covariance observée entre deux variables $\mathbf{x}$ et $\mathbf{y}$ est

$$\operatorname{cov}(\mathbf{x},\mathbf{y}) = \sigma_{\mathbf{xy}} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{x}\bar{y}.$$

et le *coefficient de r de Bravais-Pearson* ou coefficient de corrélation est donné par

$$\operatorname{cor}(\mathbf{x},\mathbf{y}) = r_{\mathbf{xy}} = \frac{\sigma_{\mathbf{xy}}}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} = \frac{\operatorname{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\operatorname{var}(\mathbf{x})}\sqrt{\operatorname{var}(\mathbf{y})}}.$$

**Propriétés**

- $\text{cov}(\mathbf{x},\mathbf{x}) = \text{var}(\mathbf{x})$ et $\text{cor}(\mathbf{x},\mathbf{x}) = 1$
- $\text{cov}(\mathbf{x},\mathbf{y}) = \text{cov}(\mathbf{y},\mathbf{x})$ et donc $\text{cor}(\mathbf{x},\mathbf{y}) = \text{cor}(\mathbf{y},\mathbf{x})$.

## Propriétés du coefficient de corrélation

**Borne** On a toujours (inégalité de Cauchy-Schwarz)

$$-1 \leq \text{cor}(\mathbf{x},\mathbf{y}) \leq 1.$$

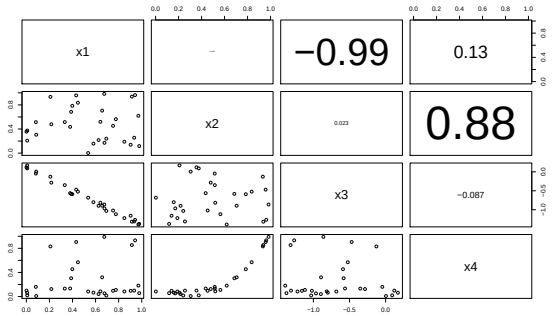
**Variables liées** $|\text{cor}(\mathbf{x},\mathbf{y})| = 1$ si et seulement si $\mathbf{x}$ et $\mathbf{y}$ sont linéairement liées :

$$ax_i + by_i = c, \text{ pour tout } 1 \leq i \leq n.$$

En particulier, $\text{cor}(\mathbf{x},\mathbf{x}) = 1$.

**Variables décorrélées** si $\text{cor}(\mathbf{x},\mathbf{y}) = 0$, on dit que les variables sont *décorrélées*. Cela ne veut pas dire qu'elles sont indépendantes !

## Le coefficient de corrélation par l'exemple




**Interprétation** on a 4 variables numériques avec 30 individus. Les variables 1 et 2 sont « indépendantes » ; les variables 1 et 3 ont une relation linéaire ; les variables 2 et 4 ont une relation non-linéaire.

## Que signifie une corrélation linéaire ?

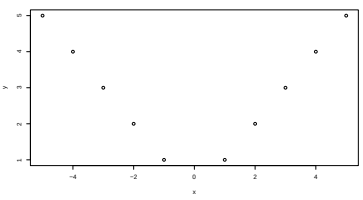
**Qu'est ce qui est significatif ?** si on a assez de données, on peut considérer qu'une corrélation supérieure à $0,5$ est forte, et une corrélation entre $0,3$ et $0,5$ est moyenne.

Une corrélation égale à un indique que les deux variables sont équivalentes.

**Qu'est-ce que cela veut dire ?** une corrélation significative indique une liaison entre deux variables, mais pas nécessairement un lien de causalité. Exemple :

> Le nombre de pompiers présents pour combattre un incendie est corrélé aux dégâts de l'incendie. *Mais ce ne sont pas les pompiers qui causent les dégâts.*

**Et une décorrélation ?** voici un exemple ou $\text{cor}(\mathbf{x},\mathbf{y}) = 0$



# Partie II. Données : vision matricielle

## Pense-bête matrices (1/2)

**Matrice** tableau de données, noté par un lettre majuscule grasse (ex : $\mathbf{A}$).

**Vecteur** matrice à une seule colonne, noté par une lettre minuscule grasse (ex : $\mathbf{x}$).

**Cas particuliers** matrices zéro ($n \times p$), identité ($n \times n$) et vecteur unité de taille $n$ :

$$\mathbf{0}_{np} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \\ 0 & & 0 \end{bmatrix}, \quad \mathbf{I}_n = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}, \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

**Addition** Possible quand les dimensions sont égales ; on ajoute les coefficients.

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}, \qquad \mathbf{A} + \mathbf{0} = \mathbf{A}$$

**Produit** Contrainte lignes/colonnes : $\underset{(n\times p)}{\mathbf{A}} \times \underset{(p\times k)}{\mathbf{B}} \Longrightarrow \underset{(n\times k)}{\mathbf{C}}$

Nombre de colonnes de la première matrice égal au nombre de lignes de la seconde

$$\mathbf{AB} \neq \mathbf{BA}, \qquad (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$$
$$\mathbf{I}_n\mathbf{A} = \mathbf{A}\mathbf{I}_p = \mathbf{A} \qquad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$

## Pense-bête matrices (2/2)

**Transposition** échange des lignes et des colonnes d'une matrice ; on note $\mathbf{A}'$ la transposée de $\mathbf{A}$.

$$(\mathbf{A}')' = \mathbf{A}, \qquad (\mathbf{A}+\mathbf{B})' = \mathbf{A}' + \mathbf{B}', \qquad (\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$

**Trace** la trace d'une matrice carrée est la somme des termes de sa diagonale

$$\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA}),$$
$$\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{CAB}) = \text{Tr}(\mathbf{BCA}) \neq \text{Tr}(\mathbf{CBA})$$

**Inverse** si $\mathbf{A}$ et $\mathbf{B}$ sont carrées de taille $n$, alors

$\mathbf{AB} = \mathbf{I}_n \implies \mathbf{BA} = \mathbf{I}_n$ On note $\mathbf{B} = \mathbf{A}^{-1}$ (inverse de $\mathbf{A}$)

## Tableau de données

On note $x_i^j$ la valeur de la *variable* $\mathbf{x}^j$ pour le $i$-ème individu. $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^p)$ est une matrice rectangulaire à $n$ lignes et $p$ colonnes.

$$\mathbf{x}^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \\ \\ \vdots \\ x_n^j \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_1^1 & x_1^2 & & & \cdots & x_1^p \\ x_2^1 & x_2^2 & & & & \\ & & & \vdots & & \\ & & \cdots & x_i^j & & \\ \vdots & & & & \ddots & \\ x_n^1 & & & & & x_n^p \end{bmatrix}.$$

Un *individu* est représenté par

$$\mathbf{e}_i' = [x_i^1, \ldots, x_i^j, \ldots, x_i^p]$$

## La matrice des poids

**Définition** on associe aux individus un poids $p_i$ tel que

$$p_1 + \cdots + p_n = 1$$

que l'on représente par la matrice diagonale de taille $n$

$$\mathbf{D_p} = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{bmatrix}.$$

**Cas uniforme** tous les individus ont le même poids $p_i = 1/n$ et $\mathbf{D_p} = \frac{1}{n}\mathbf{I}_n$.

## Point moyen et tableau centré

**Point moyen** c'est le vecteur $\mathbf{g}$ des moyennes arithmétiques de chaque variable :

$$\mathbf{g}' = (\bar{x}^1, \ldots, \bar{x}^p), \text{ où } \bar{x}^j = \sum_{i=1}^{n} p_i x_i^j.$$

On peut écrire sous forme matricielle

$$\mathbf{g} = \mathbf{X}'\mathbf{D_p}\mathbf{1}_n.$$

**Tableau centré** il est obtenu en centrant les variables autour de leur moyenne

$$y_i^j = x_i^j - \bar{x}^j$$

ou, en notation matricielle,

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}_n\mathbf{g}' = (\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'\mathbf{D_p})\mathbf{X}$$

## Matrice de variance-covariance

**Définition** c'est une matrice *carrée* de dimension $p$

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & & & \\ \vdots & & \ddots & \\ \sigma_{p1} & & & \sigma_p^2 \end{bmatrix},$$

où $\sigma_{kl}$ est la covariance des variables $\mathbf{x}^k$ et $\mathbf{x}^\ell$ et $\sigma_j^2$ est la variance de la variable $\mathbf{x}^j$

**Formule matricielle**

$$\mathbf{V} = \mathbf{X}'\mathbf{D_p}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D_p}\mathbf{Y}.$$

## Matrice de corrélation

**Définition** Si l'on note $r_{k\ell} = \sigma_{k\ell}/\sigma_k\sigma_\ell$, c'est la matrice $p \times p$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & & \\ \vdots & & \ddots & \\ r_{p1} & & & 1 \end{bmatrix},$$

**Formule matricielle** $\mathbf{R} = \mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{V}\mathbf{D}_{1/\boldsymbol{\sigma}}$, où

$$\mathbf{D}_{1/\boldsymbol{\sigma}} = \begin{bmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_p} \end{bmatrix}$$

# Partie III. Données : vision géométrique

## L'analyse de composantes principales (ACP)

**Contexte** chaque individu est considéré comme un point d'un espace vectoriel $F$ de dimension $p$. Ses coordonnées dans $F$ sont

$$(x_i^1, x_i^2, \ldots, x_i^p).$$

L'ensemble des individus est un *nuage de points* dans $F$ et $\mathbf{g}$ est son *centre de gravité*.

**Principe** on cherche à réduire le nombre $p$ de variables tout en préservant au maximum la structure du problème.

**Pour cela on projette le nuage de points sur un sous-espace de dimension inférieure.**

## Distance entre individus

**Motivation** afin de pouvoir considérer la structure du nuage des individus, il faut définir une distance, qui induira une géométrie.

**Distance euclidienne classique** la distance la plus simple entre deux points de $\mathbb{R}^p$ est définie par

$$d^2(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^{p} (u_j - v_j)^2 = \|\mathbf{u} - \mathbf{v}\|^2$$

**Généralisation simple** on donne un poids $m_j > 0$ à la variable $j$

$$d^2(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^{p} m_j (u_j - v_j)^2$$

Cela revient à multiplier la coordonnée $j$ par $\sqrt{m_j}$

## Métrique

**Définition** soit $\mathbf{M} = \mathrm{diag}(m_j)$, où $m_1, \ldots, m_p$ sont des réels strictement positifs. On pose

$$\|\mathbf{u}\|_{\mathbf{M}}^2 = \mathbf{u}'\mathbf{M}\mathbf{u} = \sum_{j=1}^{p} m_j u_j^2,$$

$$d_{\mathbf{M}}^2(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_{\mathbf{M}}^2.$$

**Espace métrique** il est défini par le produit scalaire

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} = \mathbf{u}'\mathbf{M}\mathbf{v} = \sum_{j=1}^{p} m_j u_j v_j.$$

On notera que $\|\mathbf{u}\|_{\mathbf{M}}^2 = \langle \mathbf{u}, \mathbf{u} \rangle_{\mathbf{M}}$.

**Orthogonalité** on dit que $\mathbf{u}$ et $\mathbf{v}$ sont $\mathbf{M}$-orthogonaux si $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} = 0$.

## Propriétés du produit scalaire

**Le produit scalaire est commutatif**

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} = \langle \mathbf{v}, \mathbf{u} \rangle_{\mathbf{M}}$$

**Le produit scalaire est linéaire**

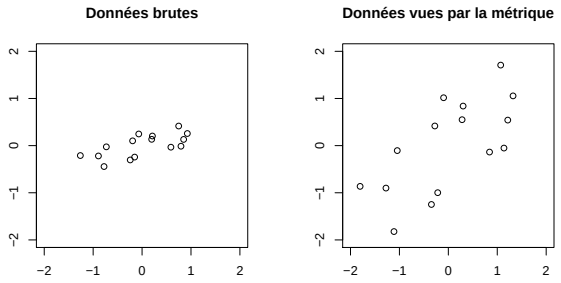
$$\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle_{\mathbf{M}} = \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} + \langle \mathbf{u}, \mathbf{w} \rangle_{\mathbf{M}},$$
$$\langle \mathbf{u}, \lambda \mathbf{v} \rangle_{\mathbf{M}} = \lambda \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} \quad \text{pour tout } \lambda \in \mathbb{R}.$$

**Identité remarquable**

$$\|\mathbf{u} + \mathbf{v}\|^2_{\mathbf{M}} = \|\mathbf{u}\|^2_{\mathbf{M}} + \|\mathbf{v}\|^2_{\mathbf{M}} + 2\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}}$$

## Utilisation des métriques

Utiliser une métrique est donc équivalent à « tordre » les données, par exemple pour les rendre comparables




## Le cas de la métrique $\mathbf{D}_{1/\sigma^2}$

**Données centrées réduites** c'est le tableau $\mathbf{Z}$ contenant les données

$$z_i^j = \frac{y_i^j}{\sigma_j} = \frac{x_i^j - \bar{x}^j}{\sigma_j},$$

qui se calcule matriciellement comme $\mathbf{Z} = \mathbf{Y}\mathbf{D}_{1/\sigma}$.

**Pourquoi réduites ?**

- pour que les distances soient indépendantes des unités de mesure
- pour qu'elles ne privilégient pas les variables dispersées.

**Équivalence avec une métrique** diviser les variables par $\sigma_j$ est équivalent à prendre $m_j = \sigma_j^2$. On a $\mathbf{D}_{1/\sigma^2} = \mathbf{D}_{1/\sigma}\mathbf{D}_{1/\sigma}$ et donc

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{D}_{1/\sigma^2}} = \mathbf{u}'\mathbf{D}_{1/\sigma^2}\mathbf{u} = \mathbf{u}'\mathbf{D}_{1/\sigma}\mathbf{D}_{1/\sigma}\mathbf{u} = \langle \mathbf{D}_{1/\sigma}\mathbf{u}, \mathbf{D}_{1/\sigma}\mathbf{v} \rangle.$$

Travailler avec la métrique $\mathbf{D}_{1/\sigma^2}$ est équivalent à diviser chaque variable par son écart-type et à utiliser la métrique $\mathbf{I}$.

## Inertie

**Définition** l'inertie en un point $\mathbf{v}$ du nuage de points est

$$I_{\mathbf{v}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{v}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{v})'\mathbf{M}(\mathbf{e}_i - \mathbf{v}).$$

**Inertie totale** La plus petite inertie possible est $I_{\mathbf{g}}$, donnée par

$$I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{g}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{g})'\mathbf{M}(\mathbf{e}_i - \mathbf{g})$$

qui est la seule intéressante puisque $I_{\mathbf{v}} = I_{\mathbf{g}} + \|\mathbf{v} - \mathbf{g}\|^2_M$.

**Autres relations** $I_{\mathbf{g}}$ mesure la moyenne des carrés des distances entre les individus

$$2I_{\mathbf{g}} = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \|\mathbf{e}_i - \mathbf{e}_j\|^2_{\mathbf{M}}.$$

**Interprétation** L'inertie totale mesure l'étalement du nuage de points

## Métriques particulières

**Forme matricielle** L'inertie totale est aussi donnée par la trace de la matrice $\mathbf{VM}$ (ou $\mathbf{MV}$)

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{VM}) = \mathrm{Tr}(\mathbf{MV})$$

**Métrique usuelle** $\mathbf{M} = \mathbf{I}_p$ correspond au produit scalaire usuel et

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{V}) = \sum_{j=1}^{p} \sigma_i^2$$

**Métrique réduite** obtenue quand $\mathbf{M} = \mathbf{D}_{1/\sigma^2} = \mathbf{D}^2_{1/\sigma}$

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{D}_{1/\sigma^2}\mathbf{V}) = \mathrm{Tr}(\mathbf{D}_{1/\sigma}\mathbf{V}\mathbf{D}_{1/\sigma}) = \mathrm{Tr}(\mathbf{R}) = p.$$

## L'analyse de composantes principales (version 2)

**Principe** on cherche à projeter orthogonalement le nuage de points sur un espace $F_k$ de dimension $k < p$, sous la forme

$$\mathbf{e}_i^* - \mathbf{g} = c_{i1}\mathbf{a}_1 + c_{i2}\mathbf{a}_2 + \cdots + c_{ik}\mathbf{a}_k$$

Les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ définissent l'espace $F_k$ et les $c_{i\ell}$ sont les coordonnées de $\mathbf{e}_i^*$.

**Critère** on veut que la moyenne des carrés des distances entre les points $\mathbf{e}_i$ et leur projetés $\mathbf{e}_i^*$ soit minimale. Comme on a toujours (théorème de Pythagore)

$$\|\mathbf{e}_i - \mathbf{g}\|^2 = \|\mathbf{e}_i - \mathbf{e}_i^*\|^2 + \|\mathbf{e}_i^* - \mathbf{g}\|^2,$$

cela revient à maximiser l'inertie du nuage projeté.

**On cherche donc $F_k$, sous espace de dimension $k$ de $F_p$, tel que l'inertie du nuage projeté sur $F_k$ soit maximale.**

# Partie IV. Calcul de l'ACP

## Résultat principal (incompréhensible)

**Théorème principal** *Le sous-espace $F_k$ de dimension $k$ portant l'inertie maximale est engendré par les $k$ vecteurs propres de $\mathbf{VM}$ associés aux $k$ plus grandes valeurs propres.*

**Problème** Qu'est-ce qu'une valeur propre ? ? ?

## Matrices : valeurs propres et vecteurs propres

**Définition** un vecteur $\mathbf{v} \neq \mathbf{0}$ 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

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

$\lambda$ 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.

**Non unicité des vecteur propres** Si $\mathbf{v}$ est un vecteur propre de $\mathbf{A}$ associé la valeur propre $\lambda$, alors, pour tout $\alpha \in \mathbb{C}$, $\alpha\mathbf{v}$ est aussi vecteur propre de $\mathbf{A}$ :

$$\mathbf{A}(\alpha\mathbf{v}) = \alpha\mathbf{A}\mathbf{v} = \alpha\lambda\mathbf{v} = \lambda(\alpha\mathbf{v}).$$

## Valeurs et vecteurs propres : un exemple concret

La matrice

$$\begin{pmatrix} 5 & 1 & -1 \\ 2 & 4 & -2 \\ 1 & -1 & 3 \end{pmatrix}$$

a pour vecteurs propres

$$\mathbf{v}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \ \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \ \mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

On vérifie facilement que les valeurs propres associées sont

$$\lambda_1 = 2, \ \lambda_2 = 4, \ \lambda_3 = 6.$$

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

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 $\mathbf{A}$ de taille $p \times p$ qui a $p$ valeurs propres distinctes est diagonalisable et

$$\mathrm{Tr}(\mathbf{A}) = \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$

## Quelques matrices diagonalisables

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

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

**Matrice 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 réel et ses valeurs propres sont elles aussi réelles

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

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

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

## Analyse de VM

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

**Axes principaux d'inertie** ce sont les $p$ vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_p$ tels que

$$\mathbf{VM}\mathbf{a}_k = \lambda_k \mathbf{a}_k, \qquad \text{avec } \langle \mathbf{a}_k, \mathbf{a}_\ell \rangle_{\mathbf{M}} = 1 \text{ si } k = \ell, \ 0 \text{ sinon.}$$

Ils sont $\mathbf{M}$-orthonormaux.

**Signe des valeurs propres** les valeurs propres de $\mathbf{VM}$ sont positives (preuve plus tard) et on peut les classer par ordre décroissant

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

## Résultat principal (admis)

**Théorème principal** *Le sous-espace $F_k$ de dimension $k$ portant l'inertie maximale est engendré par les $k$ vecteurs propres de $\mathbf{VM}$ associés aux $k$ plus grandes valeurs propres.*

**Interprétation du théorème**

- l'ACP sur $k$ variables est obtenue en se limitant aux $k$ plus grandes valeurs propres.
- Le calcul ne dépend pas du nombre de variables retenues.

**Idée du lien avec l'inertie** on sait que

$$I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{VM}) = \lambda_1 + \cdots + \lambda_p.$$

Si on ne garde que les données relatives à $\mathbf{a}_1, \ldots, \mathbf{a}_k$, on gardera l'inertie $\lambda_1 + \cdots + \lambda_k$, et c'est le mieux qu'on puisse faire.

# Partie V. Les éléments de l'ACP

## Les composantes principales

**Coordonnées des individus** supposons que $\mathbf{e}_i - \mathbf{g} = \sum_{\ell=1}^{p} c_{i\ell}\mathbf{a}_\ell$, alors

$$\langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k\rangle_{\mathbf{M}} = \sum_{\ell=1}^{p} c_{i\ell}\langle \mathbf{a}_\ell, \mathbf{a}_k\rangle_{\mathbf{M}} = c_{ik}$$

La coordonnée de l'individu centré $\mathbf{e}_i - \mathbf{g}$ sur l'axe principal $\mathbf{a}_k$ est donc donné par la projection $\mathbf{M}$-orthogonale

$$c_{ik} = \langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k\rangle_{\mathbf{M}} = (\mathbf{e}_i - \mathbf{g})'\mathbf{M}\mathbf{a}_k.$$

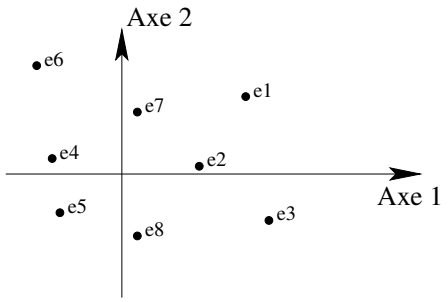
**Composantes principales** ce sont les variables $\mathbf{c}_k = (c_{1k}, \dots, c_{nk})$ de taille $n$ définies par

$$\mathbf{c}_k = \mathbf{Y}\mathbf{M}\mathbf{a}_k.$$

Chaque $\mathbf{c}_k$ contient les coordonnées des projections $\mathbf{M}$-orthogonales des individus centrés sur l'axe défini par les $\mathbf{a}_k$.

## Représentation des individus dans un plan principal

**Qu'est-ce que c'est ?** pour deux composantes principales $\mathbf{c}_1$ et $\mathbf{c}_2$, on représente chaque individu $i$ par un point d'abscisse $c_{i1}$ et d'ordonnée $c_{i2}$.




**Quand ?** Elle est utile pour des individus discernables.

## Propriétés des composantes principales

**Moyenne arithmétique** les composantes principales sont centrées :

$$\bar{c}_k = \mathbf{c}_k'\mathbf{D_p}\mathbf{1}_n = \mathbf{a}_k'\mathbf{M}\mathbf{Y}'\mathbf{D_p}\mathbf{1}_n = 0$$

car $\mathbf{Y}'\mathbf{D_p}\mathbf{1}_n = \mathbf{0}$ (les colonnes de $\mathbf{Y}$ sont centrées).

**Variance** la variance de $\mathbf{c}_k$ est $\lambda_k$ car

$$\begin{aligned}\operatorname{var}(\mathbf{c}_k) &= \mathbf{c}_k'\mathbf{D_p}\mathbf{c}_k = \mathbf{a}_k'\mathbf{M}\mathbf{Y}'\mathbf{D_p}\mathbf{Y}\mathbf{M}\mathbf{a}_k \\ &= \mathbf{a}_k'\mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}_k = \lambda_k\mathbf{a}_k'\mathbf{M}\mathbf{a}_k = \lambda_k.\end{aligned}$$

*Par conséquent on a toujours* $\lambda_k \geq 0$

**Covariance** de même, pour $k \neq \ell$,

$$\operatorname{cov}(\mathbf{c}_k, \mathbf{c}_\ell) = \mathbf{c}_k'\mathbf{D_p}\mathbf{c}_\ell = \cdots = \lambda_\ell\mathbf{a}_k'\mathbf{M}\mathbf{a}_\ell = 0.$$

*Les composantes principales ne sont pas corrélées entre elles.*

## Interprétation dans l'espace des variables

On peut transposer le tableau de données et étudier un nuage de $p$ points de $\mathbb{R}^n$ où chaque point est une variable.

**Métrique $\mathbf{D_p}$** il faut munir l'espace des variables d'une métrique raisonnable. On choisit toujours la métrique $\mathbf{D_p}$ des poids :

$$\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{D_p}} = \mathbf{x}'\mathbf{D_p}\mathbf{y}, \qquad \|\mathbf{x}\|^2_{\mathbf{D_p}} = \mathbf{x}'\mathbf{D_p}\mathbf{x}.$$

**Covariance et produit scalaire** pour deux variables *centrées* $\mathbf{x}$ et $\mathbf{y}$, on a

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{D_p}}, \qquad \operatorname{var}(\mathbf{x}) = \|\mathbf{x}\|^2_{\mathbf{D_p}},$$
$$\operatorname{cor}(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{D_p}}}{\|\mathbf{x}\|_{\mathbf{D_p}}\|\mathbf{y}\|_{\mathbf{D_p}}} = \cos(\widehat{\mathbf{x}\mathbf{y}}).$$

**Exemple** les vecteurs $\mathbf{c}_k/\sqrt{\lambda_k}$ forment une base $\mathbf{D_p}$-orthonormale

$$\left\langle \frac{\mathbf{c}_k}{\sqrt{\lambda_k}}, \frac{\mathbf{c}_\ell}{\sqrt{\lambda_\ell}} \right\rangle_{\mathbf{D_p}} = \operatorname{cor}(\mathbf{c}_k, \mathbf{c}_\ell) = \begin{cases} 1, & \text{si } k = \ell, \\ 0, & \text{sinon.} \end{cases}$$

## Facteurs principaux

**Définition** on associe à $\mathbf{a}_k$ le facteur principal $\mathbf{u}_k = \mathbf{M}\mathbf{a}_k$ de taille $p$. C'est un vecteur propre de $\mathbf{MV}$ car

$$\mathbf{M}\mathbf{V}\mathbf{u}_k = \mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}_k = \lambda_k\mathbf{M}\mathbf{a}_k = \lambda_k\mathbf{u}_k$$

**Calcul** en pratique, on calcule les $\mathbf{u}_k$ par diagonalisation de $\mathbf{MV}$, puis on obtient les $\mathbf{c}_k = \mathbf{Y}\mathbf{u}_k$. Les $\mathbf{a}_k$ ne sont pas intéressants.

**Interprétation** Si on pose $\mathbf{u}_k' = (u_{1k}, \dots, u_{pk})$, on voit que la matrice des $u_{jk}$ sert de matrice de passage entre la nouvelle base et l'ancienne

$$c_{ik} = \sum_{j=1}^{p} y_i^j u_{jk}, \qquad \mathbf{c}_k = \sum_{j=1}^{p} \mathbf{y}^j u_{jk}$$

## Formules de reconstitution

**Reconstitution** Par définition des $\mathbf{c}_k$,on a $\mathbf{e}_i - \mathbf{g} = \sum_{k=1}^{p} c_{ik}\mathbf{a}_k$, et donc

$$y_i^j = \sum_{k=1}^{p} c_{ik}a_{kj}, \qquad \mathbf{y}^j = \sum_{k=1}^{p} \mathbf{c}_k a_{kj}, \qquad \mathbf{Y} = \sum_{k=1}^{p} \mathbf{c}_k\mathbf{a}_k'$$

Les $a_{kj}$ forment de matrice de passage entre l'ancienne base et la nouvelle.

**Approximation** Les $k$ premiers termes fournissent la meilleure approximation de $\mathbf{Y}$ par une matrice de rang $k$ au sens des moindres carrés (théorème de Eckart-Young).

# Partie VI. Aspects pratiques

## L'ACP sur les données centrées réduites

**Matrice de variance-covariance** c'est la matrice de corrélation car

$$\mathbf{Z}'\mathbf{D_p}\mathbf{Z} = \mathbf{D}_{1/\sigma}\mathbf{Y}'\mathbf{D_p}\mathbf{Y}\mathbf{D}_{1/\sigma} = \mathbf{D}_{1/\sigma}\mathbf{V}\mathbf{D}_{1/\sigma} = \mathbf{R}.$$

**Métrique** on prend la métrique $\mathbf{M} = \mathbf{I}_p$.

**Facteurs principaux** Les $\mathbf{u}_k = \mathbf{M}\mathbf{a}_k = \mathbf{a}_k$ sont les $p$ vecteurs propres orthonormés de $\mathbf{R}$,

$$\mathbf{R}\mathbf{u}_k = \lambda_k \mathbf{u}_k, \text{ avec } \langle \mathbf{u}_k, \mathbf{u}_\ell \rangle = 1 \text{ si } k = \ell, \text{ 0 sinon.}$$

Les valeurs propres vérifient

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0 \quad \text{et} \quad \lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_p = p$$

**Composantes principales** elles sont données par $\mathbf{c}_k = \mathbf{Z}\mathbf{u}_k$.

## Nombre d'axes à retenir

**Dimension de l'espace des individus** L'ACP visant à réduire la dimension de l'espace des individus, on veut conserver aussi peu d'axes que possible. Il faut pour cela que les variables d'origine soient raisonnablement corrélées entre elles.

*Les seuls critères utilisables sont empiriques.*

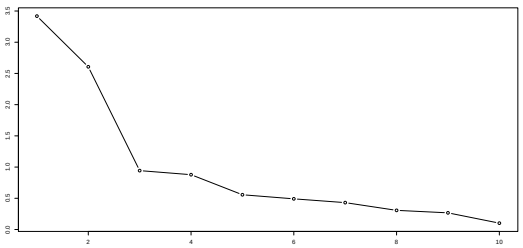
**Interprétation des axes** on s'efforce de ne retenir que des axes à propos desquels une forme d'interprétation est possible (soit directement, soit en terme des variables avec lesquels ils sont très corrélés). On donnera des outils à cet effet plus loin dans le cours.

**Critère de Kaiser (variables centrées-réduites)** on ne retient que les axes associés à des valeurs propres supérieures à 1, c'est-à-dire dont la variance est supérieure à celle des variables d'origine.

Une autre interprétation est que la moyenne des valeurs propres étant 1, on ne garde que celles qui sont supérieures à cette moyenne.

## Nombre d'axes à retenir (suite)

**Éboulis des valeurs propres** on cherche un « coude » dans le graphe des valeurs propres



## Corrélation entre composantes et variables initiales

Sur les variables centrées-réduites, cette corrélation s'écrit

$$\operatorname{cov}(\mathbf{z}^j, \mathbf{c}_k) = \operatorname{cov}\Big(\sum_{\ell=1}^{p} a_{\ell j}\mathbf{c}_\ell, \mathbf{c}_k\Big) = \sum_{\ell=1}^{p} a_{\ell j} \operatorname{cov}(\mathbf{c}_\ell, \mathbf{c}_k) = \lambda_k a_{kj}$$

$$\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_k) = \frac{\operatorname{cov}(\mathbf{z}^j, \mathbf{c}_k)}{\sqrt{\operatorname{var}(\mathbf{c}_k)}} = \frac{\lambda_k a_{kj}}{\sqrt{\lambda_k}} = \sqrt{\lambda_k} u_{jk}$$

**Position dans un plan** On sait que $\operatorname{var}(\mathbf{z}^j) = 1$, mais on peut aussi écrire

$$\operatorname{var}(\mathbf{z}^j) = \operatorname{cov}(\mathbf{z}^j, \mathbf{z}^j) = \operatorname{cov}\Big(\mathbf{z}^j, \sum_{k=1}^{p} a_{kj}\mathbf{c}_k\Big) = \sum_{k=1}^{p} a_{kj} \operatorname{cov}(\mathbf{z}^j, \mathbf{c}_k)$$

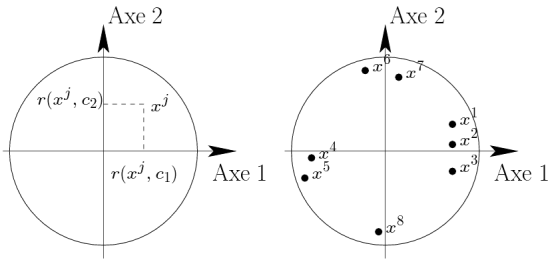
$$= \sum_{k=1}^{p} \lambda_k a_{kj}^2 = \sum_{k=1}^{p} \big[\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_k)\big]^2.$$

Par conséquent, les 2 premières coordonnées sont dans un disque de rayon 1, puisque

$$\big[\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_1)\big]^2 + \big[\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_2)\big]^2 \leq 1$$

## Le cercle des corrélations

**Qu'est-ce que c'est ?** c'est une représentation où, pour deux composantes principales, par exemple $\mathbf{c}_1$ et $\mathbf{c}_2$, on représente chaque variable $\mathbf{z}^j$ par un point d'abscisse $\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_1)$ et d'ordonnée $\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_2)$.




## Le cercle des corrélations (suite)

**Interprétation** Les variables qui déterminent les axes sont celles dont la corrélation est supérieure en valeur absolue à une certaine limite (0,9, 0,8... selon les données) ; on essaie d'utiliser la même limite pour tous les axes.

**Remarques**

- les points sont la projection orthogonale dans $\mathbf{D_p}$ des variables dans le plan défini par les composantes principales $\mathbf{c}_1$ et $\mathbf{c}_2$.
- Il ne faut interpréter la proximité des points que s'ils sont proches de la circonférence.

**Effet « taille »** quand toutes les variables ont le même signe de corrélation avec la première composante principale (positif ou négatif). Cette composante est alors appelée « facteur de taille », la seconde « facteur de forme ».

- un effet de taille indique un consensus sur une variable. Le facteur correspondant ne nous apprend pas toujours quelque chose.
- il n'y a effet de taille que sur le premier axe !
- il n'y a pas d'« effet de forme » !

### Contribution d'un individu à une composante

**Définition** On sait que $\text{var}(\mathbf{c}_k) = \lambda_k = \sum_{i=1}^n p_i c_{ik}^2$. La contribution de l'individu $i$ à la composante $k$ est donc

$$\frac{p_i c_{ik}^2}{\lambda_k}$$

**Interprétation** la contribution d'un individu est importante si elle excède d'un facteur $\alpha$ le poids $p_i$ de l'individu concerné, c'est-à-dire

$$\frac{p_i c_{ik}^2}{\lambda_k} \geq \alpha p_i,$$

ou de manière équivalente

$$|c_{ik}| \geq \sqrt{\alpha \lambda_k}$$

**Choix de $\alpha$** selon les données, on se fixe en général une valeur de l'ordre de 2 à 4, que l'on garde pour *tous* les axes

### Individus sur-représentés

**Qu'est-ce que c'est ?** c'est un individu qui joue un rôle trop fort dans la définition d'un axe, par exemple

$$\frac{p_i c_{ik}^2}{\lambda_k} > 0,25$$

**Effet** il « tire à lui » l'axe $k$ et risque de perturber les représentations des autres points sur les axes de rang $\geq k$. Il est donc surtout problématique sur les premiers axes. Un tel individu peut être le signe de données erronées.

**Solution** on peut le retirer de l'analyse et le mettre en « individu supplémentaire ».

# Partie VII. Qualité de l'analyse

### Qualité globale de la représentation

**Calcul de l'inertie** on se souvient que $I_{\mathbf{g}} = \text{Tr}(\mathbf{VM})$ ; comme la trace d'une matrice est la somme de ses valeurs propres, on a

$$I_{\mathbf{g}} = \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$

**Définition** la qualité de la représentation obtenue par $k$ valeurs propres est la proportion de l'inertie expliquée

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

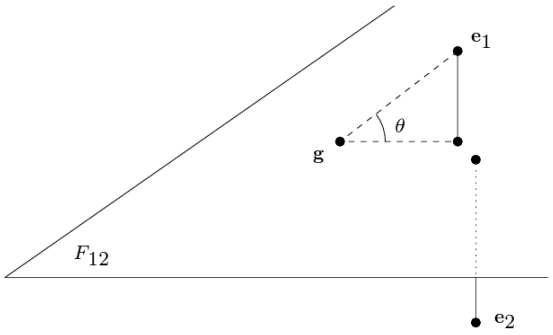
Si par exemple $\lambda_1 + \lambda_2$ est égal 90% de $I_{\mathbf{g}}$, le nuage de points est aplati autour du premier plan principal.

**Variables centrées réduites** On a $I_{\mathbf{g}} = \text{Tr}(\mathbf{R}) = p$ : la somme des valeurs propres est le nombre de variables.

**Utilisation** cette valeur sert seulement à évaluer la projection retenue, pas à choisir le nombre d'axes à garder.

### Qualité locale de la représentation

**But** on cherche à déterminer si le nuage de points est très aplati par la projection sur les sous-espaces principaux. Dans ce cas, deux individus éloignés pourraient artificiellement sembler proches les uns des autres.




### Angle entre un individu et un axe principal

Il est défini par son cosinus carré. Le cosinus de l'angle entre l'individu centré $i$ et l'axe principal $k$ est

$$\cos(\widehat{\mathbf{e}_i, \mathbf{a}_k}) = \frac{\langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k \rangle_{\mathbf{M}}}{\|\mathbf{e}_i - \mathbf{g}\|_{\mathbf{M}}}.$$

car les $\mathbf{a}_k$ forment une base orthonormale. Comme $\langle \mathbf{e}_i - \mathbf{g}, \mathbf{a}_k \rangle_{\mathbf{M}} = c_{ik}$,

$$\cos^2(\widehat{\mathbf{e}_i, \mathbf{a}_k}) = \frac{c_{ik}^2}{\sum_{\ell=1}^p c_{i\ell}^2}.$$

Cette grandeur mesure la qualité de la représentation de l'individu $i$ sur l'axe principal $\mathbf{a}_j$.

### Angle entre un individu et un sous-espace principal

C'est l'angle entre l'individu et sa projection orthogonale sur le sous-espace. La projection de $\mathbf{e}_i - \mathbf{g}$ sur le sous-espace $F_q$, $q \leq p$, est $\sum_{k=1}^q c_{ik}\mathbf{a}_k$, et donc

$$\cos^2(\widehat{\mathbf{e}_i, F_q}) = \frac{\sum_{k=1}^q c_{ik}^2}{\sum_{k=1}^p c_{ik}^2}.$$

La qualité de la représentation de l'individu $i$ sur le plan $F_q$ est donc la somme des qualités de représentation sur les axes formant $F_q$.

**Critères** Un $\cos^2$ égal à $0,9$ correspond à un angle de 18 degrés. Par contre, une valeur de $0,5$ correspond à un angle de 45 degrés !

On peut considérer les valeurs supérieures à $0,80$ comme bonnes et des valeurs inférieures à $0,5$ comme mauvaises. Une mauvaise qualité n'est significative que quand le point projeté n'est pas trop près de $\mathbf{0}$.

# Partie VIII. Interprétation externe

## Variables supplémentaires quantitatives

**Motivation** les composantes principales étant définies pour maximiser les contributions, le fait que les corrélations obtenues soient proches de 1 peut ne pas être significatif. Par contre, une corrélation forte entre une composante principale et une variable n'ayant pas participé à l'analyse est très significative.

**Méthode** on « met de coté » certaines variables pour qu'elles ne soient pas utilisées dans l'analyse (on diminue donc la dimension de $\mathbf{R}$ en enlevant des lignes et des colonnes). On cherche ensuite à savoir si elles sont liées à un axe donné.

**Corrélation** on calcule la corrélation de la variable avec les composantes principales et on la place dans le cercle des corrélations. Si $\hat{\mathbf{z}}$ est le vecteur centré-réduit correspondant à cette variable, on calcule

$$\mathrm{cor}(\hat{\mathbf{z}}, \mathbf{c}_k) = \frac{\mathrm{cov}(\hat{\mathbf{z}}, \mathbf{c}_k)}{\sqrt{\mathrm{var}(\mathbf{c}_k)}} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} p_i \hat{z}_i c_{ik}.$$

On peut éventuellement utiliser un test statistique pour déterminer si une corrélation est significative.

## Variables supplémentaires qualitatives

**Représentation** on peut représenter par des symboles différents les individus de chaque catégorie sur les axes principaux. Pour savoir si les étiquettes sont liées à l'axe $k$, on peut calculer la coordonnée $\hat{c}_k$ de leur barycentre sur cet axe. Problème : comment l'interpréter ?

**Valeur-test** on considère les $\hat{n}$ individus parmi $n$ ayant une certaine caractéristique (homme, femme...) et la coordonnée $\hat{c}_k$ de leur barycentre sur la $k$-ième composante principale. La valeur-test est

$$\hat{c}_k \sqrt{\frac{\hat{n}}{\lambda_k}} \sqrt{\frac{n-1}{n-\hat{n}}}.$$

**Usage** Elle est significative si :
- $\hat{n}$ et $n - \hat{n}$ sont assez grands (en général> 30, pour que le théorème central limite s'applique)
- sa valeur absolue est supérieure à 2 ou 3.

Sinon, on dira qu'on ne peut pas affirmer si la catégorie est liée à l'axe

**Idée du calcul** Si les $\hat{n}$ individus étaient pris au hasard, $\hat{c}_k$ serait une variable aléatoire centrée (les $\mathbf{z}$ sont de moyenne nulle) et de variance $\frac{\lambda_k}{\hat{n}} \frac{n-\hat{n}}{n-1}$ car le tirage est sans remise.

## Individus supplémentaires

**Méthode** on « met de coté » certains individus pour qu'ils ne soient pas utilisées dans l'analyse (ils ne sont pas pris en compte dans le calcul des covariances). On cherche ensuite à savoir si ils sont liés à un axe donné.

**Cas des individus sur-représentés** on peut décider d'utiliser ces points en individus supplémentaires, en particulier quand les points constituent un échantillon et ne présentent pas d'intérêt en eux-mêmes.

**Représentation** on les ajoute à la représentation sur les plans principaux. Pour calculer leur coordonnée sur un axe fixé, on écrit

$$\hat{c}_k = \sum_{j=1}^{p} \hat{z}^j u_{jk},$$

où les $\hat{z}^j$ sont les coordonnées centrées-réduites d'un individu supplémentaire $\hat{\mathbf{z}}$.

Ces individus peuvent servir d'échantillon-test pour vérifier les hypothèses tirées de l'ACP sur les individus actifs.

# Partie IX. L'ACP en trois transparents

## Un

**Données** les données représentent les valeurs de $p$ variables mesurées sur $n$ individus ; les individus peuvent avoir un poids. En général (et dans ce résumé), on travaille sur des données centrées réduites $\mathbf{Z}$ (on retranche la moyenne et on divise par l'écart type).

**Matrice de corrélation** c'est la matrice $\mathbf{R}$ de variance-covariance des variables centrées réduites. Elle possède $p$ valeurs propres $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

**Inertie totale** c'est la moitié de la moyenne des distances au carré entre les individus ; elle mesure l'étendue du nuage de points. C'est la grandeur qu'on cherche à garder maximale et elle peut s'écrire

$$I_{\mathbf{g}} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = p.$$

**Facteurs principaux $\mathbf{u}_k$** ce sont des vecteurs propres orthonormés de $\mathbf{R}$ associés aux $\lambda_k$ : $\mathbf{R}\mathbf{u}_k = \lambda_k \mathbf{u}_k$. Leur $j$-ième composante (sur $p$) $u_{jk}$ est le poids de la variable $j$ dans la composante $k$.

**Composantes principales $\mathbf{c}_k$** ce sont les vecteurs $\mathbf{Z}\mathbf{u}_k$ de dimension $n$. Leur $i$-ième coordonnée $c_{ik}$ est la valeur de la composante $k$ pour l'individu $i$. Les $\mathbf{c}_k$ sont décorrélées et leur variance est $\mathrm{var}(\mathbf{c}_k) = \lambda_k$.

## Deux

**Nombre d'axes** on se contente en général de garder les axes *interprétables* de valeur propre supérieure à 1 (critère de Kaiser).

**Cercle des corrélations** il permet de visualiser comment les variables sont corrélées (positivement ou négativement) avec les composantes principales. À partir de là, on peut soit trouver une signification physique à chaque composante, soit montrer que les composantes séparent les variables en paquets.

**Représentation des individus** pour un plan principal donné, la représentation des projections des individus permet de confirmer l'interprétation des variables. On peut aussi visualiser les individus aberrants (erreur de donnée ou individu atypique).

**Contribution d'un individu à une composante** c'est la part de la variance d'une composante principale qui provient d'un individu donné. Si cette contribution est supérieur de 2 à 4 fois au à son poids, l'individu définit la composante. Si elle est très supérieure aux autres, on dit qu'il est *sur-représenté* et on peut avoir intérêt à mettre l'individu en donnée supplémentaire.

## Trois

**Qualité globale de la représentation** c'est la part de l'inertie totale $I_{\mathbf{g}}$ qui est expliquée par les axes principaux qui ont été retenus. Elle permet de mesurer la précision et la pertinence de l'ACP.

**Qualité de la représentation d'un individu** elle permet de vérifier que tous les individus sont bien représentés par le sous-espace principal choisi ; elle s'exprime comme le carré du cosinus de l'angle entre l'individu et sa projection orthogonale.

**Individus supplémentaires** quand un individu est sur-représenté sur un des premiers axes, on peut le supprimer de l'analyse et le réintroduire dans la représentation comme individu supplémentaire.

**Variables supplémentaires quantitatives** certaines variables peuvent être mises de coté lors de l'ACP et reportées séparément sur le cercle des corrélation.

**Variables supplémentaires qualitatives** elles peuvent être représentées sur la projection des individus, et leur liaison aux axes est donnée par les valeurs-test.