

Variables quantitatives : analyse en composantes principales

Jean-Marc Lasgouttes

<http://ana-donnees.lasgouttes.net/>

Préambule : 3 approches des données

Décrire les données de 3 manières complémentaires

- *statistique* : chaque colonne représente une variable mesurée sur différent individus,
- *matricielle* : le tableau complet de données est une matrice de nombres réels,
- *géométrique* : 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

- *vision statistique* : moyenne, variance, corrélation ;
- *vision matricielle* : valeurs propres, vecteurs propres ;
- *vision géométrique* : 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 $p_1 + \dots + p_n = 1$, notamment quand les individus n'ont pas la même importance (échantillons redressés, données regroupées,...). On a souvent $p = 1/n$.

Moyenne arithmétique On note

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

ou pour des données non pondérées

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} [x_1 + x_2 + \dots + x_n].$$

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

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

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

Propriétés La variance satisfait la formule suivante

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

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

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

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

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

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)

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

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

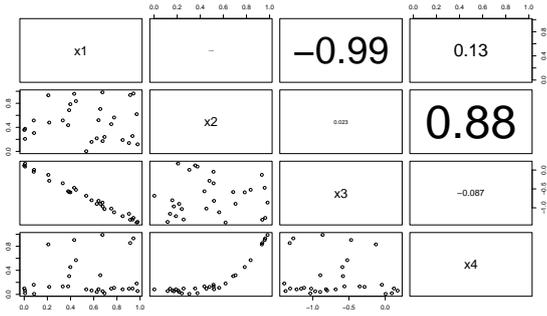
$$ax_i + by_i = c, \text{ pour tout } 1 \leq i \leq n.$$

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

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

Le coefficient de corrélation par l'exemple

<http://www.tylervigen.com/spurious-correlations>.



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 ?

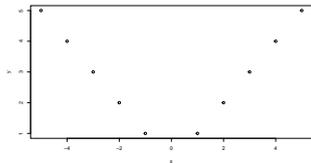
Qu'est ce qui est significatif ? si on a assez de données, on peut considérer qu'une corrélation supérieure à 0,5 est significative, et une corrélation entre 0,3 et 0,5 est faible.

Une corrélation égale à 1 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 :

En 2016, 59,2 % des décès ont eu lieu dans des établissements de santé (hôpital ou clinique) et 26% à domicile. L'hôpital est-il dangereux pour la santé ?

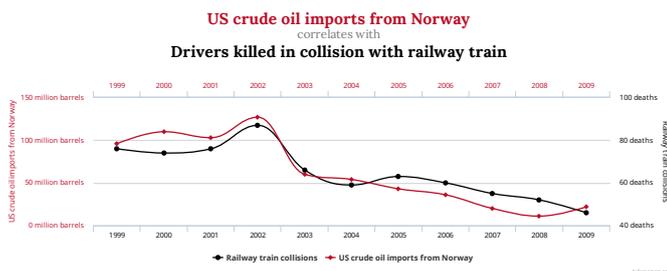
Et une décorrélation ? voici un exemple où $\text{cor}(\mathbf{x}, \mathbf{y}) = 0$



Fausse corrélation

Quand ? Elles peuvent se trouver quand on a peu de données

Exemple Importations de pétrole brut de la Norvège vers les États-Unis et nombre de conducteurs tués par une collision avec un train : $r = 0,95$ entre 1999 et 2009.



Exemple issu du site *Spurious Correlations*

Partie II. Données : vision matricielle

Notation

Matrice tableau de données, notée 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 :

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

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

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

$$\text{Tr}(\alpha \mathbf{A}) = \alpha \text{Tr}(\mathbf{A}), \quad \text{Tr}(\mathbf{A} + \mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B}),$$

$$\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA}),$$

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

Tableau de données

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

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

Un *individu* est représenté par

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

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

$$\mathbf{D}_p = \begin{bmatrix} p_1 & & 0 \\ & p_2 & \\ & & \ddots \\ 0 & & & p_n \end{bmatrix}$$

Symétrie La matrice \mathbf{D}_p est diagonale et donc symétrique : $\mathbf{D}'_p = \mathbf{D}_p$.

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 :

$$\mathbf{g}' = (\bar{x}^1, \dots, \bar{x}^p) = \sum_{i=1}^n p_i \mathbf{e}'_i.$$

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

$$y_i^j = x_i^j - \bar{x}^j, \quad \text{c'est-à-dire} \quad \mathbf{y}^j = \mathbf{x}^j - \bar{x}^j \mathbf{1}_n$$

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 $\sigma_{j\ell}$ est la covariance des variables \mathbf{x}^j et \mathbf{x}^ℓ et σ_j^2 est la variance de la variable \mathbf{x}^j

Symétrie Comme $\sigma_{j\ell} = \sigma_{\ell j}$, la matrice \mathbf{V} est symétrique : $\mathbf{V}' = \mathbf{V}$.

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 $r_{j\ell} = \sigma_{j\ell}/\sigma_j\sigma_\ell$, c'est la matrice $p \times p$

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

Symétrie Comme $r_{j\ell} = r_{\ell j}$, la matrice \mathbf{R} est symétrique : $\mathbf{R}' = \mathbf{R}$.

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

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

Les données centrées réduites

Définition c'est la matrice \mathbf{Z} contenant les données

$$z_i^j = \frac{y_i^j}{\sigma_j} = \frac{x_i^j - \bar{x}^j}{\sigma_j}, \quad \text{c'est-à-dire} \quad \mathbf{z}^j = \frac{\mathbf{y}^j}{\sigma_j}$$

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 ne pas privilégier les variables dispersées.

Covariances Les covariances des \mathbf{z}^j sont des corrélations :

$$\text{cov}(\mathbf{z}^k, \mathbf{z}^\ell) = \sum_{i=0}^n p_i z_i^k z_i^\ell = \frac{1}{\sigma_k \sigma_\ell} \sum_{i=0}^n p_i y_i^k y_i^\ell = \text{cor}(\mathbf{x}^k, \mathbf{x}^\ell),$$

où on a utilisé le fait que $\bar{z}^j = \bar{y}^j = 0$. La matrice de variance-covariance des variables centrées-réduites est donc la matrice de corrélation \mathbf{R} .

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

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

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 $\mathbf{M} = \text{diag}(m_j)$, où m_1, \dots, m_p sont des réels strictement positifs. On pose

$$\|\mathbf{u}\|_{\mathbf{M}}^2 = \sum_{j=1}^p m_j u_j^2 = \mathbf{u}'\mathbf{M}\mathbf{u}, \quad d_{\mathbf{M}}^2(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_{\mathbf{M}}^2.$$

Espace métrique il est défini par le produit scalaire

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{M}} = \sum_{j=1}^p m_j u_j v_j = \mathbf{u}'\mathbf{M}\mathbf{v}, \quad \langle \mathbf{u}, \mathbf{u} \rangle_{\mathbf{M}} = \|\mathbf{u}\|_{\mathbf{M}}^2.$$

Propriétés Le produit scalaire est commutatif, linéaire et satisfait l'identité

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

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

Cas particuliers

Métrique usuelle Si $m_1, \dots, m_p = 1$, alors $\mathbf{M} = \mathbf{I}_p$ et on note $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{I}}$.

Métrique réduite diviser les variables par σ_j est équivalent à prendre $m_j = 1/\sigma_j^2$. On a $\mathbf{D}_{1/\sigma^2} = \mathbf{D}_{1/\sigma}\mathbf{D}_{1/\sigma}$ et donc

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

Travailler avec la métrique \mathbf{D}_{1/σ^2} , c'est comme utiliser la métrique \mathbf{I} sur des variables réduites.

La plupart du temps en ACP, on fait l'analyse avec la métrique usuelle sur les données centrées-réduites.

Partie IV. L'analyse en composantes principales

Inertie

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

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

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

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

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

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

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

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

Calcul de l'inertie

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

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

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

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

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

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

Variables centrées réduites On se retrouve encore dans le cas où

$$I_{\mathbf{g}} = \text{Tr}(\mathbf{R}) = p.$$

L'analyse de composantes principales (version 2)

Principe on cherche à projeter \mathbf{M} -orthogonalement le nuage de points sur un espace F_k de dimension $k < p$, sous la forme

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

Les vecteurs $\mathbf{a}_1, \dots, \mathbf{a}_k$ 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 , qui maximise l'inertie du nuage projeté sur F_k .

Résultat principal

Propriété Il existe p réels $\lambda_1, \dots, \lambda_p$ positifs ou nuls et p vecteurs $\mathbf{a}_1, \dots, \mathbf{a}_p$, tels que

$$\mathbf{VM}\mathbf{a}_k = \lambda_k \mathbf{a}_k.$$

— Les λ_k sont les *valeurs propres* de \mathbf{VM} et sont classées par ordre décroissant :

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

— Les \mathbf{a}_k sont les *vecteurs propres* de \mathbf{VM} et sont « \mathbf{M} -orthonormaux » :

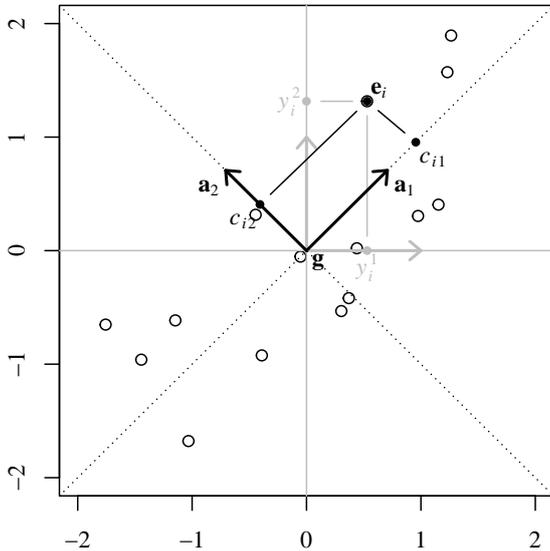
$$\langle \mathbf{a}_k, \mathbf{a}_k \rangle_{\mathbf{M}} = 1, \quad \langle \mathbf{a}_k, \mathbf{a}_\ell \rangle_{\mathbf{M}} = 0 \text{ si } k \neq \ell.$$

Théorème principal La projection sur k variables qui maximise l'inertie projetée est obtenue en considérant les k premières valeurs propres $\lambda_1, \dots, \lambda_k$ et les $\mathbf{a}_1, \dots, \mathbf{a}_k$ correspondants, appelés axes principaux.

Le calcul ne dépend pas du nombre de variables retenues.

Partie V. Les éléments de l'ACP

Changement de coordonnées



$$\mathbf{e}_i - \mathbf{g} = (y_i^1, y_i^2)' = y_i^1(1, 0)' + y_i^2(0, 1)' = c_{i1}\mathbf{a}_1 + c_{i2}\mathbf{a}_2$$

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

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

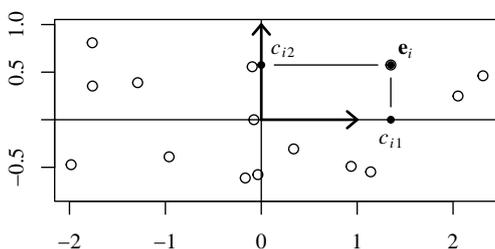
Composantes principales ce sont les variables $\mathbf{c}_k = (c_{1k}, \dots, c_{nk})$ de taille n définies par

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

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 λ_k car

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

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

$$\text{cov}(\mathbf{c}_k, \mathbf{c}_\ell) = \mathbf{c}_k' \mathbf{D}_p \mathbf{c}_\ell = \dots = \lambda_\ell \mathbf{a}_k' \mathbf{M} \mathbf{a}_\ell = 0.$$

Les composantes principales ne sont pas corrélées entre elles.

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 $\mathbf{M} \mathbf{V}$ 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 $\mathbf{M} \mathbf{V}$, 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

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

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

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

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

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

Métrie on prend la métrie $\mathbf{M} = \mathbf{I}_p$.

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

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

Les valeurs propres vérifient

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

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.

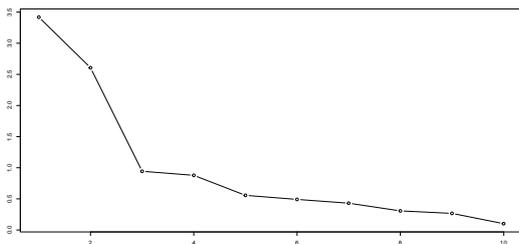
Les seuls critères utilisables sont empiriques.

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.

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



Cas des variables liées

Contexte Il arrive que plusieurs variables soient liées, par exemple parce que leur somme est connue (ex. 100% pour des pourcentages).

Redondance des variables On pourrait alors vouloir retirer une des variables, qui peut être retrouvée par les autres. Mais on perdrait l'interprétation de la variable.

Effet sur l'ACP Il n'y a pas de réel problème

- pour chaque relation entre les variables, on aura une valeur propre nulle.
- le nombre de valeurs propres retournées par le logiciel sera souvent réduit d'autant, même si la somme des valeurs propres reste toujours égale à p .

Remarque Il est important de repérer de telles relations dans la phase initiale d'étude des données.

Corrélation entre composantes et variables initiales

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

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

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

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

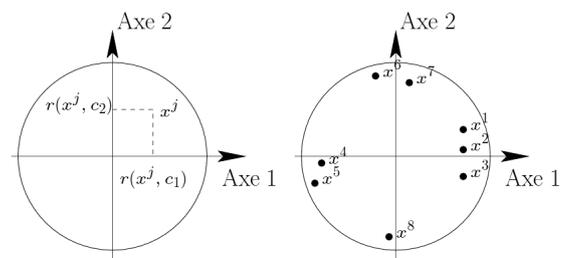
$$\begin{aligned} \text{var}(\mathbf{z}^j) &= \text{cov}(\mathbf{z}^j, \mathbf{z}^j) = \text{cov}\left(\mathbf{z}^j, \sum_{k=1}^p a_{kj} \mathbf{c}_k\right) = \sum_{k=1}^p a_{kj} \text{cov}(\mathbf{z}^j, \mathbf{c}_k) \\ &= \sum_{k=1}^p \lambda_k a_{kj}^2 = \sum_{k=1}^p [\text{cor}(\mathbf{z}^j, \mathbf{c}_k)]^2. \end{aligned}$$

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

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

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 $\text{cor}(\mathbf{z}^j, \mathbf{c}_1)$ et d'ordonnée $\text{cor}(\mathbf{z}^j, \mathbf{c}_2)$.



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.

Remarque 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 α 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 α 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

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

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

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

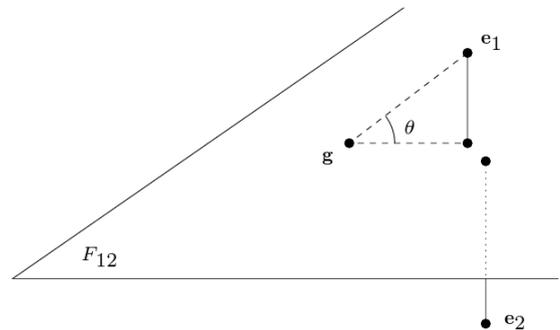
Si par exemple $\lambda_1 + \lambda_2$ est égal 90% de $I_{\mathbf{g}}$, le nuage de points est aplati autour du premier plan principal.

Variables centrées réduites On a $I_{\mathbf{R}} = \text{Tr}(\mathbf{R}) = p$: 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

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

et comme les \mathbf{a}_k forment une base orthonormale,

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

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

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.

Attention ! Une mauvaise qualité n'est significative que quand le point projeté n'est pas trop près de $\mathbf{0}$. Sinon on ne peut pas conclure à partir de ce simple nombre.

Qualité d'un individu vs. contribution

Importance d'un individu sur un axe On peut considérer que plus $p_i c_{ik}^2$ est grand, plus l'individu i est important sur l'axe k .

Problème grand par rapport à quoi ?

Contribution on compare aux autres individus en divisant par la somme sur la colonne k , ce qui donne

$$\frac{p_i c_{ik}^2}{p_1 c_{1k}^2 + p_2 c_{2k}^2 + \dots + p_n c_{nk}^2} = \frac{p_i c_{ik}^2}{\lambda_k}.$$

On retrouve alors la formule de la contribution de l'individu i à l'axe k .

Qualité on compare aux autres axes en divisant par la somme sur la ligne i , qui est

$$\frac{p_i c_{ik}^2}{p_i c_{i1}^2 + p_i c_{i2}^2 + \dots + p_i c_{ip}^2} = \frac{c_{ik}^2}{c_{i1}^2 + c_{i2}^2 + \dots + c_{ip}^2}.$$

C'est la qualité de représentation de l'individu i par l'axe k .

Partie VIII. Interprétation externe

Variables supplémentaires quantitatives

Motivation 1 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.

Motivation 2 les variables peuvent naturellement se séparer en deux paquets : offre/demande, produits détenus par des clients et données personnelles (âge, nombre d'enfants, revenu), etc. On cartographie le premier paquet et projette le second dessus.

Méthode on « met de côté » 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 : si $\hat{\mathbf{z}}$ est le vecteur centré-réduit correspondant à cette variable, c'est

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

On peut 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^{e} 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 (un peu significative) ou 3 (significative).

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 côté » certains individus pour qu'ils ne soient pas utilisés 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 $\hat{\mathbf{z}}$.

Autre utilisation 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 $\lambda_1 \geq \dots \geq \lambda_p \geq 0$.

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

$$I_g = \lambda_1 + \lambda_2 + \dots + \lambda_p = p.$$

Facteurs principaux \mathbf{u}_k ce sont des vecteurs propres orthonormés de \mathbf{R} associés aux λ_k : $\mathbf{R}\mathbf{u}_k = \lambda_k\mathbf{u}_k$. Leur j^{e} 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^{e} coordonnée c_{ik} est la valeur de la composante k pour l'individu i . Les \mathbf{c}_k sont décorrélés et leur variance est $\text{var}(\mathbf{c}_k) = \lambda_k$.

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_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.

Résumé des notations

Notation	taille	description
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	$n \times p$	données brutes/centrées/centrées-réduites
$\mathbf{x}^j, \mathbf{y}^j, \mathbf{z}^j$	n	variable brute/centrée/centrée-réduite
\mathbf{p}	n	poids p_1, \dots, p_n des individus (de somme égale à 1).
\mathbf{D}_p	$n \times n$	matrice de poids des individus (diagonale)
σ_j^2	réel > 0	variance de \mathbf{x}^j
$\sigma_{j\ell}, r_{j\ell}$	réel	covariance/corrélation de \mathbf{x}^j et \mathbf{x}^ℓ
\mathbf{V}, \mathbf{R}	$p \times p$	matrice de variance-covariance/corrélation de \mathbf{X}
\mathbf{M}	$p \times p$	métrique sur les variables (diagonale)
\mathbf{c}_k	n	composante principale (nouvelle variable)
λ_k	réel ≥ 0	Variance de \mathbf{c}_k . On a $\lambda_1 \geq \dots \geq \lambda_p \geq 0$.
\mathbf{a}_k	p	axe principal : poids de \mathbf{c}_k dans chaque variable \mathbf{z}^j
\mathbf{u}_k	p	facteur principal : poids de chaque variable dans \mathbf{c}_k