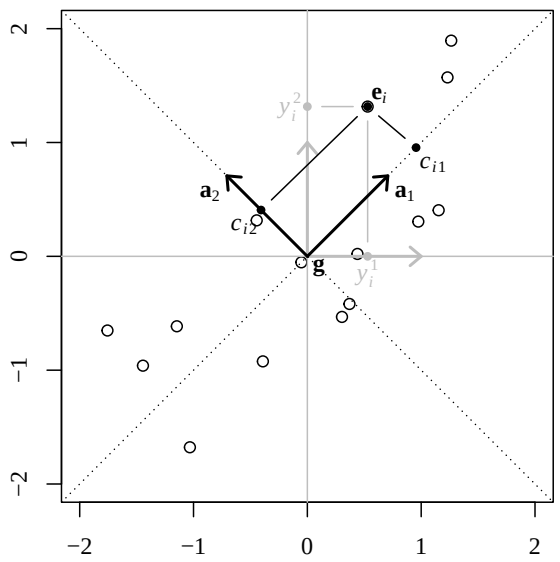
Partie V. Les éléments de l'ACP

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

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

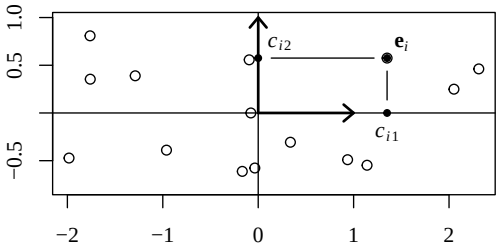
Composantes principales ce sont les variables $\mathbf{c}_k = (c_{1k}, \ldots, c_{nk})$ de taille n définies par

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

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

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

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

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

Les composantes principales ne sont pas corrélées entre elles.

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

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

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

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

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.

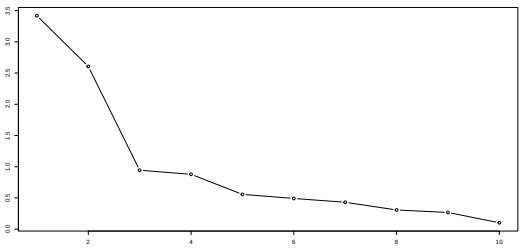
Les seuls critères utilisables sont empiriques.

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.

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

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

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

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

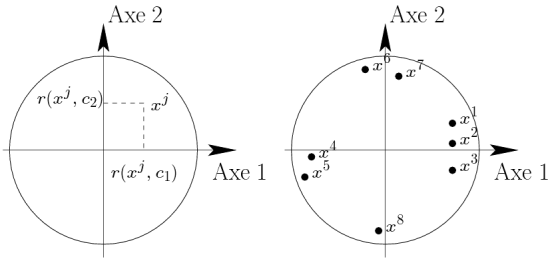
$$\operatorname{var}(\mathbf{z}^j) = \operatorname{cov}(\mathbf{z}^j, \mathbf{z}^j) = \operatorname{cov}\Big(\mathbf{z}^j, \sum_{k=1}^{p} a_{kj}\mathbf{c}_k\Big) = \sum_{k=1}^{p} a_{kj} \operatorname{cov}(\mathbf{z}^j, \mathbf{c}_k)$$
$$= \sum_{k=1}^{p} \lambda_k a_{kj}^2 = \sum_{k=1}^{p} \big[\operatorname{cor}(\mathbf{z}^j,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

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 $\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_2)$.



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 » !