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 $\mathbf{R}$,

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

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.

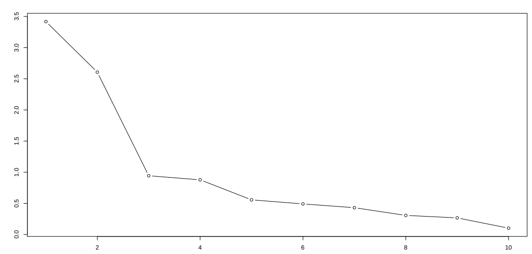
Les seuls critères utilisables sont empiriques.

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 retournée par le logiciel sera souvent réduit d'autant, même si la somme des variables 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

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

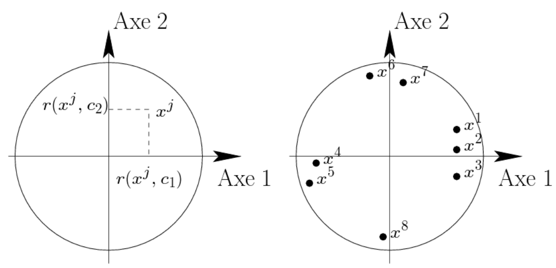
$$= \sum_{k=1}^{p} \lambda_k a_{kj}^2 = \sum_{k=1}^{p} \big[\operatorname{cor}(\mathbf{z}^j, \mathbf{c}_k)\big]^2.$$

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 » !

Contribution d'un individu à une composante

Définition On sait que $\mathrm{var}(\mathbf{c}_k) = \lambda_k = \sum_{i=1}^n p_i c_{ik}^2$. 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 $I_{\mathbf{g}} = \mathrm{Tr}(\mathbf{VM})$; 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

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

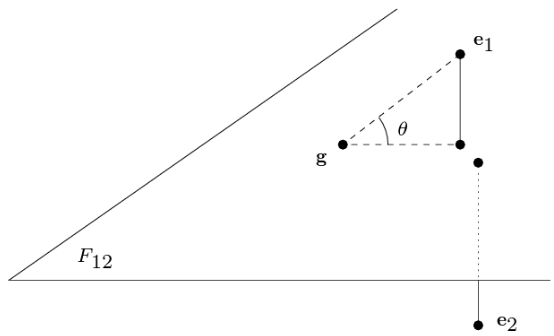
Si par exemple $\lambda_1 + \lambda_2$ est égal 90% de $I_{\mathbf{g}}$, le nuage de points est aplati autour du premier plan principal.

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