Corrélation entre composantes et variables initiales

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

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

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

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

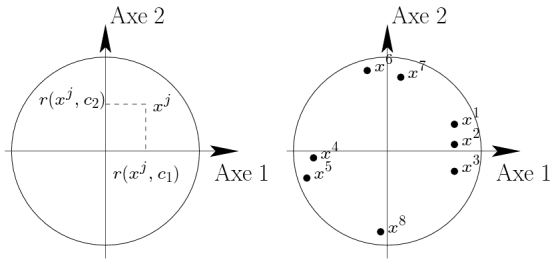
$$\mathrm{var}(\mathbf{z}^j) = \mathrm{cov}(\mathbf{z}^j, \mathbf{z}^j) = \mathrm{cov}\Big(\mathbf{z}^j, \sum_{k=1}^{p} a_{kj}\mathbf{c}_k\Big) = \sum_{k=1}^{p} a_{kj}\,\mathrm{cov}(\mathbf{z}^j, \mathbf{c}_k)$$
$$= \sum_{k=1}^{p} \lambda_k a_{kj}^2 = \sum_{k=1}^{p} \big[\mathrm{cor}(\mathbf{z}^j, \mathbf{c}_k)\big]^2.$$

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

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

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



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 » !

Un autre interprétation de l'ACP

Propriété $\mathbf{c}_1$ est la variable la plus liée aux $\mathbf{x}^j$ au sens des sommes des carrés des corrélations

$$\sum_{j=1}^{p} \big[\mathrm{cor}(\mathbf{v}, \mathbf{x}^j)\big]^2 \text{ est maximal pour } \mathbf{v} = \mathbf{c}_1$$

Si on se restreint aux $\mathbf{v}$ orthogonaux à $\mathbf{c}_1$, le maximum sera atteint pour $\mathbf{c}_2$, et ainsi de suite.

Interprétation de l'ACP normée elle revient à remplacer les variables corrélées $\mathbf{x}^1, \ldots, \mathbf{x}^p$ par de nouvelles variables $\mathbf{c}^1, \ldots, \mathbf{c}^p$ non corrélées entre elles, de variance maximale et les plus liées en un certain sens aux $\mathbf{x}^j$.

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

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_{k^*}}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

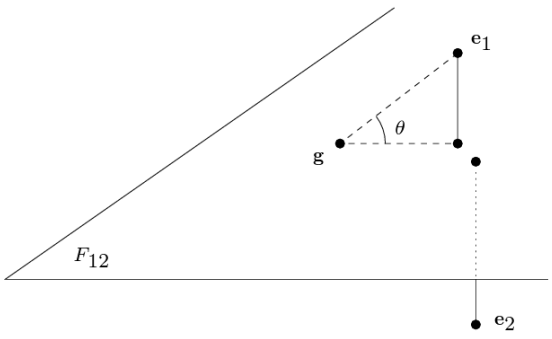
Si par exemple $\lambda_1 + \lambda_2$ est égal 90% de $I_{\mathbf{g}}$, le nuage de points est aplati autour du premier plan principal.

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_{k^*}, $k^* \leq p$, est $\sum_{k=1}^{k^*} c_{ik}\mathbf{a}_k$, et donc

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

La qualité de la représentation de l'individu i sur le plan F_{k^*} est donc la somme des qualités de représentation sur les axes formant F_{k^*}.

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

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

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

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

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