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

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

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

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

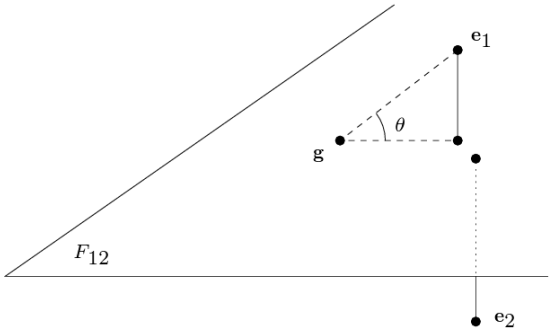
Si par exemple $\lambda_1 + \lambda_2$ est égal 90% de $I_{\mathbf{g}}$, le nuage de points est aplati autour du premier plan principal.

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

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

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

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

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.

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 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 : si $\hat{\mathbf{z}}$ est le vecteur centré-réduit correspondant à cette variable, c'est

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

On peut 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 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 $\hat{\mathbf{z}}$.

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