

**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 **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 **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

**Idee 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  $\hat{\mathbf{z}}$ .

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  $\lambda_k$  :  $\mathbf{R}\mathbf{u}_k = \lambda_k\mathbf{u}_k$ . Leur  $j^{\text{e}}$  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  $i^{\text{e}}$  coordonnée  $c_{ik}$  est la valeur de la composante  $k$  pour l'individu  $i$ . Les  $\mathbf{c}_k$  sont décorréliées 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 côté lors de l'ACP et reportées séparément sur le cercle des corrélations.

**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)
$\sigma_j^2$	réel $> 0$	variance de $\mathbf{x}^j$
$\sigma_{j\ell}, r_{j\ell}$	réel	covariance/corrélation de $\mathbf{x}^j$ et $\mathbf{x}^\ell$
$\mathbf{V}, \mathbf{R}$	$p \times p$	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)
$\lambda_k$	réel $\geq 0$	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 $\mathbf{c}_k$