Analyse en composantes principales
{ACP}

Who? I. Laroussi

From? [1]Département de Mathématiques

When? Cours de Master I, 2022

Plans du cours

Motivation

Lorsqu'on étudie simultanément un nombre important de variables quantitatives, comment faire un graphique global ?
La difficulté vient du fait que les individus étudiés ne sont plus représentés dans un plan (espace de dimension 2), mais dans un espace avec une dimension assez grande (par exemple 4). Notre travail est consacré à l'étude de ce genre de tableaux et l'intérêt principal est l'extraction de la plus grande quantité d'information contenue dans les différentes mesures des caractères.

Plans du cours

Tableau des données

L'ACP s'applique à des tableaux croisant des individus et des variables quantitatives, à l'intersection de la ligne i et de la colonne j se trouve la valeur de la variable j pour l'individu i noté x_{ij} et le tableau de données sur lequel on va faire l'analyse est le suivant

Tableau des données

$$X = \begin{array}{c} \\ U_1 \\ U_2 \\ . \\ U_i \\ . \\ U_n \end{array} \begin{array}{c} \begin{array}{cccccc} V_1 & V_2 & \dots & V_j & \dots & V_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ . & . & \dots & . & \dots & . \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ . & . & \dots & . & \dots & . \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \end{array} .$$

Tableau des données

On peut représenter chaque unité par le vecteur de ses mesures sur les p variables par

$$U_i^t = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \end{bmatrix}$$

ce qui donne

$$U_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ x_{ij} \\ . \\ x_{ip} \end{bmatrix}.$$

Tableau des données

On peut représenter chaque variable par le vecteur de ses mesures sur les n individus par

$$V_j^t = \begin{bmatrix} x_{1j} & x_{2j} & \dots & x_{ij} & \dots & x_{nj} \end{bmatrix}$$

ce qui donne

$$V_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ x_{ij} \\ . \\ x_{nj} \end{bmatrix}.$$

Les objectifs de l'ACP

L'idée générale des méthodes factorielles est de trouver un système d'axes et de plans tels que les projections de ces nuages de points sur ces axes et ces plans permettent de reconstituer les positions et les distances des points les uns par rapport aux autres. Pour faire une représentation géométrique, il faut choisir une distance entre deux points de l'espace. La distance usuelle dans l'analyse en composantes principales est la distance euclidienne classique.

Les objectifs de l'ACP

1er but

Les deux voies principales de cette exploration sont

- Un bilan des ressemblances entre individus tel que deux individus se ressemblent d'autant plus qu'ils possèdent des valeurs proches pour l'ensemble des variables ou en ACP, la distance $d^2(u_i; u_l)$ entre deux individus i et l est défnie par

$$d^2(u_i; u_l) = \sum_{i=1}^{p} (x_{ij} - x_{lj})^2 :$$

Les objectifs de l'ACP

A cette distance on associe un produit scalaire entre deux vecteurs u_i et u_l par

$$\langle \overrightarrow{u_i}, \overrightarrow{u_l} \rangle = \sum_{j=1}^{p} x_{ij} x_{lj} = u_i^t u_l,$$

Ainsi que la norme d'un vecteur est donnée par

$$\| \overrightarrow{u_i} \|^2 = \sum_{j=1}^{p} x_{ij}^2 = u_i^t u_i.$$

Les objectifs de l'ACP

On peut alors définir l'angle α entre deux vecteurs par le cosinus

$$\cos(\alpha) = \frac{\langle \overrightarrow{u_i}, \overrightarrow{u_l} \rangle}{\| \overrightarrow{u_i} \| \| \overrightarrow{u_l} \|} = \frac{\sum_{j=1}^{p} x_{ij} x_{lj}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2 \sum_{j=1}^{p} x_{lj}^2}} = \frac{u_i^t u_l}{\sqrt{u_i^t u_i u_l^t u_l}}$$

Les objectifs de l'ACP

1er but

Plus généralement, on cherche à mettre en évidence des groupes homogènes d'individus dans le cadre d'une typologie des individus.

Les objectifs de l'ACP

2ième but

- En deuxième plan, on cherche un bilan des liaisons entre variables et cela en marquant la liaison entre deux variables$(V_j; V_k)$ est en mesurant cette corrélation par le coefficient de corrélation linéaire, notéρ définit par

$$\rho(V_j; V_k) = \frac{Cov(V_j; V_k)}{\sqrt{Var(V_j)Var(V_k)}}.$$

Ce paramètre est défini pour avoir les variables qui vont dans le même sens, dans un sens opposé et d'une manière indépendante. Cela au lieu d'analyser le tableau à travers p variables.

Remarque

Les poids pour les individus

Si les données sont aléatoires à probabilités égales, alors les individus ont la même importance $\frac{1}{n}$, dans ce cas la masse totale de ces individus égale à 1, il y a des cas où on affecte des poids différents aux individus. Cette situation se présente lorsque les individus représentent chacun une sous-population; on affecte alors à un individu un poids proportionnel à l'effectif de la sous-population qu'il représente noté p_i, on utilise ce poids pour calculer la moyenne et la variance de chaque variable, et la liaison entre deux variables comme suit

Remarque

Les poids pour les variables

$$\bar{V}_j = \sum_{i=1}^{n} p_i x_{ij},$$

$$s_j^2 = \sum_{i=1}^{n} p_i (x_{ij} - \bar{V}_j)^2$$

$$\text{et } \rho(V_j, V_k) = \sum_{i=1}^{n} p_i \left(\frac{x_{ij} - \bar{v}_j}{s_j} \right) \left(\frac{x_{ik} - \bar{v}_k}{s_k} \right).$$

Remarque

Les poids pour les variables

Jusqu'ici on donne la même importance aux différentes variables, c'est rare qu'on affecte des importances différentes aux variables. Cette importance peut être modulée à l'aide d'un coefficient appelé poids de la variable. En notant m_j le poids de la variable j, la distance entre deux individus i et l est définie par

$$d^2(u_i; u_l) = \sum_{j=1}^{p} m_j (x_{ij} - x_{lj})^2.$$

Transformation des données

En ACP le tableau de données dépend de quelques transformations nécessaires puisque le nuage des points-individus peut être centré ou non, réduit ou non. Pour cela, l'analyse en composantes principales normée (centré-réduit) est certainement la plus utilisée ainsi toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale.

Transformation des données

Le centre de gravité

Le centre de gravité G du nuage des individus est alors le point dont les coordonnées sont les valeurs moyennes des variables j

$$G_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad = x_{\bullet j}.$$

Transformation des données

Alors le centre de gravité G du nuage des individus est le point des valeurs moyennes des variables

$$G = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ .... \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ ... \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} = \begin{pmatrix} x_{\bullet 1} \\ x_{\bullet 2} \\ . \\ . \\ . \\ x_{\bullet j} \\ . \\ . \\ . \\ x_{\bullet p} \end{pmatrix}.$$

Les variables centrées

A chaque valeur numérique, on soustrait la moyenne de la variable en cause. Le tableau obtenu est alors de terme général

$$X_c = \begin{bmatrix} x_{11} - x_{\bullet 1} & \dots & x_{1j} - x_{\bullet j} & \dots & x_{1p} - x_{\bullet p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} - x_{\bullet 1} & \dots & x_{ij} - x_{\bullet j} & \dots & x_{ip} - x_{\bullet p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} - x_{\bullet 1} & \dots & x_{nj} - x_{\bullet j} & \dots & x_{np} - x_{\bullet p} \end{bmatrix}$$

Cette transformation ne change rien à la problématique, l'ACP peut être réalisée sur des données seulement centrées.

Plans du cours

Le moment d'inertie

Inertie totale

Nous allons à présent nous intéressé à la déformation du nuage des points lorsque nous avons changé de point de vue. On note I_G le moment d'inertie du nuage de points des individus par rapport au centre de gravité G.

$$\begin{aligned} I_G &= \frac{1}{n}\sum_{i=1}^{n} d^2(G, u_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - x_{\bullet j})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} u_{ci}^t u_{ci}, \end{aligned}$$

où u_{ci} est le vecteur centré u_i.

Le moment d'inertie

L'inertie est une mesure de dispersion de nuage des individus par rapport au centre de gravité. Si le moment d'inertie est petit, alors le nuage est bien centré sur l'origine et s'il est grand le nuage est écarté ou éloigné du centre de gravité. I_G peut aussi s'écrire sous la forme suivante

$$I_G = \sum_{j=1}^{p} \left[\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - x_{\bullet j})^2 \right] = \sum_{j=1}^{p} Var(V_j),$$

où $Var(V_j)$ est la variance empirique de la variable V_j, donc l'inertie total est égale à la trace de la matrice de variance-covariance noté Σ qui résume la structure des dépendances linéaires 2 à 2 des p variables.

Le moment d'inertie

Inertie du nuage des individus par rapport à un axe passant par G

L'inertie du nuage des individus par rapport à un axe Δ passant par G est égale, par définition, à

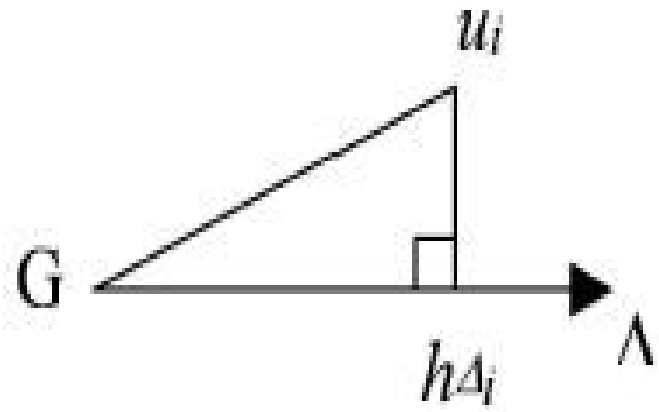
$$I_\Delta = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta_i}, u_i),$$

où h_{Δ_i} est la projection orthogonale de u_i sur l'axe Δ.

Le moment d'inertie

Inertie du nuage des individus par rapport à un axe passant par G

Cette inertie mesure la proximité à l'axe Δ du nuage des individus.



Le moment d'inertie

Inertie du nuage des individus par rapport à un sous-espace vectoriel

L'inertie du nuage par rapport à un sous-espace vectoriel $\mathbf{W}$ passant par G est, par définition, égale à

$$I_\Delta = \frac{1}{n} \sum_{i=1}^{n} d^2(h_{\mathbf{W}i}, u_i),$$

où $h_{\mathbf{W}i}$ est la projection orthogonale de u_i sur le sous-espace $\mathbf{W}$.

Le moment d'inertie

Décomposition de l'inertie totale

Si on note $\mathbf{W}^*$ le complémentaire orthogonal de $\mathbf{W}$ dans $\mathbb{R}^p$ et $h_{\mathbf{W}_i^*}$ la projection orthogonale de u_i sur $\mathbf{W}^*$, en appliquant le théorème de Pythagore, on peut écrire

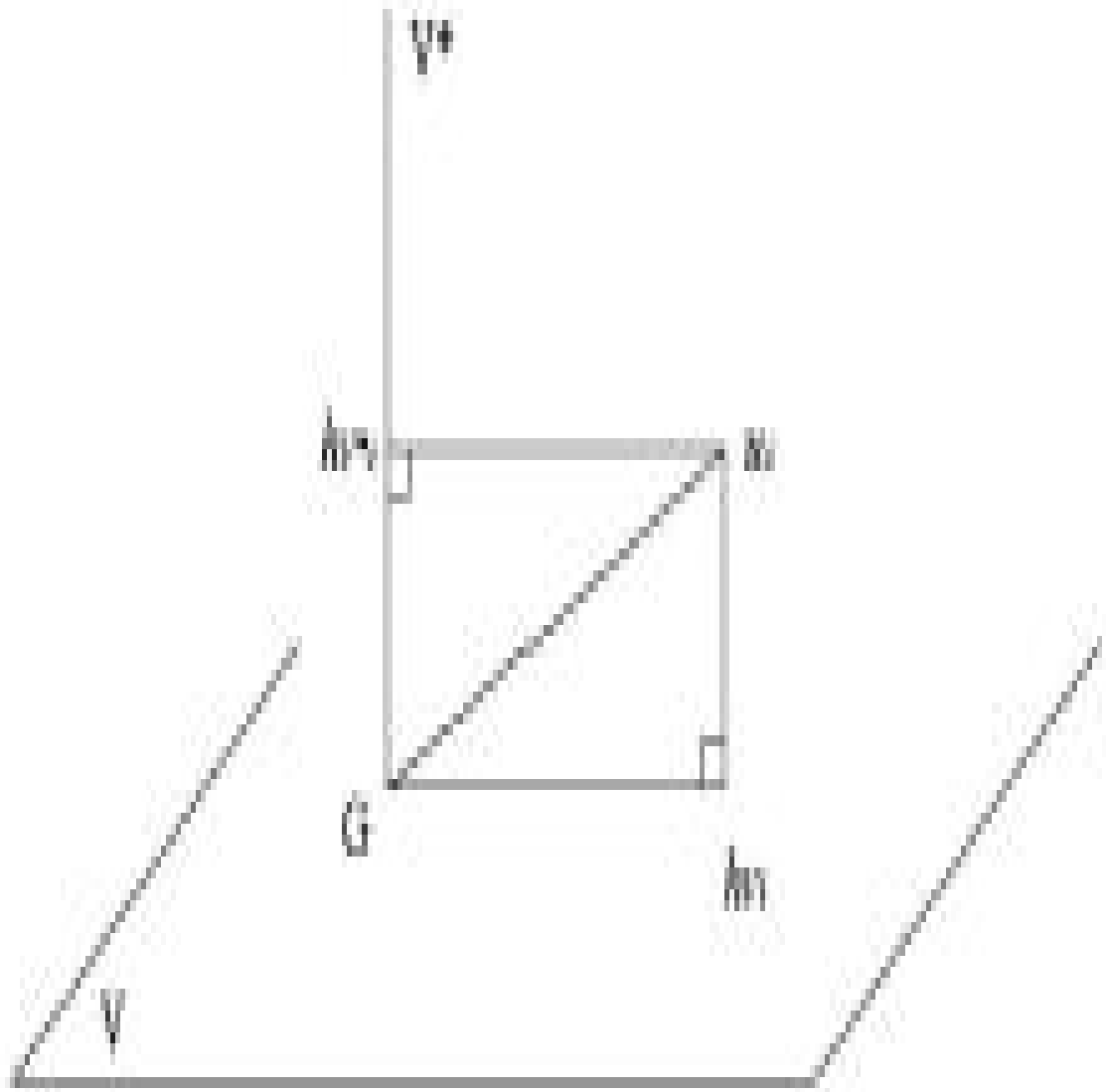
$$\begin{aligned} d^2(h_{\mathbf{W}_i}, u_i) + d^2(h_{\mathbf{W}_i^*}, u_i) &= d^2(G, u_i) \\ &= d^2(G, h_{\mathbf{W}_i}) + d^2(G, h_{\mathbf{W}_i^*}). \end{aligned}$$

On en déduit, par le théorème de Huygens, que

$$I_{\mathbf{W}} + I_{\mathbf{W}^*} = I_G$$

Le moment d'inertie

Décomposition de l'inertie totale



Le moment d'inertie

Décomposition de l'inertie totale

En projetant le nuage des individus sur un sous-espace $\mathbf{W}$, on perd l'inertie mesurée par $I_{\mathbf{W}}$, on ne conserve que celle mesurée par $I_{\mathbf{W}^*}$. De plus, si on décompose l'espace $\mathbb{R}^p$ comme la somme des sous-espaces de dimension 1 et orthogonaux entre eux

$$\Delta_1 \oplus \Delta_2 \oplus \cdots \oplus \Delta_p.$$

On peut alors écrire

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + \cdots + I_{\Delta_p^*}.$$

Plans du cours

Recherche des axes

Le premier axe

On cherche, d'abord, un axe Δ_1 passant par G d'inertie I_{Δ_1} minimum car c'est l'axe le plus proche de l'ensemble des points du nuage des individus, et donc, si l'on doit projeter ce nuage sur cet axe, c'est lui qui donnera l'image la moins déformée du nuage. Si on utilise la relation entre les inerties donnée au paragraphe précédent, rechercher Δ_1 tel que I_{Δ_1} est minimum, est équivalent à chercher Δ_1 tel que $I_{\Delta_1^*}$ est maximum.

$$I_{\Delta_1} \text{ est minimum } \iff I_{\Delta_1^*} \text{ est maximum.}$$

Recherche des axes

Le premier axe

On définit l'axe Δ_1 par son vecteur directeur unitaire $\overrightarrow{Ga_1}$. Il faut donc trouver $\overrightarrow{Ga_1}$ tel que $I_{\Delta_1^*}$ est maximum sous la contrainte que $\|\overrightarrow{Ga_1}\|^2 = 1$.
Les expressions algébriques de $I_{\Delta_1^*}$ et de $\|\overrightarrow{Ga_1}\|^2$ sont données par

$$d^2(G, h_{\Delta_{1i}^*}) = \langle \overrightarrow{Gu_i}, \overrightarrow{Ga_1} \rangle^2 = a_1^t u_{ci} u_{ci}^t a_1 .$$

Recherche des axes

Le premier axe

En utilisant la symétrie du produit scalaire, on en déduit

$$I_{\Delta_1^*} = \frac{1}{n}\sum_{i=1}^{n} a_1^t u_{ci} u_{ci}^t a_1 = a_1^t \left[\frac{1}{n}\sum_{i=1}^{n} u_{ci} u_{ci}^t\right] a_1 .$$

L'expression entre crochets est la matrice de covariance empirique Σ des p variables, donc

$$I_{\Delta_1^*} = a_1^t \Sigma a_1 .$$

Et

$$\|\overrightarrow{Ga_1}\|^2 = a_1^t a_1 .$$

Recherche des axes

Le premier axe

Notre problème est de trouver a_1 tel que $a_1^t \Sigma a_1$ soit maximum avec la contrainte $a_1^t a_1 = 1$ et c'est un problème de la recherche d'un optimum d'une fonction de plusieurs variables liées par une contrainte (les inconnues sont les composantes de a_1). La méthode des multiplicateurs de Lagrange peut, alors, être utilisée.
Dans le cas de la recherche de a_1, il faut calculer les dérivées partielles de

$$g(a_1) = g(a_{11}, a_{12}, \ldots, a_{1p}) = a_1^t \Sigma a_1 - \lambda_1 (a_1^t a_1 - 1).$$

Recherche des axes

Le premier axe

En utilisant la dérivée matricielle, on obtient

$$\frac{\partial g(a_1)}{\partial a_1} = 2\Sigma a_1 - 2\lambda_1 a_1 = 0.$$

Le système à résoudre est donc

$$\begin{cases} \Sigma a_1 - \lambda_1 a_1 = 0 & \dots \quad (1) \\ a_1^t a_1 - 1 = 0 & \dots \quad (2) \end{cases}$$

Recherche des axes

Le premier axe

De l'équation matricielle (1), de ce système, on déduit que a_1 est un vecteur propre de la matrice Σ associé à la valeur propre λ_1. En multipliant les deux côtés de l'expression (1), à gauche, par a_1^t, on obtient

$$a_1^t \Sigma a_1 - \lambda_1 (a_1^t a_1 = 1) = 0.$$

En utilisant l'équation (2), on trouve que

$$a_1^t \Sigma a_1 = \lambda_1.$$

Le premier membre de l'équation précédente est égal à l'inertie $I_{\Delta_1^*}$ qui doit être maximum. Cela signifie que la valeur propre λ_1 est la plus grande valeur propre de la matrice de covariance Σ et que cette valeur propre est égale à l'inertie portée par l'axe Δ_1. L'axe Δ_1, pour lequel le nuage des individus a une inertie minimum, a comme vecteur directeur unitaire le premier vecteur propre associé à la plus grande valeur propre de la matrice de covariance Σ.

Recherche des axes

Le deuxième axe

On cherche, ensuite, un deuxième axe Δ_2 orthogonal au premier et d'inertie minimum. On peut, comme dans le paragraphe précédent, définir l'axe Δ_2 passant par le centre de gravité G par son vecteur directeur unitaire a_2. L'inertie du nuage des individus par rapport à son complémentaire orthogonal est égale à

$$I_{\Delta_2^*} = a_2^t \Sigma a_2,$$

et elle doit être maximale avec les deux contraintes suivantes

$$a_2^t a_2 = 1 \quad \text{et} \quad a_2^t a_1 = 0.$$

Recherche des axes

Le deuxième axe

La deuxième contrainte exprime que le deuxième axe doit être orthogonal au premier et donc que le produit scalaire des deux vecteurs directeurs est nul.
En appliquant la méthode des multiplicateurs de Lagrange, cette fois avec deux contraintes, on trouve que a_2 est le vecteur propre de Σ correspondant à la deuxième plus grande valeur propre. On peut montrer que le plan défini par les axes Δ_1 et Δ_2 est le sous-espace de dimension 2 qui porte l'inertie maximum. On peut chercher de nouveaux les axes suivants par la même procédure.

Recherche des axes

Le deuxième axe

Les nouveaux axes sont tous vecteurs propres de Σ correspondant aux valeurs propres ordonnées. La matrice de covariance Σ étant une matrice symétrique réelle. Elle possède p vecteurs propres réels, formant une base orthogonale de $\mathbb{R}^p$

$$\begin{cases} \Delta_1 & \perp & \Delta_2 & \perp & \dots & \perp & \Delta_p \\ a_1 & \perp & a_2 & \perp & \dots & \perp & a_p \\ \lambda_1 & \geq & \lambda_2 & \geq & \dots & \geq & \lambda_p \\ \Delta_1^* & \geq & \Delta_2^* & \geq & \dots & \geq & \Delta_p^* \end{cases}.$$

On passera de la base orthogonale initiale des variables centrées à la nouvelle base orthogonale des vecteurs propres de Σ. On appelle les nouveaux axes, axes principaux.

Plans du cours

En utilisant le théorème de Huygens, on peut décomposer l'inertie totale du nuage des individus comme suit

$$I_G = I_{\Delta_1^*} + I_{\Delta_2^*} + \cdots + I_{\Delta_p^*} = \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$

La contribution absolue de l'axe Δ_k à l'inertie totale du nuage des individus est égale à la valeur propre qui lui est associé,

$$ca(\Delta_k / I_G) = \lambda_k.$$

Sa contribution relative est égale à

$$cr(\Delta_k / I_G) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}.$$

On emploie souvent l'expression "pourcentage d'inertie expliquée par Δ_k".

On peut étendre ces définitions à tous les sous-espaces engendrés par les nouveaux axes. Ainsi, le pourcentage d'inertie expliqué par le plan engendré par les deux premiers axes Δ_1 et Δ_2 est égal à

$$cr(\Delta_1 \oplus \Delta_2 / I_G) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}.$$

Ces pourcentages d'inertie sont des indicateurs qui rendent compte de la part de variabilité du nuage des individus expliquée par ces sous-espaces. Si les dernières valeurs propres ont des valeurs faibles, on pourra négliger la variabilité qu'expliquent les axes correspondants.

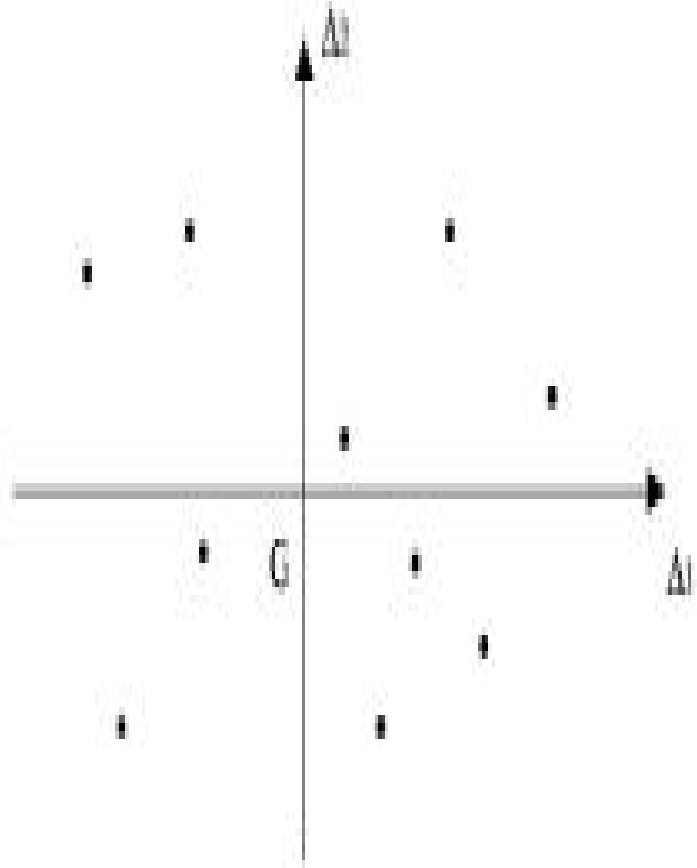
On se contente souvent de faire des représentations du nuage des individus dans un sous-espace engendré par les “d” premiers axes si ce sous-espace explique un pourcentage d’inertie proche de 1. On peut ainsi réduire l’analyse à un sous-espace de dimension $d < p$.

Plans du cours

Pour faire la représentation des individus dans les plans définis par les nouveaux axes, il suffit de calculer les coordonnées des individus dans les nouveaux axes. Pour obtenir y_{ik}, coordonnée de l'unité u_i sur l'axe Δ_k. On projette orthogonalement le vecteur $\overrightarrow{Gu_i}$ sur cet axe et on obtient

$$\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \rangle = a_k^t u_{ci} \quad et \quad y_i = A^t u_{ci},$$

où y_i est le vecteur des coordonnées de l'unité u_i et A est la matrice du changement de base. C'est la matrice des vecteurs propres orthogonaux et de norme 1. C' est une matrice orthogonale, son inverse est égale à sa transposée.


Δ2
G
Δ1

Qualité de la représentation des individus

Lorsque les points de projections des individus sont éloignés sur un axe (ou sur un plan), on peut assurer que les points représentants ces individus sont éloignés dans l'espace. En revanche, deux individus dont les projections sont proches sur un axe (ou sur un plan) peuvent ne pas être proches dans l'espace.

Qualité de la représentation des individus

Pour interpréter correctement la proximité des projections de deux individus sur un plan, il faut donc s'assurer que ces individus sont bien représentés dans le plan. Pour que l'individu lui même soit bien représenté sur un axe (ou sur un plan, ou un sous-espace), il faut que l'angle entre le vecteur $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous-espace) soit petit. On calcule donc le cosinus de cet angle, ou plutôt le carré de ce cosinus.

Qualité de la représentation des individus

En effet, en utilisant le théorème de Pythagore, on peut montrer que le carré du cosinus de l'angle d'un vecteur avec un plan engendré par deux vecteurs orthogonaux, est égal à la somme des carrés des cosinus des angles du vecteur avec chacun des deux vecteurs qui engendrent le plan. Cette propriété se généralise à l'angle d'un vecteur avec un sous-espace de dimension k quelconque.

Qualité de la représentation des individus

Si le carré du cosinus de l'angle entre $\overrightarrow{Gu_i}$ et l'axe (ou le plan, ou le sous-espace) est proche de 1, alors on pourra dire que l'individu est bien représenté par sa projection sur l'axe (ou le plan, ou le sous-espace). Et si deux individus sont bien représentés en projection sur un axe (ou un plan, ou un sous-espace) et ont des projections proches, alors on pourra dire que ces deux individus sont proches dans l'espace.

Qualité de la représentation des individus

Le carré du cosinus de l'angle α_{ik} entre $\overrightarrow{Gu_i}$ et un axe Δ_k du vecteur directeur unitaire a_k est égal à

$$\begin{aligned}\cos^2(\alpha_{ik}) &= \frac{\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k} \rangle^2}{\| \overrightarrow{Gu_i} \|^2} \\ &= \frac{a_k^t u_{ci}^t u_{ci} a_k}{u_{ci}^t u_{ci}} \\ &= \frac{\left[\sum_{j=1}^{p} (x_{ij} - x_{.j}) a_{kj} \right]^2}{\sum_{i=1}^{p} (x_{ij} - x_{.j})^2}.\end{aligned}$$

Qualité de la représentation des individus

En utilisant le théorème de Pythagore on peut calculer le carré du cosinus de l'angle $\alpha_{ikk'}$ entre $\overrightarrow{Gu_i}$ et le plan engendré par deux axes $\Delta_k \oplus \Delta_{k'}$

$$\cos^2(\alpha_{ikk'}) = \cos^2(\alpha_{ik}) + \cos^2(\alpha_{ik'}).$$

Après l'étude, des pourcentages d'inertie expliqués par les sous-espaces successifs engendrés par les nouveaux axes, sont calculés. on ne peut retenir qu'un sous-espace de dimension $d < p$. On pourra calculer la qualité de la représentation d'un individu en calculant le carré du cosinus de l'angle de $\overrightarrow{Gu_i}$ avec ce sous-espace.

Qualité de la représentation des individus

Si un individu est très proche du centre de gravité dans cet espace, c'est-à-dire si $\|\overrightarrow{Gu_i}\|^2$ est très petit, le point représentant cet individu sur un axe (ou un plan, ou un sous-espace) sera bien représenté.

Qualité de la représentation des individus

Interprétation des nouveaux axes en fonction des individus

Lorsqu'on calcule l'inertie $I_{\Delta_k^*}$ portée par l'axe Δ_k, on peut voir quelle est la part de cette inertie due à un individu u_i particulier.

Qualité de la représentation des individus

Contribution absolue d'un individu à un axe

Nous savons, comme il est énoncé précédemment, que l'inertie portée par l'axe Δ_k^* est

$$I_{\Delta_k^*} = \frac{1}{n}\sum_{i=1}^{n} d^2(h_{\Delta_{ki}}, G),$$

alors la contribution absolue de l'individu u_i est égale à

$$ca(u_i/_{\Delta_k}) = \frac{1}{n} d^2(h_{\Delta_{ki}}, G).$$

Qualité de la représentation des individus

Contribution absolue d'un individu à un axe

puisque tous les individus ont le même poids, alors chacun d'eux contribuera autant plus à la confection d'un axe, que sa projection sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribuera faiblement à l'inertie portée par cette axe. On se sert de ces contributions pour interpréter les nouveaux axes de l'ACP en fonction des individus.

Qualité de la représentation des individus

Contribution relative d'un individu à un axe

Pour un individu particulier u_i, La contribution relative à l'inertie portée par un axe Δ_k est

$$\begin{aligned} cr(u_i/_{\Delta_k}) &= \frac{\frac{1}{n}d^2(h_{\Delta_{ki}}, G)}{I_{\Delta_k^*}} \\ &= \frac{\frac{1}{n}\langle \overrightarrow{Gu_i}, \overrightarrow{Ga_k}\rangle^2}{\lambda_k} \\ &= \frac{\frac{1}{n}a_k^t u_{ci} u_{ci}^t a_k}{\lambda_k}. \end{aligned}$$

L'examen de ces contribution permet d'interpréter les axes principaux avec les individus. On peut remarquer que $\sum_{i=1}^{n} cr(u_i/\Delta_k) = 1$.

Représentation des variables

On peut envisager le problème de la représentation des variables de façon complètement symétrique de celle des individus. Les raisonnements se font dans $\mathbb{R}^n$ au lieu de $\mathbb{R}^p$. Mais dans l'ACP, au delà de la symétrie formelle entre les individus et les variables, on peut utiliser la dissymétrie liée à la sémantique: les variables n'ont pas la même signification que les individus. On peut alors faire le raisonnement suivant

Représentation des variables

On représente les individus dans l'espace des anciennes variables, et on fait un changement de base dans cet espace. Les nouveaux axes sont des combinaisons linéaires des anciens axes et peuvent donc être considérés comme de nouvelles variables combinaisons linéaires des anciennes. On appelle communément ces nouvelles variables" composantes principales".

Représentation des variables

On note $Z_1, Z_2, \ldots, Z_k, \ldots, Z_p$ les composantes principales, Z_k étant la nouvelle variable correspondante à l'axe Δ_k,

$$Z_k = \sum_i^p a_{kj} V_{cj} = X_c a_k,$$

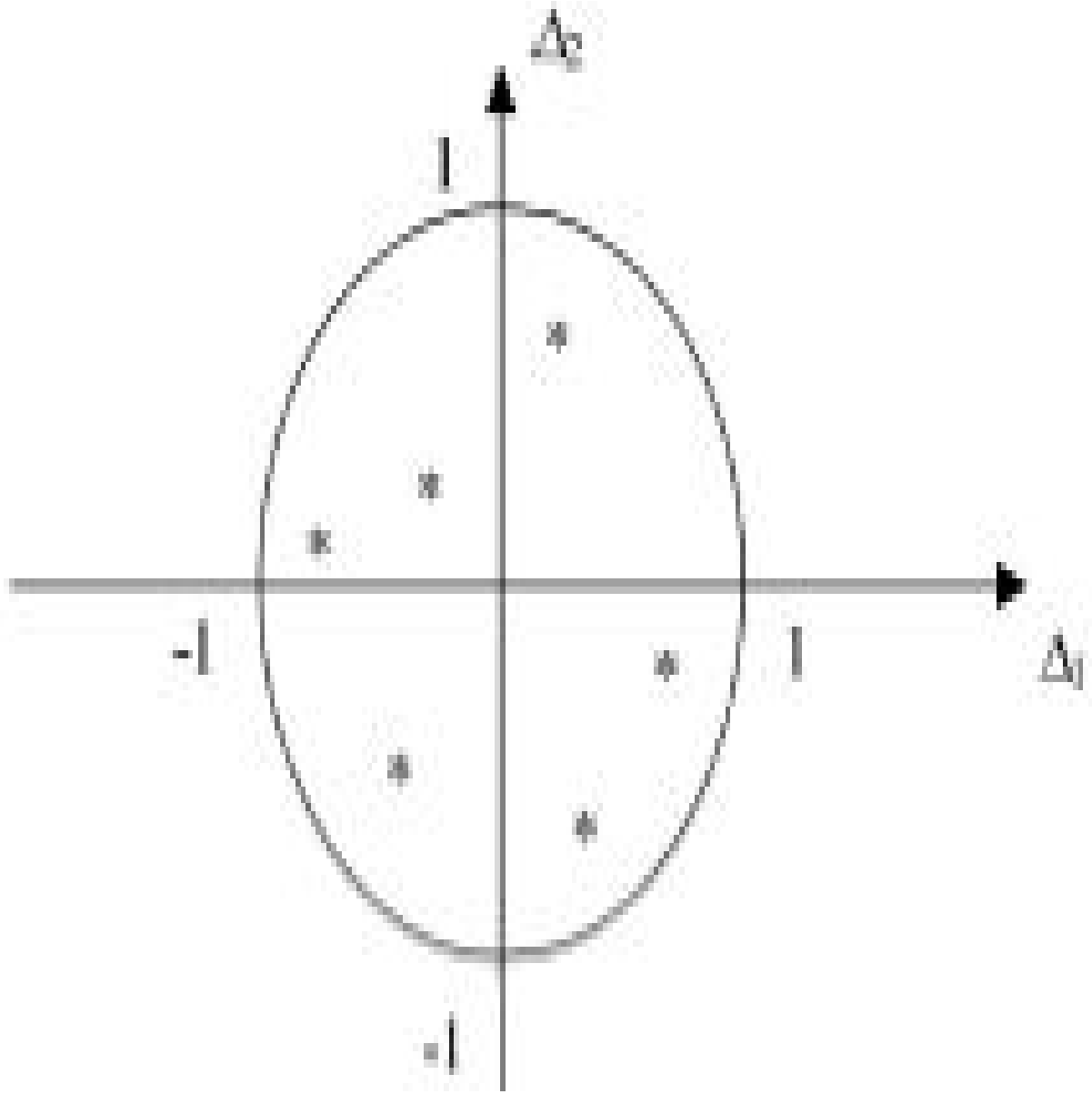
et de façon générale

$$Z = [Z_1, Z_2, \ldots, Z_k, \ldots, Z_p] = X_c A.$$

Représentation des variables

Il est alors intéressant de voir comment les anciennes variables sont liées aux nouvelles, et pour cela on calcule les corrélations des anciennes variables avec les nouvelles. La représentation des anciennes variables se fera en prenant comme coordonnées des anciennes variables, leurs coefficients de corrélation avec les nouvelles variables. On obtient alors ce qu'on appelle communément le "cercle des corrélations", dénomination qui vient du fait qu'un coefficient de corrélation variant entre -1 et $+1$. Les représentations des variables de départ sont des points qui se trouvent à l'intérieur d'un cercle de rayon 1,c'est on fait la représentation sur un plan.

Représentation des variables



Représentation des variables

On peut montrer que les variances; les covariances, et les coefficients de corrélation empiriques des composantes principales entre elles ou avec les variables de départ sont

$$Var(Z_k) = \frac{1}{n} a_k^t X_c^t X_c a_k = a_k^t \Sigma a_k = \lambda_k.$$

Représentation des variables

Et

$$Cov(Z_k, V_{cj}) = \frac{1}{n} a_k^t X_c^t V_{cj}$$

$$= \frac{1}{n} a_k^t X_c^t X_c \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = a_k^t \Sigma \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

$$= \lambda_k a_k^t \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \lambda_k a_k.$$

Représentation des variables

Enfin, on trouve

$$Cor(Z_k, V) = \sqrt{\lambda_k} \frac{a_{kj}}{\sqrt{Var(V_j)}}$$

où a_{kj} est la j ème coordonnée du vecteur directeur unitaire a_k de Δ_k.

Représentation des variables

De façon générale, la matrice de covariances des composantes principales est égale à Σ_Z,

$$\Sigma_Z = \frac{1}{n} A^t X_c^t X_c^t A = A^t \Sigma A = \Lambda,$$

où Λ est la matrice diagonale des valeurs propres de $\sum$

$$\Lambda = \begin{pmatrix} \lambda_1 & \dots & (0) \\ \vdots & \ddots & \vdots \\ (0) & \dots & \lambda_p \end{pmatrix},$$

Représentation des variables

et la matrice des covariance entre les composantes principales et les anciennes variables vaut

$$Cov(Z, V) = \frac{1}{n} X_c^t X_c A = \Sigma A = A\Lambda.$$

Si on remarque que la variance empirique d'une variable est égale au carré de la norme du vecteur qui la représente dans la géométrie euclidienne choisie et que le coefficient de corrélation empirique de deux variables est égal au produit scalaire des deux vecteur qui les représentent, on pourra interpréter les angles des vecteurs comme des corrélations.

Représentation des variables

Interprétation des nouveaux axes en fonction des anciennes variables

On peut interpréter les axes principaux en fonction des anciennes variables. Une ancienne variable V_j expliquera d'autant mieux un axe principal qu'elle sera fortement corrélée avec la composante principale correspondant à cet axe.

Représentation des variables

Qualité de la représentation des variables

Pour les mêmes raisons qui ont poussé à se préoccuper de la représentation des individus, il faut se préoccuper de la qualité de la représentation des variables sur un axe, un plan ou un sous-espace. Une variable sera d'autant mieux représentée sur un axe que sa corrélation avec la composante principale correspondante est en valeur absolue proche de 1. Le coefficient de corrélation empirique entre une ancienne variable V_{cj} et une nouvelle variable Z_k n'est autre que le cosinus de l'angle du vecteur joignant l'origine au point V_j représentant la variable sur l'axe avec ce vecteur.

Représentation des variables

Qualité de la représentation des variables

Une variable sera bien représentée sur un plan, si elle est proche du bord du cercle des corrélation, car cela signifie que le cosinus de l'angle du vecteur joignant l'origine au point représentant la variable avec le plan est en valeur absolue, proche de 1, ... etc.

Représentation des variables

Étude des liaisons entre les variables

Sur le graphique du cercle des corrélations, on peut aussi interpréter les positions des anciennes variables les un par rapport aux autres en termes de corrélation. Deux points très proches du cercle des corrélations, donc bien représentés dans le plan, seront très corrélées positivement entre elles. Si elles sont proches du cercle, mais dans des positions symétriques par rapport à l'origine, elles seront très corrélées négativement.

Représentation des variables

Étude des liaisons entre les variables

Deux variables proche du cercle des corrélation et dont les vecteurs qui les joignent à l'origine forment un angle droit, ne seront pas corrélées entre elles. Il faut, pour interpréter correctement ces graphiques des cercles de corrélation, se souvenir qu'un coefficient de corrélation est une mesure de liaison linéaire entre deux variables, et qu'il peut arriver que deux variables très fortement liées ont un coefficient de corrélation nul ou très faible, si leur liaison n'est pas linéaire.

Plans du cours

Analyse en composantes principales normée

Dans les paragraphes précédents, nous avons étudié l'ACP simple, pour laquelle, non seulement tous les individus ont le même poids dans l'analyse, mais aussi, toutes les variables sont traitées de façon symétrique (on leur fait jouer le même rôle) et les nouveaux axes sont issus de la matrice de covariance empirique des variables.

Analyse en composantes principales normée

Le Cercle pose parfois des problèmes. Le premier reproche fait par des praticiens est que, si les anciennes variables sont hétérogènes, comme par exemple des poids, des tailles et des ages, quel sens peut-on donner aux composantes principales qui sont alors des combinaisons linéaires de variables hétéroclites? Le deuxième reproche, est que, si on change d'unités sur ces variables, on peut changer complètement les résultats de l'ACP. Le dernier reproche vient du fait qu'une variable contribuera d'autant plus à la confection des premiers axes, que sa variance est forte.

Analyse en composantes principales normée

Pour échapper à tous ces problèmes, on cherchera à normaliser les variables et à travailler sur des variables sans dimension. Il y a plusieurs façons de normaliser les variables, mais la plus couramment utilisée est celle qui consiste à diviser les valeurs des variables par leur écart-type, c'est-à-dire que l'on travaille sur des variables centrées et réduites.

Analyse en composantes principales normée

Cela revient à faire la même analyse que pour l'ACP simple, mais à choisir une autre distance euclidienne entre les individus que la distance euclidienne classique. La distance choisie est

$$d^2(u_i, u_{i\prime}) = \sum_{j=1}^{p} \frac{1}{\sigma_j^2}(x_{ij} - x_{i'j})^2.$$

Cette nouvelle distance ne traite plus les variables de façon symétrique, mais elle permet de faire jouer un rôle plus équitable à chacune d'entre elles.

Analyse en composantes principales normée

Si on reprend tous les calculs de l'ACP simple, mais en remplaçant les variables de départ par les variables centrées réduites, on voit que ce n'est plus la matrice de covariance, mais la matrice de corrélation $\mathbb{R}$ qui intervient pour la recherche des nouveaux axe. Les particularités de l'ACP normée par rapport à l'ACP simple proviennent du fait que la matrice de corrélation $\mathbb{R}$ n'a que des 1 sur sa diagonale principale. Cela entraîne que sa trace est toujours égale à p.

Analyse en composantes principales normée

On a vu que la trace de la matrice est égale à l'inertie totale du nuage calculée avec la distance euclidienne que l'on a choisie. L'inertie totale du nuage des individus dans $\mathbb{R}^p$ est donc toujours égale à p dans toute l'ACP normée. Cette particularité donne une règle supplémentaire pour choisir le nombre d'axes que l'on va garder pour les interprétations, fondée sur le raisonnement suivant:

Analyse en composantes principales normée

on a p valeurs propres dont la somme vaut p (puisque l'on a vu que l'inertie totale est aussi égale à la somme des valeurs propres); on peut ne considérer comme significatives que les valeurs propres supérieure à 1, puisque la valeur moyenne des valeurs propres vaut 1 et leur somme vaut p. C'est bien sur une règle empirique mais qui peut servir de guide pour le choix de la dimension du sous-espace que l'on veut garder.

Analyse en composantes principales normée

Une autre particularité de l'ACP normée est que la représentation des variables avec le cercle de corrélation corresponde exactement à la représentation des variables dans $\mathbb{R}^n$ que l'on aurait construite si l'on avait adopté la même démarche que celle qui a servi pour la représentation des individus dans $\mathbb{R}^p$.

Plans du cours

Dans la plus part des cas, notre intérêt se porte sur la liaisons linéaire de certaines variables et non pas sur toutes les informations disponibles (variable moyenne générale des notes des étudiants par exemple par rapport aux résultats) ou bien des variables qualitatives qui ne peuvent être utiliser dans une ACP.
Les variables actives sont celles utilisées dans l'ACP.
Les variables supplémentaires ne participent pas aux calcules des axes factorielles mais qui seront représentées à la fin à l'aide de la matrice A.