Analyse des Données (1)

Introduction à l'analyse des données

Le but de l'analyse de données est de synthétiser, structurer l'information contenue dans des données multidimensionnelles.
Deux groupes de méthodes :
– méthodes de classification : réduire la taille de l'ensemble des individus en formant des groupes homogènes ;
– méthodes factorielles : réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques. Dans ce cours, nous verrrons essentiellement l'ACP, Analyse en Composantes Principales, dans le cas oú les variables sont quantitatives. Il existe d'autres méthodes factorielles :
– l'AC, Analyse des Correspondances si les variables sont qualitatives, où on cherchera les liens entre les modalités, – l'AFC Analyse Factorielle des Correspondances (simples) dans le cas oú on dispose de 2 variables – l'ACM Analyse des Correspondances Multiples dans le cas oú on dispose de plus de 2 variables. Une bonne référence : Jean-Marie Bouroche et Gilbert Saporta, *L'analyse des données*, Que Sais-je ?, Presses Universitaires de France, 2002.

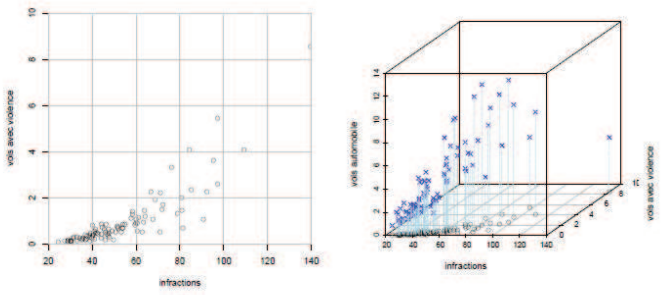
Exemple, ville et (in)sécurité

Le palmarès des départements : où vit-on en securité? dans L'Express (no 2589, 15 fevrier 2001)
infra: Nombre d'infractions totale pour 1000 habitants (2000)
vvi: Nombre de vols avec violence pour 1000 habitants (2000)
auto: Nombre de vols d'automobiles pour 1000 habitants (2000)

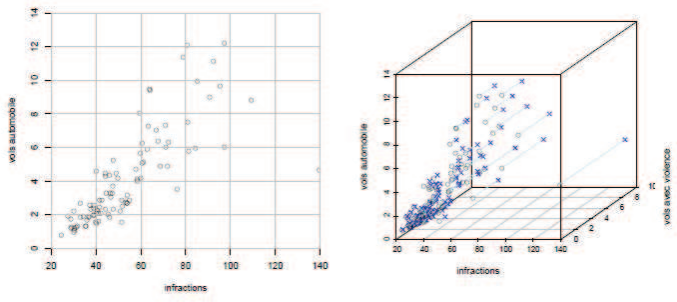
	infra	vvi	auto
D1	44.11	0.27	4.47
D2	45.97	0.55	4.39
D3	38.83	0.41	2.39
D4	49.68	0.21	4.17
D5	47.67	0.33	2.35
D6	109.21	4.10	8.83
...	...	...	...

```
> donnees=read.table("http://www.univ-montp3.fr/
                     miap/ens/site/uploads/Misashs.AD/securite.txt",header=TRUE)
> base=donnees[,2:ncol(donnees)]
> rownames(base)=donnees$dep
> base=base[,c(1,6,9)]
> print(head(base))
```

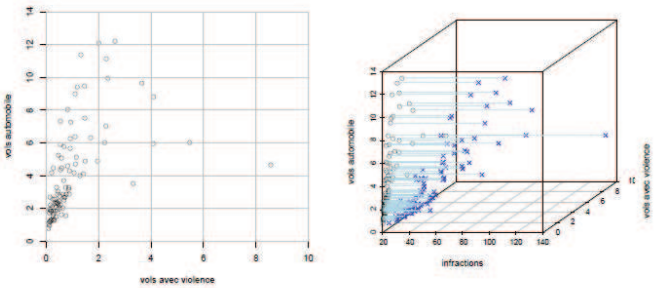

Exemple, ville et (in)sécurité



Exemple, ville et (in)sécurité



Exemple, ville et (in)sécurité



Exemple, ville et (in)sécurité

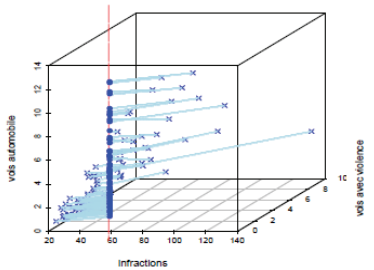
Les variables semblent plutôt correlées positivement, comme on peut le voir sur la matrice des corrélations :

	infra	vvi	auto
infra	1.000	0.858	0.781
vvi	0.858	1.000	0.503
auto	0.781	0.503	1.000

Supposons que l'on cherche à regrouper les villes "proches". Comme on a du mal à voir dans $\mathbb{R}^3$, on va essayer de projeter le nuage.

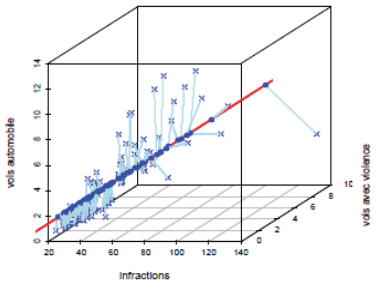
- projection sur un axe (droite)
- projection sur un plan

Exemple, ville et (in)sécurité



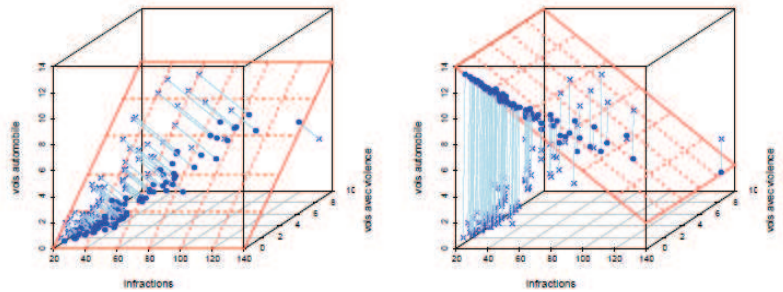
Exemple, ville et (in)sécurité

Recherche de la projection la plus représentative : idée des moindres carrés, qui minimise l'erreur de projection.



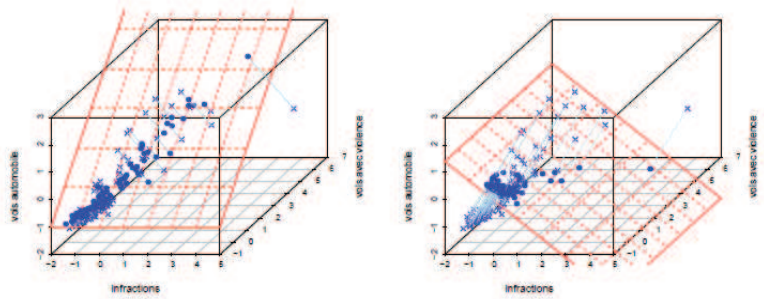
Exemple, ville et (in)sécurité

On peut aussi essayer de projeter sur un plan :



Exemple, ville et (in)sécurité

On peut aussi vouloir "normer" les axes afin de les rendre comparables :



Les données

On observe n individus, et q variables (quantitatives, sur $\mathbb{R}$). Le nuage de points peut se décomposer de deux manières,

- l'espace des individus, i.e. $\mathbb{R}^q$
- l'espace des variables, i.e. $\mathbb{R}^n$

On note x_{ij} l'observation de la jème variable sur le ième individu. On a donc le tableau de données

		variables				
		1	$\cdots$	j	$\cdots$	q
individus	1	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1q}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{iq}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	n	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{nq}

Le tableau de données $X = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq q}$ est une matrice rectangulaire de taille $n \times q$.

Espace des individus

Chaque individu est caractérisé par une ligne $\ell_i = (x_{i1}, \cdots, x_{iq})$, appartenant à $\mathbb{R}^q$, exprimé dans la base canonique.

Définition

L'ensemble des individus dans l'espace vectoriel $\mathbb{R}^q$, muni de la base canonique est appelé espace des individus.

Chaque individu est considéré comme un élément de l'espace vectoriel $\mathbb{R}^q$ de dimension q. L'ensemble des individus est un nuage de points dans $\mathbb{R}^q$.

Principe :
on cherche à réduire le nombre q de variables tout en préservant au maximum la structure du problème.
A cette fin, on projette le nuage de points sur un ou des sous-espaces de dimension inférieure à q. Dans la pratique, on projette sur des plans (dimension 2).

Espace des variables

De la même manière, chaque variable est caractérisée par $x^j = (x_{1j}, \cdots, x_{nj})^t$, appartenant à $\mathbb{R}^n$, exprimé dans la base canonique. On donne un poids p_i à chaque individu avec $p_1 + p_2 + \cdots + p_n = 1$ et on représente ces poids à l'aide de la matrice diagonale :

$$D = \begin{pmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{pmatrix}$$

Dans l'espace des variables, en général, un poids identique sera donné à chaque individu et donc :

$$D = \frac{1}{n} Id_n$$

Point moyen et tableau centré

Point moyen : c'est le vecteur g des moyennes arithmétiques de chaque variable :

$$g = (\bar{x}^1, \cdots, \bar{x}^q),$$

où $\bar{x}^j = \sum_{i=1}^n p_i x_{ij} = \sum_{i=1}^n \frac{1}{n} x_{ij}$
On peut aussi écrire :

$$g = \mathbf{1}_n DX.$$

Tableau centré : $Y = (y_{ij})_{1 \leq i \leq n, 1 \leq j \leq q}$ est obtenu en centrant les variables autour de leur moyenne :

$$y_{ij} = x_{ij} - \bar{x}^j$$

ou, en notation matricielle,

$$Y = X - \mathbf{1}_n^t g$$

N.B : $\mathbf{1}_n =$ est le vecteur ligne de dimension n dont chaque coordonnée est 1.

Matrice de variance-covariance

C'est une matrice carrée de dimension q

$$V = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1q} \\ s_{21} & & & \\ \vdots & & \ddots & \\ s_{q1} & & & s_q^2 \end{pmatrix}$$

où

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

est la variance de la variable x^j et

$$s_{k\ell} = \sum_{i=1}^n p_i (x_{ik} - \bar{x}^k)(x_{i\ell} - \bar{x}^\ell) = \sum_{i=1}^n p_i x_{ik} x_{i\ell} - \bar{x}^k \bar{x}^\ell$$

est la covariance des variables x^k et x^ℓ.
Formule matricielle : $V = X^t DX - gg^t = Y^t DY$.

Matrice de corrélation

Si l'on note $r_{k\ell} = s_{k\ell}/s_k s_\ell$, c'est la matrice de dimension q :

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1q} \\ r_{21} & 1 & & \\ \vdots & & \ddots & \\ r_{q1} & & & 1 \end{pmatrix}$$

Formule matricielle : $R = D_{1/s} V D_{1/s}$, où

$$D_{1/s} = \begin{pmatrix} s_1^{-1} & & 0 \\ & \ddots & \\ 0 & & s_q^{-1} \end{pmatrix}$$

Distance entre individus

Comment mesurer la distance entre deux individus ?

Définition

Soit M une matrice diagonale $q \times q$, dont les éléments diagonaux sont strictement positifs ($m_{jj} > 0$ pour $j = 1, \cdots, q$). Alors la fonction $\varphi : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ définie par

$$(u, v) \mapsto u^t M v = \sum_{j=1}^{q} m_{jj} u_j v_j$$

est un produit scalaire, noté $\langle .; . \rangle_M$.

Définition

Soit M une telle matrice diagonale $q \times q$, et $\langle .; . \rangle_M$ le produit scalaire associé. On note alors $\|.\|_M$ la norme associée,

$$\|u\|_M = \sqrt{\langle u; u \rangle_M} = \sqrt{\sum_{j=1}^{q} m_{jj} u_j^2}$$

et $d_M(.; .)$ la distance associée,

$$d_M(u; v) = \|u - v\|_M.$$

Exemples de produits scalaires

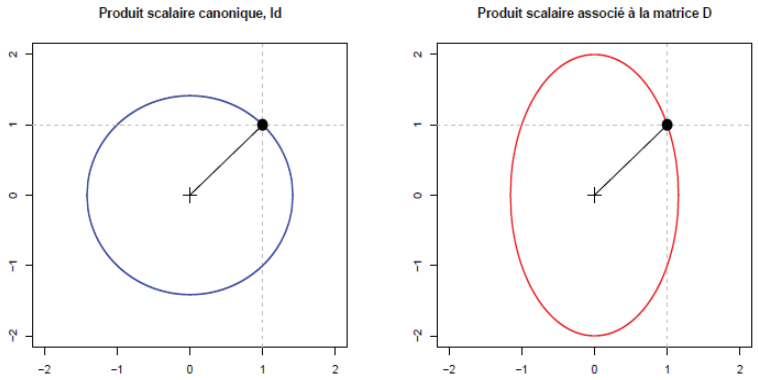
- Dans la suite, on prendra $M = Id_q$.

  Cela correspond au produit scalaire canonique, $\langle u; v \rangle = \sum_{j=1}^{q} u_j v_j$
- Soit $M = \begin{pmatrix} 1/4 & 0 \\ 0 & 3/4 \end{pmatrix}$ et $\langle u; v \rangle_M$ le produit scalaire associé Les points à égale distance de l'origine 0 sont les points $P = (x; y) \in \mathbb{R}^2$ tels que :

  $$\|\vec{OP}\|_M = cste > 0 \ i.e \ \frac{1}{4}x^2 + \frac{3}{4}y^2 = cste.$$

  soit une ellipse de $\mathbb{R}^2$.

Déformation de l'espace

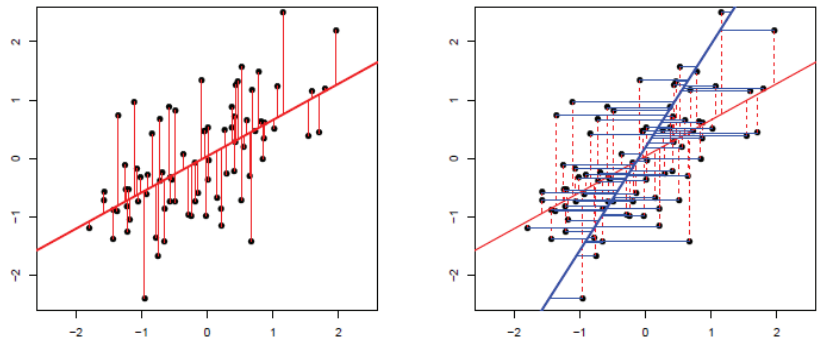


Les variables, cas de la dimension 2

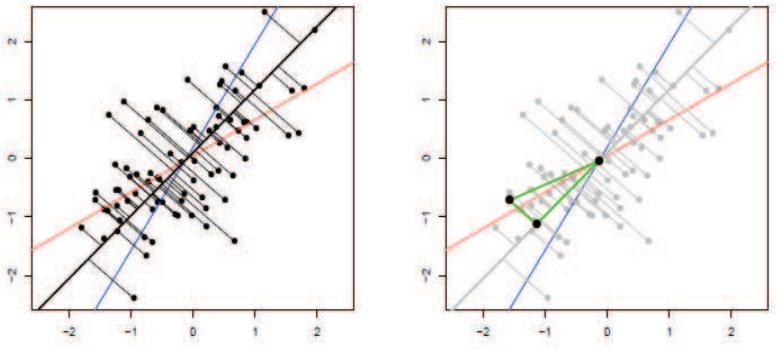
On cherche ici à mesurer une distance, ou une proximité, entre les variables. Intuitivement, cette notion doit être proche de la notion de corrélation.

Soient deux variables X_1 et X_2 continues.
Remarque: La régression propose d'étudier le lien entre deux variables, dans l'optique d'en utiliser une pour prévoir l'autre.

Droites de régression de y en x (rouge) et de x en y (bleue).



Ici, l'objectif est différent : on s'intéresse davantage à des projections (orthogonales). On parlera alors de direction principale du nuage. On peut montrer que cet axe passe par le centre de gravité du nuage (comme les deux autres régressions en rouge et bleu). Changeons les coordonnées pour simplifier, $Y_1 = X_1 - \bar{X}_1$ et $Y_2 = X_2 - \bar{X}_2$. On notera O ce barycentre, X les points d'origine et P les projections orthogonales.



Posons ℓ_i point de coordonnées (x_{i1}, x_{i2}) et O point moyen de coordonnées $(\bar{X}_1, \bar{X}_2)$.
On cherche la droite sur laquelle projeter le nuage de façon à minimiser la quantité :

$$\sum_{i=1}^{n} \|\ell_i \vec{P}_i\|^2 = \sum_{i=1}^{n} \|\vec{O\ell}_i\|^2 - \sum_{i=1}^{n} \|\vec{OP}_i\|^2$$

ce qui revient à maximiser la quantité $\sum_{i=1}^{n} \|\vec{OP}_i\|^2$ qu'on appellera l'inertie projetée du nuage.

Proposition

L'axe principal d'un nuage de points bivarié est le vecteur propre associé à la plus grande valeur propre de la matrice de variance-covariance des deux variables. (Ce résultat se généralise en plus grande dimension.)

Projeter un nuage de points

Projeter des points, la notion d'inertie

Considérons le tableau de données $X = (x_{ij})_{1 \le i \le n; 1 \le j \le q} = \{\ell_1, \cdots, \ell_n\}$.

Définition

On appelle inertie du nuage des points $\{\ell_1, \cdots, \ell_n\}$ la quantité

$$I(X) = \sum_{i=1}^{n} \frac{1}{n} \|\ell_i\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{q} \frac{1}{n} x_{ij}^2$$

$$I(X, M) = \sum_{i=1}^{n} p_i \|\ell_i\|_M^2 = \sum_{i=1}^{n} \sum_{j=1}^{q} p_i m_{jj} x_{ij}^2$$

L'inertie expliquée par un axe

Considérons le tableau de données $X = (x_{ij})_{1 \le i \le n; 1 \le j \le q} = \{\ell_1, \cdots, \ell_n\}$.

Définition

Soit $u \in \mathbb{R}^q$.
On appelle inertie du nuage des points $\{\ell_1, \cdots, \ell_n\}$ expliquée par l'axe dirigé par u la quantité $I(X; u)$ correspondant à l'inertie du nuage projeté orthogonalement sur cet axe.

D'après le théorème de Pythagore :

$$I(X) \ge I(X; u)$$

inertie totale $\ge$ inertie expliquée par un axe

Principe :
On cherche des axes ou des plans tels que l'inertie expliquée soit maximale.

L'inertie expliquée par un axe

Ecriture matricielle :

$$I(X) = \sum_{i=1}^{n} \frac{1}{n} \|\ell_i\|^2 = \frac{1}{n} X\, X^t$$

$$I(X;u) = \sum_{i=1}^{n} \frac{1}{n} \|pr(\ell_i)\|^2 = \frac{1}{n} (Xu)^t\, Xu = \frac{1}{n} u^t X^t\, Xu = u^t\, V\, u$$

Le (premier) axe principal

Définition

L'axe principal, ou premier axe principal, pour un nuage d'individus $\{\ell_1, \cdots, \ell_n\}$ est une droite dirigée par le vecteur unitaire $u \in \mathbb{R}^q$ qui maximise l'inertie $I(X;u)$

Proposition

L'axe principal est le sous-espace propre (droite) associé à la plus grande valeur propre λ_1 de la matrice de variance-covariance $V = \frac{1}{n} X^t X$.

Le (premier) axe principal

Corollaire

Il existe deux vecteurs unitaires $u \in \mathbb{R}^q$ maximisant $I(X;u)$ (i.e. au signe près). Notons u_1 l'un des deux. Alors u_1 est un vecteur propre associé à la plus grande valeur propre λ_1 de la matrice de variance-covariance V.

L'inertie expliquée par cet axe vaut alors λ_1 : $I(X;u_1) = \lambda_1$

Démonstration :

$$I(X;u_1) = u_1^t\, V\, u_1 \;= u_1^t \lambda_1 u_1 \;= \lambda_1 u_1^t u_1 \;= \lambda_1 < u_1, u_1 > = \lambda_1 \times 1 \quad \square$$

Les autres axes principaux

Les valeurs propres de la matrice variance-covariance V sont rangées par ordre décroissant et peuvent être considérées comme toutes différentes :

$$\lambda_1 > \lambda_2 > ..... > \lambda_q$$

Le 1èr axe principal est :

- un axe dirigé par u_1 unitaire
- u_1 est un vecteur propre associé à la première valeur propre λ_1 de la matrice V.
- $I(X;u_1) = \lambda_1$

Le 2ème axe principal est :

- un axe orthogonal à u_1 dirigé par u_2 unitaire
- u_2 est un vecteur propre associé à la seconde valeur propre λ_2 de la matrice V.
- $I(X;u_2) = \lambda_2$

...

Le kème axe principal est

- un axe orthogonal à $u_1, \cdots, u_{k-1}$ dirigé par u_k unitaire
- u_k est un vecteur propre associé à la $k$iéme valeur propre λ_k de la matrice V.
- $I(X;u_k) = \lambda_k$

Le (premier) plan principal

Considérons le tableau de données $X = (x_{ij})_{1 \leq i \leq n; 1 \leq j \leq q} = \{\ell_1, \cdots, \ell_n\}$. L'espace individus (de $\mathbb{R}^q$) est muni de la métrique issue de M.

Définition

Le plan principal, ou premier plan principal, pour un nuage d'individus $\{\ell_1, \cdots, \ell_n\}$ est le plan engendré par les vecteurs u_1 et u_2 qui constituent une base orthonormée de ce plan.

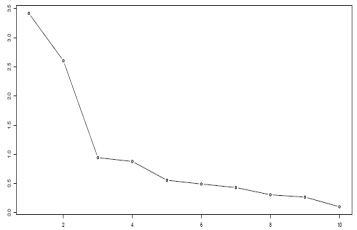
Rappel de la méthodologie

- on diagonalise V
- soient $\lambda_1 > \lambda_2 > \cdots > \lambda_q$ les valeurs propres
- les axes principaux sont dirigés par les vecteurs propres orthonormés associés

Combien d'axes principaux doit-on retenir ?

Rappelons que l'on cherche à résumer l'information apportée par les variables par un petit nombre de facteurs, en tenant compte des corrélations entre les variables.
On veut retenir peu d'axes principaux, avec

- un soucis d'interprétation : on ne garde que des axes que l'on puisse interpréter,
- des axes qui expliquent un bon pourcentage d'inertie. Pour cela, on a deux méthodes :
  – la règle de Kaiser, pour les variables centrées réduites : on ne garde que les valeurs propres supérieures à 1. (ce seuil de 1 correspond à la moyenne des valeurs propres).
  – la méthode du coude, correspondant à un décrochage au niveau des valeurs propres



Les composantes principales

On notera x pour les individus, y pour les individu centrés, et z pour les individus centrés réduits.
Les coordonnées d'un individu centré réduit $z_i^t = (z_{i1}, z_{i2}, \cdots, z_{iq}) \in \mathbb{R}^q$ sur un axe principal Δ_k dirigé par u_k sont obtenues par projection orthogonale:

$$c_{ij} = < z_i; u_k > = z_i^t u_k$$

Définition

On appellera composantes principales les nouvelles variables composites c_k, dans $\mathbb{R}^n$, définies par $c_k = Zu_k$

Il s'agit de nouvelles variables obtenues par combinaison linéaire des variables initiales.
Les coefficients sont les coordonnées de u_k.
On obtient ainsi les coordonnées des projections orthogonales sur les axes principaux.

Les composantes principales

On notera que, par construction,

$$\bar{c}_k = \frac{1}{n} \sum_{i=1}^{n} c_{ik} = 0$$

car les colonnes de Z sont centrées.
De plus, $\text{Var}(c_k) = \lambda_k$ et $\text{Cov}(c_1; c_2) = 0$, i.e. les composantes principales sont orthogonales.

Les données centrées réduites

Nous travaillerons avec les données centrées réduites car les distances entre variables sont très sensibles aux unités

Définition

On appellera nuage centré réduit le tableau $Z = (z_{ij})_{1 \leq i \leq n, 1 \leq j \leq q}$ *avec*

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Le cercle des corrélations

On rappelle que l'espace des variables est muni d'une métrique $D = (1/n)Id_n$.

$$\langle x; y \rangle_D = \frac{1}{n} \sum_{j=1}^{n} x_j y_j$$

On a donc pour x et y deux variables centrées :

$$s_x^2 = \mathrm{Var}(x) = \|x\|_D^2 \quad \text{et} \quad s_{xy} = \mathrm{cov}(x, y) = < x, y >_D$$

De plus, $r(x, y) = \frac{< x, y >_D}{\|x\|_D \|y\|_D}$.
Si les variables sont supposées centrées et réduites (Z), la corrélation entre une composante principale c_k et une variable z^j centrée, réduite est

$$r(z^j, c_k) = \frac{\mathrm{cov}(z^j, c_k)}{\sqrt{\mathrm{Var}(c_k)}} = \frac{\frac{1}{n}(z^j)^t c_k}{\sqrt{\lambda_k}}$$

donc le vecteur des corrélations du facteur c_k avec toutes les variables z^j est

$$r(Z, c_k) = \frac{\frac{1}{n} Z^t c_k}{\sqrt{\lambda_k}}$$

Contribution d'un individu à une composante

Définition

On sait que $Var(c_k) = \lambda_k = \sum_{i=1}^{n} \frac{1}{n} c_{ik}^2$.
La contribution de l'individu i à la composante k est donc :

$$\frac{\frac{1}{n} c_{ik}^2}{\lambda_k}$$

Interprétation : la contribution d'un individu est importante si elle excède d'un facteur α le poids $p_i = \frac{1}{n}$ de l'individu concerné, c'est-à-dire si :

$$\frac{\frac{1}{n} c_{ik}^2}{\lambda_k} \geq \frac{\alpha}{n} \Longrightarrow |c_{ik}| \geq \sqrt{\alpha \lambda_k}.$$

Selon les données, on se fixe une valeur de α de l'ordre de 2 à 4.

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. En pratique :

$$\frac{\frac{1}{n} c_{ik}^2}{\lambda_k} > 0,25$$

Interprétation :
il "tire trop à lui" l'axe k et risque de perturber les représentations des autres points.
Un tel individu peut être le signe de données erronées.
Solution :
on peut le retirer de l'analyse et le mettre en individu supplémentaire.

Qualité globale de la représentation

Calcul de l'inertie

On montre que $I(X) = Tr(V)$ où $Tr(V)$ est la trace de V i.e. la somme des termes diagonaux.

Proposition

La trace d'une matrice est égale à la somme de ses valeurs propres.

Proposition

Pour des données centrées réduites, $Tr(V) = q$

Proposition

$$I(X) = \lambda_1 + \lambda_2 + \cdots + \lambda_q = q.$$

Définition

La qualité de la représentation obtenue par k valeurs propres est la proportion de

Qualité de représentation d'un individu

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 Δ_k est :

$$\cos(z_i, u_k) = \frac{< z_i, u_k >}{\|z_i\|}$$

car les u_k forment une base orthonormale.
Comme $< z_i, u_k >= c_{ik}$,

$$\cos^2(z_i, u_k) = \frac{c_{ik}^2}{\sum_{k=1}^{q} c_{ik}^2}$$

Cette grandeur mesure la qualité de la représentation de l'individu i sur l'axe principal dirigé par u_k.

La qualité de la représentation de l'individu i sur le plan F_p est la somme des qualités de représentation sur les axes formant F_p.

Individus supplémentaires

On les ajoute à la représentation sur les plans principaux. Pour calculer leur coordonnée sur un axe fixé, on écrit :

$$\tilde{c}_k =< \tilde{z}, u_k >= \sum_{j=1}^{q} \tilde{z}_j u_{kj},$$

où les $\tilde{z}_j$ sont les coordonnées centrées-réduites d'un individu supplémentaire $\tilde{z}$. Ces individus serviront d'échantillon-test pour vérifier les hypothèses tirées de l'ACP sur les individus actifs.

En résumé (1)

Données : les données représentent les valeurs de q variables mesurées sur n individus ; les individus peuvent avoir un poids (en général $1/n$). Le plus souvent, on travaille sur des données centrées réduites Z (on retranche la moyenne et on divise par l'écart type).

Matrice de corrélation : c'est la matrice V de variance-covariance des variables centrées réduites. Elle possède q valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$.

Facteurs principaux u_k ce sont les vecteurs propres orthonormés de la matrice V (de dimension q) associés aux valeurs propres λ_k. Leur j-ième composante u_{jk} est le poids de la variable j dans la composante k.

Composantes principales ce sont les vecteurs $c_k = Zu_k$ de dimension n. Leur i-ième coordonnée c_{ik} est la valeur de la composante k pour l'individu i. Les c_k sont non corrélées et leur variance est $\text{Var}(c_k) = \lambda_k$.

En résumé (2)

Nombre d'axes : on se contente souvent de garder les axes interprétables de valeur propre supérieure à 1. La qualité de la représentation retenue est mesurée par la part d'inertie expliquée par ces composantes.

Cercle des corrélations il permet de visualiser comment les variables sont corrélées (positivement ou négativement) avec les composantes principales. A partir de là, on peut soit trouver une signification physique à chaque composante, soit montrer que les composantes séparent les variables en paquets. Seules les variables bien représentées (situées près du bord du cercle) doivent être interprétées.

En résumé (3)

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 données 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 très supérieure aux autres, on peut avoir intérêt à mettre l'individu en donnée supplémentaire.

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.