Analyse en composantes principales
Mme L.HAMDAD

Plan

Introduction

L'ACP est une méthode descriptive permettant de traiter des tableaux de données quantitatives $X_{n,p}$(de grandes dimension) où n représente le nombre d'individus et p le nombre de variables quantitatives. Le but de l'ACP est de résumer la grande quantité d'information contenue dans X, et cela dans un tableau de plus petite dimension $Y_{n,q}(q < p)$. Et ainsi fournir une représentation visuelle tels que :

- Y^jest une combinaison linéaire des p variables quantitatives, $X^j, j = 1, ..., p$.
- Les variables $(Y^j)_{j=1,...q}$ sont non correlées entre elles.
- Le tableau X peut être reconstitué à partir du nouveau tableau Y.
- Y contient le maximum d'informations sur X.

Exemple de tableau de données :

- Notes de n étudiants en p modules,
- Relevés des dépenses de ménages en 10 postes.
- Teneur en mineraux de certaines eaux, ect.....

Tableau de données

	X^1	$\cdots$	X^j	$\cdots$	X^p
1	$\cdots$	$\cdots$	x_1^j	$\cdots$	x_1^p
$\vdots$					
i	x_1^1		x_i^j		x_i^p
$\vdots$					
n	x_n^1		x_n^j		x_n^p

- x_1^j représente la mesure de la variable X^j sur l'individu "i".

A chaque individu "i" on associe le vecteur $X_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^p \end{pmatrix}$ et un poids

p_i, tel que $0 \leq p_i \leq 1$.

Le nuage de n individus appartenant à $\mathbb{R}^p$:

$$\aleph(I) = \{X_i \in \mathbb{R}^p, i = 1, ..., n\}.$$

L'espace $\mathbb{R}^p$ est muni d'une métrique qu'on notera M. Cette métrique peut être euclidienne c'est à dire que :

$$M = \begin{pmatrix} 1 & 0 & 0 \\ & \ddots & \vdots \\ & & 1 \end{pmatrix}$$

ou

$$M = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & 0 \\ & \ddots & \vdots \\ & & \frac{1}{\sigma p} \end{pmatrix}$$

σ_j représente l'écart type de de la variable X^j.

Remarque

Notons que le choix de l'une ou de l'autre des métriques se fera selon des cas qu'on citera ci après.

A chaque variable est associé le vecteur $X^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_n^j \end{pmatrix}$ de $\mathbb{R}^n$ et on définit le nuage de variables par :

$$\aleph(J) = \left\{ X^j \in \mathbb{R}^n, j = 1, ..., p \right\}$$

$\mathbb{R}^n$ est muni de la métrique des poids $D_p = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$. Lorsque les individus sont pris aléatoirement équiprobablement alors ; $p_i = \frac{1}{n}, \forall i = 1, \ldots, n.$

Le centre de gravité du nuage N(I)

Il est défini par

$$g = \frac{1}{n}\sum_{i=1}^{n} x_i = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^1 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_i^p \end{pmatrix} = \begin{pmatrix} \overline{x}^1 \\ \vdots \\ \overline{x}^p \end{pmatrix}$$

où $\overline{x}^j$ représente la moyenne arithmétique de la $j^{ième}$ variable.
L'inertie est une mesure de dispersion multidimentionnelle, elle est défini par :

$$I_g = \sum_{i=1}^{n} p_i \|x_i - g\|_M^2.$$

La mesure de dispersion dans le cas unidimentionnel n'est rien d'autre que l' écart type.

Formulation du problème d'ACP

Le principe est d'obtenir une représentation approchée du nuage $N(I)$ ($N(J)$) dans un sous espace de plus faible dimension par projection.
Ainsi, formellement :
1- On commence par rechercher un sous espace vectoriel de dimension 1, $E_1 = \Delta u_1$ engendré par un vecteur unitaire u_1, qui ajuste au mieux le $N(I)$ de $\mathbb{R}^n$
2- Ensuite rechercher un sous espace vectoriel de dimension 2, E_2 en déterminant Δu_2 orthogonal à Δu_1 qui ajuste au mieux le $N(I)$ de $\mathbb{R}^n$
3- En général rechercher un sous espace vectoriel E_k de dimension k en déterminant Δu_k orthogonal à Δu_{k-1} qui ajuste au mieux le $N(I)$ de $\mathbb{R}^n$ avec

$$E_k = \Delta u_k \oplus \Delta u_{k-1}$$

Détermination des axes factoriels

A partir de maintenant, on suppose que le tableau X est centré.
Ajustement sur $(\mathbb{R}^P, M)$: Dans ce cas le nuage $N(I)$ est ajusté.
On recherche le sous espace vectoriel de $dim1$, Δu_1 passant par l'origine et engendré par le vecteur unitaire u_1 qui ajuste au mieux le nuage N(I). Cela se fait, en déterminant u_1 qui maximise l'inertie du nuage $N(I)$, défini précedemment.
Notons par α_i la valeur de projection du vecteur individu X_i du nuage $N(I)$ sur l'axe Δu_k engendré par le vecteur unitaire u_k, α_i est donnée par :

$$\alpha_i = \langle X_i, u_1 \rangle_M = X_i^t M u_1,$$

Le vecteur de projection de tout les individus est donc donné par :

$$Y = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} X_1^t M u_1 \\ \vdots \\ X_n^t M u_1 \end{pmatrix} = XMu_1$$

Y est appelé composante principale.

Ainsi l'inertie du nuage $N(I)$ défini plus haut s'écrit :

$$I = \phi(u) = \| \ Y \ \|^2_{D_p} = u_1^t M X^t D_p X M u_1.$$

$V = X^t D_p X$ représente la matrice de variance covariance des p variables.
Dans la suite, nous déterminons u_1 unitaire qui maximise $\phi(u)$.
Commençons par écrire la fonction de Lagrange correspondant à notre problème d'optimisation :

$$L(u) = \phi(u) - \lambda(u^t M u - 1)$$

u_1 est solution du système suivant :

$$\begin{cases} \frac{dL(u)}{du}(u_1) = 0 \\ u_1^t M u_1 = 1 \end{cases}$$

Aprés résolution, nous obtenons l'équation suivante :

$$VMu_1 = \lambda u_1$$

. u_1 est vecteur propre de la matrice VM associé à la valeur propre λ.
Laquelle des valeurs propres de VM ?
En utilisant le contrainte sur le vecteur propre et en multipliant chaque coté de l'équation précédente par $u_1^T M$, nous obtenons :

$$u_1^t MVMu_1 = \lambda u_1^t M u_1 = \lambda.$$

Pour retrouver le sous espace vectoriel de dimension 2 qui ajuste au mieux le nuage de points $N(I)$, il suffit de trouver u_2 vecteur propre unitaire orthogonale à u_1 qui maximise $\phi(u)$.
Dans ce cas la fonction de Lagrange sous deux contraintes s'écrit :

$$L(u, v) = L(u) = \phi(u) - \lambda(u^t M u - 1) - \alpha u^t M v.$$

u_2 est solution du systhème

$$\begin{cases} \frac{dL(u,v)}{du}(u_2, u_1) = 0 \\ \frac{dL(u,v)}{dv}(u_2, u_1) = 0 \\ u_2^t M u_2 = 1, u_2^t M u_1 = 0 \end{cases}$$

Aprés résolution du système et en prenant en considération les contraintes, on déduit que u_2 est vecteur propre de VM associ é à la deuxième plus grande valeur propre.
En général, le sous espace vectoriel de dimension k qui ajuste au mieux le nuage de points $N(I)$ est engendré par les vecteurs propres $u_1, ..., u_k$ de VM unitaires et deux à deux orthogonaux associés aux valeurs propres $\lambda_1, ..., \lambda_k$, ordonnées de manière décroissantes, c'est Ã dire que $\lambda_1 \geq ... \geq \lambda_k$.

Ajustement sur $(\mathbb{R}^n, D_p)$

Dans ce cas le nuage $N(J)$ des variables est ajusté.
On recherche le sous espace vectoriel de dim 1, Δv_1 engendré par le vecteur unitaire v_1 qui ajuste au mieux le nuage $N(J)$ et ceci en déterminant v_1 qui maximise l'inertie du nuage $N(J)$, défini dans ce qui suit.
le sous espace vectoriel de dimension k qui ajuste au mieux le nuage de points $N(J)$ est engendrée par les vecteurs propres $v_1, ..., v_k$ de TD_p unitaires et deux à deux orthogonaux associés aux valeurs propres $\lambda_1, ..., \lambda_k$, ordonnées de manière décroissantes.

Remarque

Pour éviter la différence dans l'echelle de mesure de variables et pour faire jouer à chaque variable un rôle identique dans la définition des proximités entre individus, on passe à l'ACP normé qui consiste réduire les variables, c'est à dire :

$X_i^j \longrightarrow \frac{X_i^j}{\sigma_{j,}}$

ou bien utiliser la métrique $M = \begin{pmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_j} \end{pmatrix}$

Propriétés des composantes principales

Nous rappelons que $Y_\alpha(i)$ qui représente le vecteur de projection des individus sur l'axe factoriel δ_α est appelé composante principale ou nouvelle variable, ses propriétés sont :

$$\begin{aligned} \forall\alpha &= 1,...,p, \quad \overline{y}_\alpha = 0, \\ \|y_\alpha\|^2 &= vary_\alpha = \lambda_\alpha, \\ cov(y_\alpha, y_{\alpha'}) &= 0. \end{aligned}$$

Représentation d'un individu supplémentaire Soit x_i un individu supplémentaire, sa représentation est donnée par :

$$\alpha_{x_i} = \widetilde{x_i}^t u, \text{ tel que } \widetilde{x_i} = x_i - g^t.$$

Représentation d'une variable supplémentaire

Soit x^j une variable supplémentaire, sa représentation est donnée par : $\alpha_{x_i} = X^{jt} D_p v$, tel que $\widetilde{x^j} = x^j - \overline{x}^j$ est la variable centrée.

Remarque

Si l'ACP est normée en plus d'être centré les vecteurs sont réduits.

Formules de transitions

Ces dernières permettent de passer de l'analyse d'un nuage à un autre.

Proposition

: Les matrices XX^tD_p et X^tD_pX ont les mêmes valeurs propres.

Les aides à l'interprétation :
Qualité globale de représentation d'un axe factoriel : Elle est mesurée par le pourcentage d'inertie et elle est donnée par

$$I = \frac{\lambda_\alpha}{\sum_{r=1}^{p} \lambda_r} \times 100$$

Qualité d'un individu (variable) par un axe factoriel
a- **Individu** :

$$C_{re}^{\alpha}(i) = \cos_i^2(\theta) = \frac{(y_\alpha(i))^2}{\|x_i\|^2}$$

b-**Variable** :

$$C_{re}^{\alpha}(j) = \cos_j^2(\theta) = \frac{(V_\alpha(j))^2}{\|x^j\|_{D_p}^2}$$

Dés que $\cos_i^2(\theta) \simeq 1$, on dira que l'individu ou la variable sont trés bien représenté par le $\alpha^{tème}$axe factoriel.

Remarque

Il y a une relation très étroite entre le coéffcient de corrélation entre l'ancienne et la nouvelle variable et la projection de cette dernière sur l'axe factoriel, en effet

$$r(X^j, Y_\alpha) = \frac{V_\alpha(j)}{\sigma_j}$$

Ceci implique que lorsque l'ACP est normée, les variables varient à l'intérieur d'un cercle appelé cercle de corrélation.
✓ *Si les variables sont proches du cercle, alors elles seront bien représentées par le plan factoriel.*

Reconstitution du tableau de données : Le tableau de données est complètement reconstitué à partir de la formule suivante :

$$X = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha} v_\alpha u_\alpha^t$$

En effet à partir des formules de transition, on a

$$v = \frac{1}{\sqrt{\lambda}} Xu \Leftrightarrow \sqrt{\lambda} v = Xu$$

On multiplie les deux cotés de l'égalité par u^t on aura

$$\sqrt{\lambda} v u^t = Xuu^t \Leftrightarrow X \sum_{\alpha} u_\alpha u_\alpha^t =_\alpha \sqrt{\lambda_\alpha} v_\alpha u_\alpha^t$$

Récapitulation
Algorithme ACP

1. Calculer les moyennes des variables $\overline{X}^j, j = 1, ..., p$.
2. Centrer le tableau X (réduire si les données sont hétérogènes).
3. Calculer la matrice de variance covariance $V = X^t D_p X = \frac{1}{n} X^t X$.
4. Calculer les valeurs propres et les vecteurs propres de V.
5. Calculer les projections des individus et des variables sur les axes factoriels : $Y_\alpha = Xu, V_\alpha = \sqrt{\lambda_\alpha} u$.
6. Représenter graphiquement les individus et les variables.
7. Interpréter les résultats de l'analyse.