UNIVERSITÉ DE VERSAILLES SAINT-QUENTIN-EN-YVELINES

MODULE METHODOLOGIQUE M5

Statistique Descriptive :
Analyse en composantes principales (ACP)

UVED
Université Virtuelle Environnement & Développement Durable

30 AVRIL 2014
VERSION 300414

STATISTIQUE DESCRIPTIVE : ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Cécile Mallet

SOMMAIRE

5.Analyse en composantes principales (ACP)

Les méthodes qui permettent de réduire le nombre de variables sont appelées méthodes d'analyses factorielles Dans les analyses factorielles, on part du principe que si les données sont dépendantes entre elles, c'est parce qu'elles sont liées à des facteurs qui leur sont communs. L'intérêt des facteurs réside dans le fait qu'un nombre réduit de facteurs explique presque aussi bien les données que l'ensemble des variables, ce qui est utile quand il y a un grand nombre de variables. L'analyse en composante principale (ACP) est la plus courante des méthodes d'analyse factorielle.

L'objectif est d'obtenir une représentation de n points d'un espace de dimensions p dans un espace de dimension plus réduit (1D, 2D ou 3D) avec une perte d'information minimale, cette visualisation permettra éventuellement de discerner des sous groupes (classification des données), de détecter des valeurs extrêmes ou aberrantes, d'obtenir la dimension réelle du problème. L'ACP fait correspondre à p variables quantitatives décrivant n individus, q < p facteurs : les composantes principales, de telle manière que la perte d'information soit minimum. Les composantes sont organisées dans l'ordre croissant des pertes d'information (la première en perdant le moins), elles sont non corrélées linéairement entre elles. Le choix du nouvel espace de représentation s'effectue selon un critère qui revient à déformer le moins possible les similitudes et les différences entre individus. Ce qui revient à chercher le sous-espace dans lequel l'inertie du nuage projeté est maximale. L'ACP est une technique mathématique permettant de réduire un système complexe de corrélations en un plus petit nombre de dimensions.

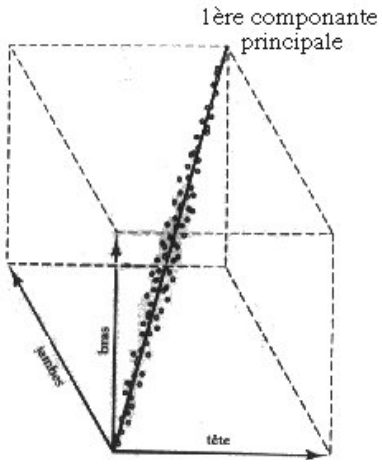
Lorsque plusieurs variables sont très corrélées, l'essentiel de l'information peut être contenu dans une seule composante, comme le montre l'exemple ci-dessous:

Exemple

Matrice de corrélations pour trois mesures très corrélées :

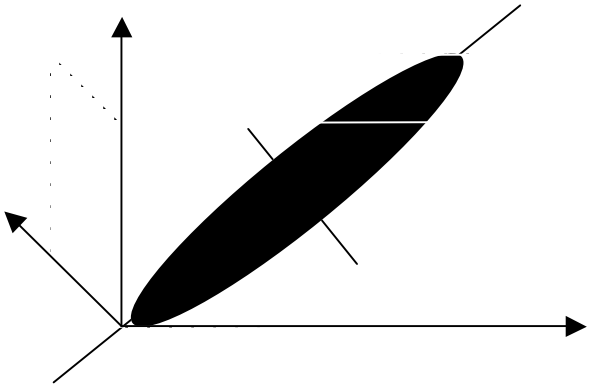
$$R=\begin{pmatrix} 1 & 0,91 & 0,72 \\ 0,91 & 1 & 0,63 \\ 0,72 & 0,63 & 1 \end{pmatrix}$$

L'essentiel de l'information peut être contenu dans une seule composante, comme le montre la figure suivante :



Représentation des individus et du premier axe principal dans un espace de dimension 3.

Dans le cas d'un nombre important de dimensions, la projection des données sur un seul axe ne suffit pas. Nous aurons besoin d'axes supplémentaires. Par convention, nous représentons la deuxième dimension par une droite perpendiculaire à la première composante principale. Ce deuxième axe, ou deuxième composante principale, se définit comme la droite qui "explique" (le mot n'a pas ici de signification causale) la plus grande partie de l'information restante (aucune autre droite qui "expliquerait" autant ou d'avantage ne pourrait être tracée perpendiculairement à la première composante principale). Si, par exemple, le nuage de point a la forme d'un ballon de rugby aplati, la première composante principale passerait par le centre suivant le grand axe de l'ellipse, et la deuxième également par le centre mais par le petit axe.



Représentation des individus et des deux premiers axes principaux dans un espace de dimension 3.

Méthode cas M = I

Concept général

Si on considère l'ensemble des données comme un nuage de n points-individus dans un espace de dimension p, dit ***espace des individus***. L'ACP consiste à définir dans cet espace p nouveaux axes appelés ***axe principaux*** du nuage choisis de manière à optimiser le nombre d'axe nécessaire pour rendre compte de presque toute la variance totale du nuage.

Variance totale

Chaque individu I_i est un point dont les coordonnées sont données par la $i^{éme}$ ligne de la matrice d'observation.

Soit X_j les variables centrées. ($X_j = X_j - \overline{X}_j$)

L'origine O de tous les axes de coordonnées est le centre G du nuage de points et la variance de chaque variable centrée X_j s'écrit $s_{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2$

Si les axes « X_j » sont orthogonaux entre eux et ont mêmes unités de mesure (base orthonormée) le vecteur $\mathbf{GI_i}$ a pour longueur $|GI_i|^2 = \sum_{j=1}^{p} x_{ij}^2$ en sommant sur l'ensemble des individus on obtient :

$$\sum_{i=1}^{n} |GI_i|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 = \sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2 = \sum_{j=1}^{p} n s_{X_j}^2$$

Soit en divisant par le nombre n d'individus :

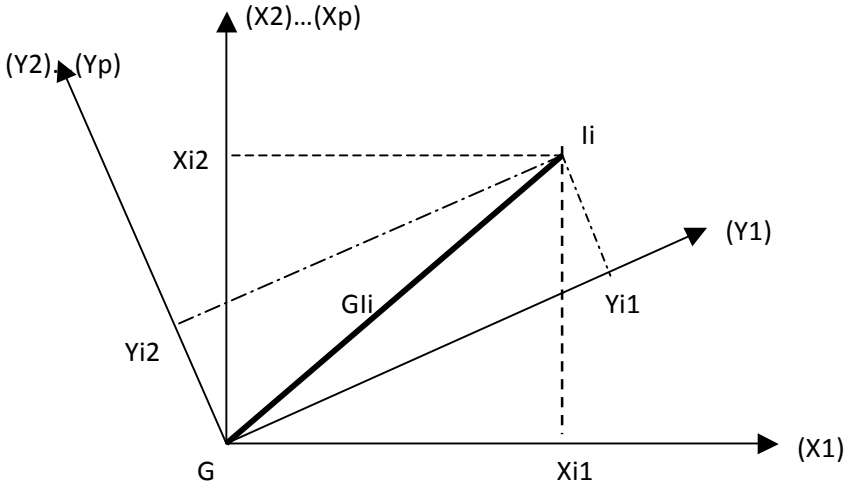
$$\Sigma^2 = \frac{1}{n}\sum_{i=1}^{n} |GI_i|^2 = \sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2\right) = \sum_{j=1}^{p} s_{X_j}^2$$

La somme des carrés des longueurs des vecteurs $\mathbf{GI_i}$ divisé par le nombre d'individus est égale à la somme des variances de toutes les variables, c'est la variance totale (ou inertie totale du nuage de points I dans le cas ou M = I)

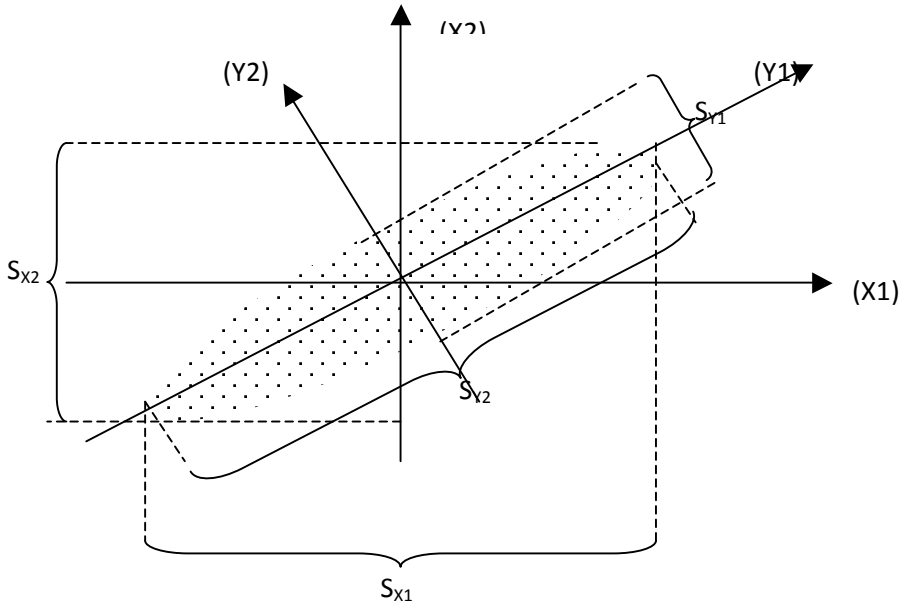
L'ACP consiste à définir comme axe de coordonnées, ceux propres au nuage de points. Ces nouveaux axes de coordonnées perpendiculaires définissent p nouvelles variables Y_k. Le passage des X_j aux Y_k se fait par une rotation du système d'axes de coordonnées autour de G dans l'espace à q dimensions. Dans cette rotation les longueurs des vecteurs $\mathbf{GI_i}$ sont inchangées. On en déduit que

$$\Sigma^2 = \sum_{j=1}^{p} s_{X_j}^2 = \sum_{k=1}^{p} s_{Y_k}^2$$

La variance totale du nuage reste inchangée dans toute rotation d'axes autour du centre G du nuage de points observés. L'ensemble des variances des variables initiales réalise une partition de la variance totale I en p variances des variables X_j, l'utilisation de nouveaux axes de coordonnées orthogonaux conduit à une nouvelle partition de la variance totale I en p variances des nouvelles variables Y_k .



Invariance de la variance totale dans une rotation du système d'axes orthogonaux autour du point moyen G.



Variances du nuage de points suivant les axes initiaux et les axes principaux.

Composantes principales

L'existence de corrélations entre les Xj se traduit par des inclinaisons des axes propres au nuage (axes principaux) par rapport aux axes de coordonnées, sur la figure on voit que quand X1 augmente X2 augmente aussi. En revanche, si on utilise les nouvelles variables Yk correspondant aux axes principaux le nuage apparaît sans inclinaison. Les variables Yk sont non corrélées. La matrice de variances-covariances des Yk est diagonale

$$S_{XX} = \begin{pmatrix} s^2_{X_1} & C_{X_1X_2} & \cdots & C_{X_1X_p} \\ C_{X_2X_1} & s^2_{X_2} & \cdots & \cdots \\ \cdots & & & \cdots \\ C_{X_pX_1} & & \cdots & s^2_{X_p} \end{pmatrix} \text{ alors que } S_{YY} = \begin{pmatrix} s^2_{Y_1} & 0 & \cdots & 0 \\ 0 & s^2_{Y_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & s^2_{Y_p} \end{pmatrix}$$

La diagonalisation est réalisée en calculant les valeurs propres et les vecteurs propres de S_{XX}. Les valeurs propres sont les éléments de la matrice diagonalisée, a savoir les variances $s^2_{Y_k}$. La variance totale étant invariante dans un changement de cordonnées la trace de S_{XX} est égale à celle de S_{YY}

les vecteurs propres donnent les nouveaux axes de coordonnées le long desquels les projections des individus définissent les nouvelles variables Y_k Les nouvelles variables Yk appelées composantes principales sont donc obtenues comme combinaison linaires des variables Xj

$$Y_k = a_{1k}X_1 + a_{2k}X_2 + ... + a_{pk}X_p \text{ ou}$$

la *k*-eme composante principale est notée Yk c'est le vecteur des coordonnées des individus sur l'axe défini par Ak (unitaire) *k*-eme vecteur propre.

$$y_{ik} = A_k^t x_i = \begin{pmatrix} a_{1k} & a_{2k} & & a_{pk} \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \\ x_{ip} \end{pmatrix} = a_{1k}x_{i1} + a_{2k}x_{i2} + \ldots + a_{pk}x_{ip}$$

Les coefficients varient avec k.

On réalise ainsi un changement de repère de X = (X1, X2, ..., Xp) à Y = (Y1, Y2, ..., Yp),

$$\begin{pmatrix} y_{11} & y_{21} & & y_{n1} \\ y_{12} & y_{22} & & y_{n2} \\ \\ y_{1p} & y_{2p} & & y_{np} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & & a_{p1} \\ a_{12} & a_{22} & & a_{p2} \\ \\ a_{1p} & a_{2p} & & a_{pp} \end{pmatrix} \begin{pmatrix} x_{11} & x_{21} & & x_{n1} \\ x_{12} & x_{22} & & x_{n2} \\ \\ x_{1p} & x_{2p} & & x_{np} \end{pmatrix} \text{ soit } Y_k = A_k X^t$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \\ x_{ip} \end{pmatrix} \quad i^{\text{éme}} \text{ individu dans l'espace des p variables } X_j = \begin{pmatrix} x_{1j} & x_{2j} & & x_{nj} \end{pmatrix}$$

$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \\ y_{ip} \end{pmatrix}$ *i^éme individu dans l'espace des p composantes principales* $Y_k = \begin{pmatrix} y_{1k} & y_{2k} & & y_{nk} \end{pmatrix}$

Classement des axes principaux

Soient $\lambda_1 > \lambda_2 > ... > \lambda_.$ les p valeurs propres de S_{XX} et aj pour j=1 à p les p vecteur propre normé de dimension p associé à $\lambda_k = s^2_{Y_j}$.

L'axe principal est donc l'axe définit par a1 vecteur propre associé à la plus grande valeur propre $\lambda_1 = s^2_{Y_1}$

Le second axe est celui, dans le plan perpendiculaire à a1, et maximisant la variabilité du nuage projeté sur cet axe, c'est l'axe définit par a2 vecteur propre associé à la plus seconde valeur propre $\lambda_2 = s^2_{Y_2}$

On définit ainsi successivement les axes principaux d'inertie comme les vecteurs propres de S.

La première composante principale est telle que l'axe Y1 maximise la dispersion des points, la seconde maximise la dispersion dans le plan orthogonal à Y1,

On peut ainsi sélectionner un nombre restreint d'axes significatifs en ne gardant que les axes qui ont une variance significative

On peut écrire $\Sigma^2 = s^2_{Y_1} + s^2_{Y_2} + ... + s^2_{Y_q} + ... + s^2_{Y_p}$ avec $s^2_{Y_1} > s^2_{Y_2} > .... + s^2_{Y_p}$ et dans certains cas on pourra garder seulement q <p axes significatifs tels que $\Sigma^2 \approx s^2_{Y_1} + s^2_{Y_2} + ... + s^2_{Y_q}$

Critère d'arrêt

Propriété des valeurs propres $\lambda_1 + \lambda_2 + .. + \lambda_.$ =trace (S)=$S_{11} + S_{22} + ... + S_{pp}$

La somme des valeurs propres des p variables est égale à la somme des variances des p variables

L'inertie totale d'un nuage de point (M = I) étant égale à la somme des variances elle est donc égale à la somme des valeurs propres

$$I_G = \sum_{j=1}^{p} \lambda_j$$

La variance d'une composante principale est égale à la valeur propre correspondante

La qualité de Yj est définie par :

$$\frac{\lambda_j}{trace(S)} = \frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}$$

généralement exprimé en %.

La perte d'information pour Yj est définie par :

$$1 - \frac{\lambda_j}{trace(S)}$$

La qualité de la représentation des deux premiers axes (Y1, Y2) est donnée par :

$$\frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{p} \lambda_k}$$

La perte d'information de (Y1, Y2) est donnée par :

$$1 - \frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{p} \lambda_k} = \frac{\sum_{k=3}^{p} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

Les premiers axes factoriels retenus pour la description du nuage fournissent le meilleur résumé du tableau de données totalisant la plus grande proportion de sa variance.

Le choix du nombre de composantes nécessaires est arbitraire

Pour limiter l'effet de la variance respective des différentes variables initiales on utilise généralement les variables centrées réduites, ce qui revient à utiliser la matrice de corrélation R au lieu de la matrice de dispersion S, autrement dit cela revient à utiliser une métrique $M = D_{\frac{1}{s^2}}$

Exemple

Soit un ensemble de données ou chaque individu est décrit par 3 variables (taille, poids, âge) on obtient la matrice de corrélation suivante :

$$\begin{pmatrix} 1 & 0,87779 & 0,81143 \\ & 1 & 0,74089 \\ & & 1 \end{pmatrix}$$

Le calcul des 3 valeurs propres fournit les informations suivantes :

Factor ou facteur	Eigenvalue ou valeur propre	Pct of Var ou pourcentage de variance	Cum Pct ou pourcentage de variance cummulé
1	2,62135	87,4	87,4
2	,26839	8,9	96,3
3	,11026	3,7	100,0

Le premier axe 'explique' 87,4% de l'information contenu dans l'ensemble des données à lui seul.

Interprétation

La méthode pour déterminer la signification d'une composante Yk est de la relier aux variables initiales (X1, X2, ..., Xp) en étudiant les corrélations entre composantes et variables initiales, c'est à dire en étudiant les coefficients de corrélation linéaires r(Yk,Xj)

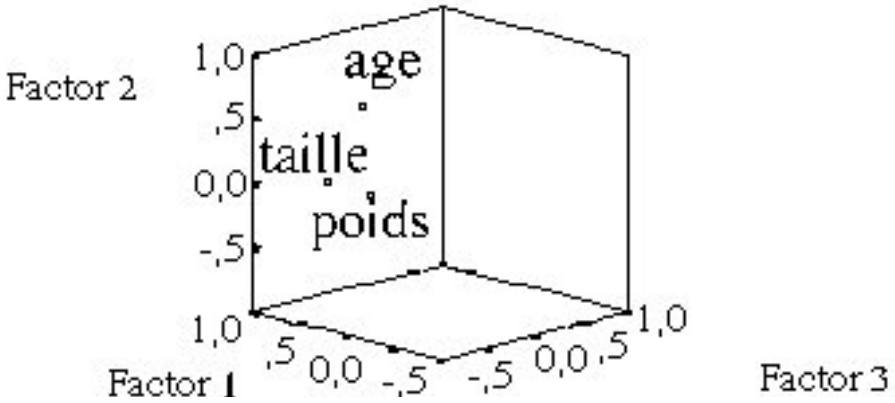
Exemple

Pour interpréter la signification des 3 axes on étudie les corrélations entre les axes principaux et les variables initiales : cosinus des angles entre les nouveaux axes et les anciens (la somme des carrés des éléments des lignes doit être égale à 1). On obtient

	Factor 1	Factor 2	Factor 3
TAILLE	,96060	-,09926	-,25961
POIDS	,93497	-,29807	,19233
AGE	,90799	,41194	,07661

Le premier facteur est très corrélé avec l'ensemble des variables c'est le facteur "de développement" où les individus les plus âgés tendent aussi à être les plus grands et les plus lourds, le second facteur oppose les individus plus âgés, mais plus petits et moins lourds. Le 3^{éme} facteur apporte très peu d'information (3,7 %) et est très difficilement à interpréter. La figure ci-dessous permet une représentation des variables initiales dans l'espace des facteurs. Le second facteur, est un facteur de

« morphologie » les individus ayant un score élevé sur ce facteur seront petit et maigre pour leur âge, alors que les individus ayant un score faible seront grand et gros pour leur âge.



Représentation des variables initiales dans l'espace des facteurs principaux.

L'analyse en composantes principales (en abrégé ACP) est donc une méthode de réduction du nombre de caractères permettant des représentations géométriques des individus et des caractères. Cette réduction ne sera possible que si les **p** caractères initiaux ne sont pas indépendants et ont des coefficients de corrélation non nuls.

L'ACP est une méthode factorielle car la réduction du nombre des caractères ne se fait pas par une simple sélection de certains d'entre eux, mais par la construction de nouveaux caractères synthétiques obtenus en combinant les caractères initiaux au moyen des "facteurs". C'est une méthode linéaire car il s'agit de combinaisons linéaires. Il existe de nombreuses autres méthodes d'analyses factorielles.