UNIVERSITE DE LORRAINE
Ecole Nationale Supérieure d'Electricité et de Mécanique

Introduction à l'Analyse en Composantes Principales

Didier Maquin

3 février 2014

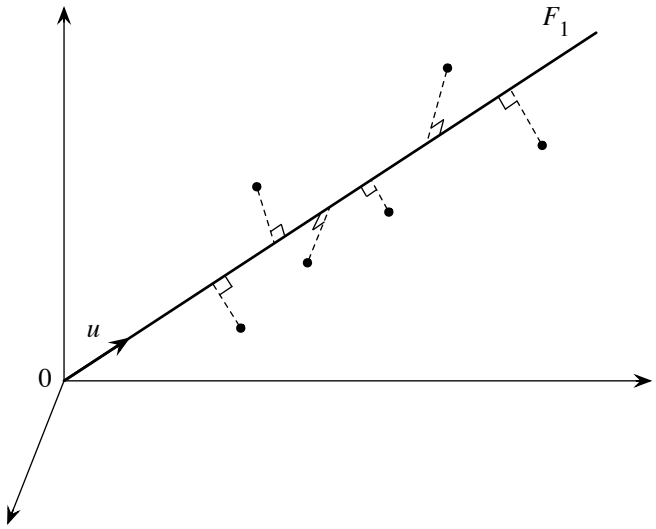


Table des matières

1 Analyses factorielles

Les différentes méthodes factorielles visent à établir des représentations synthétiques de vastes ensembles de valeurs numériques. Elles cherchent à établir un résumé descriptif de ces données qui soit appréhendable et exploitable par l'analyste tout en entraînant une perte d'information minimale. Plus exactement, on "consent une perte en information afin d'obtenir un gain en signification" (Volle, 1985).

Les méthodes factorielles ont en commun une étape de réduction de dimension qui peut se décrire ainsi. Pour un nuage de points, dont chacun est muni d'une masse, dans un espace vectoriel sur lequel est défini une métrique, calculer l'inertie totale de ce nuage, déterminer ses axes d'inertie, repérer les points dans la base formée par ces axes d'inertie.

Ainsi décrites, les analyses factorielles sont des techniques descriptives permettant d'identifier une structure de dépendance entre des observations multivariables afin d'obtenir une représentation compacte de celles-ci. Nous verrons cependant que ces techniques peuvent également être utilisés comme outil de modélisation permettant ainsi d'estimer les variables ou les paramètres du processus sur lequel les données ont été acquises.

Les objets traités par une analyse factorielle sont donc les suivants : l'espace, les points, les masses affectées aux points, la métrique et les résultats qu'elle fournit sont les axes d'inertie, les coordonnées des points sur ces axes et diverses indications annexes que l'on regroupe souvent sous l'expression "aides à l'interprétation".

Afin d'alléger la présentation qui suit, on supposera que les masses attribuées aux points sont unitaires et que la distance utilisée est la distance euclidienne (c'est le cas de l'Analyse en Composantes Principales qui fait plus particulièrement l'objet de ce document).

2 Analyse générale

2.1 Introduction

On s'attachera à résoudre dans ce paragraphe le problème d'approximation numérique suivant : étant donné un tableau rectangulaire de valeurs numériques représenté par une matrice X à n lignes et p colonnes, de terme général x_{ij}, est-il possible de reconstituer les np valeurs x_{ij} à partir d'un plus petit nombre de valeurs numériques ?

Contrairement aux diagrammes binaires qui projettent ces points sur deux des dimensions choisies plus ou moins arbitrairement, les analyses factorielles projettent les points sur une droite, un plan,.., un sous-espace à q dimension (avec $q \leq p$) choisi de façon à optimiser un certain critère. Intuitivement, on cherchera le sous-espace donnant la "meilleure visualisation" possible du nuage de points. Un bon choix consiste à rechercher la plus grande dispersion (le plus grand étalement) possible des projections dans le sous-espace choisi.

Pour débuter, supposons qu'il existe un vecteur u_1 à n composantes et un vecteur v_1 à p composantes tels que $X = u_1 v_1^T$. On aura alors reconstitué les np valeurs de X avec $n+p$ valeurs numériques seulement (dans ce cas, le rang de la matrice X est égal à 1). En pratique, il est extrêmement improbable de pouvoir obtenir une décomposition aussi simple. On cherchera donc une *approximation de rang q* pour X, c'est-à-dire une approximation de la forme :

$$X = u_1 v_1^T + u_2 v_2^T + \ldots + u_q v_q^T + E$$

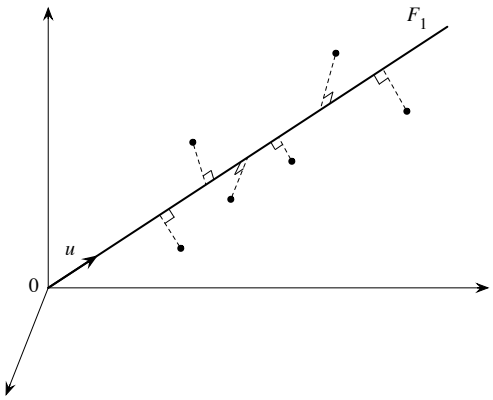
E étant une matrice $(n \times p)$ résiduelle dont les termes sont suffisamment petits pour que l'on puisse considérer que les np valeurs x_{ij} sont reconstituées de façon satisfaisante par les $q(n+p)$ valeurs des vecteurs u_i et v_i, $i = 1, \ldots, q$.

Ce problème peut être résolu en s'aidant de représentations géométriques. La matrice X donnera lieu à deux représentations : les n lignes de X peuvent être considérées comme les coordonnées de n points dans un espace à p dimensions $\mathbb{R}^p$ ou les p colonnes de X peuvent représenter les coordonnées de p points dans un espace à n dimensions.

2.2 Ajustement par un sous-espace vectoriel dans $\mathbb{R}^p$

Si le nuage des n points, représentant la matrice X dans cet espace, est contenu dans un sous-espace vectoriel de dimension q inférieure à p, il est alors possible de reconstituer les positions des n points (donc de reconstituer X) à partir des coordonnées sur q nouveaux axes et des composantes de ces nouveaux axes. On remplace ainsi np valeurs par $nq + pq$ autres valeurs.

On va donc chercher à ajuster le nuage des n points par un sous-espace vectoriel de $\mathbb{R}^p$, qui sera muni de la distance euclidienne usuelle.



Commençons par chercher la droite F_1 *passant par l'origine*, qui ajuste au mieux le nuage. Soit u un vecteur unitaire porté par cette droite, c'est-à-dire tel que $u^T u = 1$ ou encore $\sum_{j=1}^{p} u_j^2 = 1$. Puisque chaque ligne de X représente un point de $\mathbb{R}^p$, les n lignes du vecteur Xu sont les n produits scalaires de ces points avec u et sont donc les longueurs des projections de ces n points sur F_1.

Pour chaque point, le carré de sa distance à l'origine se décompose en carré de sa projection sur F_1 et en carré de sa distance à F_1. Les distances à l'origine étant données, il est équivalent de minimiser la somme des carrés des distances à F_1 ou de maximiser la somme des carrés des projections sur F_1. Si l'on veut que la somme des carrés des projections soit maximale, il faut chercher u qui rende maximale la quantité :

$$\phi = (Xu)^T Xu = u^T X^T Xu$$

Remarque : La maximisation de l'expression précédente correspond à celle de l'inertie du nuage de points expliquée par la direction portée par le vecteur u. En effet, par définition, l'inertie d'un nuage de points par rapport à un point c correspond à la somme des carrés des distances des différents points au point c. Si la mesure est effectuée par rapport à l'origine (c=0) les distances à prendre en compte sont les longueurs des vecteurs définis par les différents points. Dans ce cas, l'"inertie (par rapport à l'origine) expliquée par la direction portée par un vecteur u correspond à la somme des carrés des projections

orthogonales des différents points sur la droite portée par u. Comme indiqué précédemment, le vecteur Xu contient les longueurs des projections des points sur la droite portée par u. La somme des carrés des différentes longueurs s'écrira donc bien :

$$\|Xu\|^2 = u^T X^T X u$$

Ainsi, trouver un sous-espace vectoriel à une dimension qui ajuste au mieux, au sens des moindres carrés, le nuage de n points, revient à rendre maximale la forme quadratique $u^T X^T X u$ sous la contrainte $u^T u = 1$. On désignera ce sous-espace optimal par u_1.

On montre aisément que le meilleur sous-espace vectoriel à deux dimensions contient u_1. On le trouve en cherchant le vecteur u_2 unitaire et orthogonal à u_1 tel que $u_2^T u_1 = 0$ et $u_2^T u_2 = 1$) qui rend maximale la forme quadratique $u_2^T X^T X u_2$.

De façon analogue, à l'aide d'un raisonnement par récurrence, on verrait que le meilleur sous-espace vectoriel à q dimensions ($q \leq p$) est engendré par les vecteurs $u_1, u_2, \ldots, u_q$ où u_q est orthogonal à $u_1, u_2, \ldots, u_{q-1}$ et rend maximale la forme quadratique $u_q^T X^T X u_q$ avec $u_q^T u_q = 1$.

L'obtention des différents vecteurs u_i est aisée. Pour u_1, le problème d'optimisation à résoudre s'écrit :

$$\begin{cases} \max\limits_{u} \phi = u^T X^T X u \\ \text{sous la contrainte } u^T u = 1 \end{cases}$$

Le lagrangien associé à ce problème s'écrit :

$$\mathcal{L} = u^T X^T X u + \lambda(u^T u - 1)$$

où λ est un multiplicateur de Lagrange. La dérivation du lagrangien par rapport aux différentes composantes de u, puis l'annulation de ces dérivées conduisent à la relation matricielle :

$$2X^T X u - 2\lambda U = 0,$$

c'est-à-dire :

$$X^T X u = \lambda u.$$

Ceci montre que u est vecteur propre de la matrice $X^T X$. Remarquons également que $u^T X^T X u = \lambda u^T u = \lambda$. Le maximum cherché est donc une valeur propre de $X^T X$. Ainsi, u sera le vecteur propre u_1 correspondant à la *plus grande* valeur propre λ_1 de la matrice symétrique $X^T X$. Notons encore que cette plus grande valeur propre λ_1 représente la part d'inertie expliquée par la droite F_1

Si l'on recherche l'espace à deux dimensions qui s'ajuste au mieux au nuage, nous devons trouver une deuxième droite, portée par le vecteur unitaire t, passant par l'origine et qui maximise $t^T X^T X t$ en étant orthogonal à u_1 (c'est-à-dire $t^T u_1 = 0$ et $t^T t = 1$). Le problème d'optimisation correspondant s'écrit :

$$\begin{cases} \max\limits_{t} \phi = t^T X^T X t \\ \text{sous les contraintes } t^T t = 1 \\ \qquad\qquad\qquad\qquad t^T u_1 = 0 \end{cases}$$

Le lagrangien associé au problème d'optimisation s'exprime sous la forme suivante :

$$\mathcal{L} = t^T X^T X t + \lambda(t^T t - 1) - \mu t^T u_1$$

où λ et μ sont les multiplicateurs de Lagrange associés aux contraintes. L'annulation de ses dérivées partielles par rapport aux composantes de t conduit à la relation matricielle :

$$2X^TXt - 2\lambda t - \mu u_1 = 0$$

En prémultipliant cette relation par u_1^T et en remarquant que $u_1^T X^T X = \lambda_1 u_1^T$, on obtient :

$$2\underbrace{u_1^T X^T X t}_{=\lambda_1 u_1^T t} - 2\lambda u_1^T t - \mu \underbrace{u_1^T u_1}_{=1} = 0$$

Comme $u_1^T t = 0$, il vient $\mu = 0$; on déduit donc de l'équation initiale que :

$$X^TXt = \lambda t$$

Ainsi t est le second vecteur propre associé à la seconde plus grande valeur propre de X^TX. Le résultat s'étend aux éléments propres $\alpha = 1, 2, \ldots, r$ où r est le rang de X^TX.

Finalement, une base orthonormée du sous-espace vectoriel à q dimensions s'ajustant au mieux, au sens des moindres carrés, au nuage est constitué par les q vecteurs propres correspondant aux q plus grandes valeurs propres de la matrice symétrique X^TX.

2.3 Ajustement par un sous-espace vectoriel dans $\mathbb{R}^n$

On s'est intéressé, jusqu'à maintenant, au nuage de n points dans l'espace de dimension p. On peut également considérer le nuage de p points dans l'espace de dimension n. On cherche donc le sous-espace de dimension q ($q \leq p$) pour lequel la somme des carrés des projections est maximale.

On applique la même technique que précédemment. Le premier axe factoriel de $\mathbb{R}^n$ est porté par un vecteur unitaire v et l'on cherche à rendre maximale la somme des carrés des projections X^Tv.

$$\begin{cases} \max\limits_{v} \phi = v^T X X^T v \\ \text{sous la contrainte } v^T v = 1 \end{cases}$$

On nomme v_1 le vecteur satisfaisant le problème précédent. En suivant ensuite une démarche analogue à la précédente, on peut établir que le sous-espace vectoriel de dimension q qui assure une dispersion maximale des p points est défini par une base orthonormée formée des q vecteurs propres $v_1, v_2, \ldots, v_q$ correspondant aux q plus grandes valeurs propres de la matrice XX^T.

2.4 Lien entre les sous-espaces de $\mathbb{R}^p$ et de $\mathbb{R}^n$

Cherchons les relations qui existent entre les vecteurs u_i et v_i précédents. Par définition de v_i, on a :

$$XX^Tv_i = \mu_i v_i$$

où v_i et μ_i désignent respectivement le $i^{\text{ème}}$ vecteur propre et la $i^{\text{ème}}$ valeur propre de XX^T. Notons r le rang de la matrice X, donc le rang de la matrice XX^T. Il y a alors r valeurs propres non nulles avec $r \leq \min(n, p)$. En prémultipliant les deux membres de cette relation par la matrice X^T, on obtient :

$$X^TX(X^Tv_i) = \mu_i(X^Tv_i)$$

A chaque vecteur propre $v_i, (i \leq r)$ de XX^T correspond donc un vecteur propre $u_i = X^Tv_i$ de X^TX relatif à la même valeur propre μ_i. Toute valeur propre non nulle de XX^T est donc valeur propre de X^TX et les vecteurs propres correspondants sont liés par les relations :

$$u_i = k_i X^T v_i$$

où k_i sont des constantes. En prémultipliant de façon analogue par X les deux membres de la relation $X^T X u_i = \lambda_i u_i$, on obtient :

$$XX^T(Xu_i) = \lambda_i(Xu_i)$$

Ainsi, à tout vecteur propre $u_i, (i \leq r)$ de $X^T X$ correspond un vecteur propre Xu_i de XX^T relatif à la même valeur propre λ_i. On a donc finalement, pour tout $i \leq r$:

$$\lambda_i = \mu_i, \quad \text{et} \quad v_i = k_i' X U_i$$

où les coefficient k_i' sont constants. En écrivant que tous les vecteurs u_i et v_i sont normés, $u_i^T u_i = v_i^T v_i = 1$, on trouve $k_i = k_i' = 1/\sqrt{\lambda_i}$. On peut alors écrire, pour $i = 1, \ldots, r$, le système de relations suivant souvent appelées formules de transition :

$$\begin{aligned} u_i &= \frac{1}{\sqrt{\lambda_i}} X^T v_i \\ v_i &= \frac{1}{\sqrt{\lambda_i}} X u_i \end{aligned}$$

L'axe F_i qui porte le vecteur unitaire u_i est appelé $i^{\text{ème}}$ axe factoriel de $\mathbb{R}^p$ et l'axe G_i qui porte le vecteur unitaire v_i est appelé $i^{\text{ème}}$ axe factoriel de $\mathbb{R}^n$.

Les coordonnées des points du nuage sur le $i^{\text{ème}}$ axe dans $\mathbb{R}^p$ (respectivement dans $\mathbb{R}^n$) sont, par construction, les composantes du vecteur Xu_i (respectivement de $X^T v_i$). Il y a donc proportionnalité entre coordonnées des points sur le $i^{\text{ème}}$ axe dans un espace et composantes unitaires (cosinus directeur) du $i^{\text{ème}}$ axe de l'autre espace.

2.5 Reconstruction complète et partielle de la matrice de départ

Les formules de transition précédentes peuvent s'écrire de manière plus compacte en définissant les matrices de vecteurs propres $U = (u_1\ u_2\ \ldots\ u_r)$ (de dimension $p \times r$) et $V = (v_1\ v_2\ \ldots\ v_r)$ (de dimension $n \times r$) et la matrice diagonale des valeurs propres $\Lambda = \text{diag}(\lambda_1\ \lambda_2\ \ldots\ \lambda_r)$. On a alors :

$$U = X^T V \Lambda^{-1/2} \qquad \text{et} \qquad V = XU\Lambda^{-1/2}$$

En multipliant, par la droite, la seconde égalité par la matrice régulière $\Lambda^{1/2}$, on obtient :

$$XU = V\Lambda^{1/2}$$

ou encore en post-multipliant par la matrice U^T et en remarquant que $UU^T = I$ (les vecteurs propres d'une matrice symétrique sont orthogonaux), on a également :

$$XUU^T = V\Lambda^{1/2}U^T$$

c'est-à-dire :

$$X = V\Lambda^{1/2}U^T$$

Cette dernière expression indique clairement que l'on peut reconstruire (complètement) la matrice X si l'on connaît les valeurs propres et les vecteurs propres des matrices symétriques XX^T et $X^T X$. Cette décomposition est connue sous le nom de "décomposition en valeurs singulières" (SVD for Singular Value Decomposition).

Par définition, on appelle $i^{\text{ème}}$ composante principale le vecteur dont les composantes sont les coordonnées des points du nuage sur le $i^{\text{ème}}$ axe factoriel (cette analyse peut être effectuée dans $\mathbb{R}^p$ ou dans $\mathbb{R}^n$).

Composantes principales dans $\mathbb{R}^p$ (projections des observations sur les vecteurs propres)	$C_o = XU = V\Lambda^{1/2}$
Composantes principales dans $\mathbb{R}^n$ (projections des variables sur les vecteurs propres)	$C_v = X^T V = U\Lambda^{1/2}$

La reconstruction peut également être partielle ou approchée, on parlera alors d'*estimation*. En effet, la décomposition précédente peut s'écrire sous la forme :

$$X = \sum_{i=1}^{r} \sqrt{\lambda_i} v_i u_i^T$$

Si les valeurs propres $\lambda_i, i = 1, \ldots, r$ sont rangées par ordre décroissant, une estimation (reconstruction partielle) peut consister à ne tenir compte que des q premières valeurs propres (les q plus grandes), comme indiqué dans l'introduction, pour obtenir :

$$X \simeq \hat{X} = \sum_{i=1}^{q} \sqrt{\lambda_i} v_i u_i^T$$

On remplace ainsi les np éléments de X par un ensemble de $q(n+p)$ nombres constitué des q vecteurs $\sqrt{\lambda_i} v_i$ à n composantes et q vecteurs u_i à p composantes.

Comme chaque valeur propre λ_i mesure la somme des carrés des projections sur l'axe F_i, la part d'inertie expliquée par l'axe F_i s'explicite sous la forme [1] :

$$I(k) = \frac{\lambda_k}{\sum_{i=1}^{r} \lambda_i}$$

La qualité de la reconstruction s'appuyant sur les seules q premières valeurs propres peut donc être évaluée par la quantité :

$$\tau_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{r} \lambda_i}$$

Ce quotient, appelé *taux d'inertie*, mesure ainsi la part de la dispersion du nuage imputable au sous-espace à q dimensions retenu.

2.6 Diversification de l'analyse générale

Dans ce qui précède, on s'est essentiellement attaché à un problème d'approximation numérique. Lorsqu'il s'agit d'analyse statistique, on dispose le plus souvent d'informations complémentaires sur la nature des données et la prise en compte de ces informations conduit à procéder à des transformations préalables des données de départ. Le type de transformation que subit la matrice des données X détermine le mode d'analyse ultérieur.

L'Analyse en Composantes Principales (ACP) utilise la matrice X centrée de terme général $(x_{ij} - \overline{x_j})/\sqrt{n}$ ou $\overline{x_j}$ est la moyenne des n valeurs de la variable j. Cette transformation permet de remédier à l'hétérogénéité des moyennes des variables.

L'analyse en Composantes Principales Normée utilise la matrice X centrée et réduite. La transformation permet de s'affranchir de l'hétérogénéité des variables tant en moyenne qu'en

1. On notera que l'inertie totale du nuage est la somme des valeurs propres de $X^T X$, quantité qui est égale à la trace de cette matrice : $\text{Trace}(X^T X) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{r} \lambda_i$ car, cette matrice étant de rang égal à r, seule r valeurs propres sont différentes de zéro.

dispersion. Elle élimine donc l'effet du choix arbitraire des unités de mesure et donne la même importance aux différentes variables.

L'analyse des Rangs utilise la matrice des rangs R. Celle-ci est obtenue en remplaçant la valeur observée x_{ij} par le rang r_{ij} de l'observation i dans le classement des n valeurs de la variable j. Toutes les variables transformées ont alors même moyenne : $m = (n+1)/2$ et même variance $s^2 = (n^2-1)/12$. L'analyse générale s'effectue alors sur le tableau d'éléments $(r_{ij} - m)/s\sqrt{n}$.

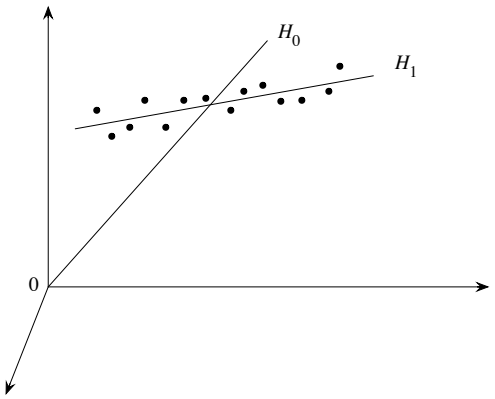
L'analyse des Correspondances est plus particulièrement dévolue à l'étude de tableaux de contingence ou le terme général x_{ij} est une fréquence (ventilation d'une population selon divers caractères par exemple). Pour de tels tableaux, les sommes en ligne et les sommes en colonne ont un sens ; les lignes et les colonnes jouent des rôles analogues et le tableau transposé est étudié de façon similaire.

3 Analyse en composantes principales normée

3.1 Usage et description

L'analyse en composantes principales s'utilise essentiellement lorsque la matrice de données initiale T représente des valeurs numériques issues de n observations de p variables.

Lorsqu'on effectue l'analyse dans $\mathbb{R}^p$, les n points dans cet espace sont les observations. On souhaite obtenir une représentation de la proximité de ces observations dans un espace de faible dimension. Or, le sous-espace à une dimension qui assure une déformation minimale (une perte d'information minimale) des proximités entre observations n'a aucune raison d'être assujetti à passer par l'origine du repère. En effet, ce n'est pas la position du nuage qui nous intéresse, mais la forme de ce nuage.



Il est clair, sur la figure précédente, que le sous-espace affine H_1 rend mieux compte des proximités entre points que le sous-espace vectoriel (droite passant par l'origine) H_0.

Cette remarque conduit à prendre comme nouvelle origine le centre de gravité du nuage dont les p composantes sont les p moyennes arithmétiques $\overline{t_j}$. On effectuera donc l'analyse générale sur la matrice de terme général $x_{ij} = (t_{ij} - \overline{t_j})/\sqrt{n}$. L'influence du niveau de chacune des variables sera ainsi éliminée. Le coefficient $1/\sqrt{n}$ n'a pour objet que de faire coïncider la matrice à diagonaliser $X^T X$ avec la *matrice des covariances expérimentales* conformément à un usage répandu.

Une modification supplémentaire de la matrice initiale peut également être nécessaire si les dispersions des variables sont très différentes ; elle conduira à l'analyse en composantes principales normée, c'est-à-dire l'analyse de la matrice de terme général :

$$x_{ij} = (t_{ij} - \overline{t_j})/s_j\sqrt{n} \qquad \text{avec} \qquad s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(t_{ij} - \overline{t_j})^2$$

En effet, dans le calcul du carré de la distance entre les observations i et i',

$$d^2(i,i') = \frac{1}{n}\sum_{j=1}^{p}(t_{ij} - t_{i'j})^2/s_j^2$$

la division de chaque terme de la somme par la variance s_j correspondante conduit à une contribution analogue de chaque variable indépendamment de sa dispersion.

En analyse en composantes principales normée, la matrice à diagonaliser X^TX coïncide alors avec la *matrice des corrélations* entre les variables.

En résumé, l'analyse du nuage de points dans $\mathbb{R}^p$ conduit à effectuer une translation d'origine au centre de gravité du nuage et à changer les échelles des différents axes. L'analyse générale s'effectue alors sur la matrice des corrélations. Les coordonnées des points-observations sur l'axe F_i sont alors les produits scalaires constituant les lignes de Xu_i.

Si l'on considère maintenant l'analyse dans $\mathbb{R}^n$, les p points du nuage sont maintenant les points-variables. L'analyse de la matrice X dans $\mathbb{R}^p$ induit, comme on l'a vu précédemment, une analyse dans $\mathbb{R}^n$. Cependant, les indices i et j ne jouent pas des rôles symétriques dans la transformation initiale. Les interprétations géométriques associées à cette transformation seront donc différentes.

Ainsi, la transformation $x_{ij} = t_{ij} - \overline{t_j}$ qui était interprétée comme une translation de l'origine dans $\mathbb{R}^p$ est maintenant, dans $\mathbb{R}^n$ une projection parallèle à la première bissectrice. Le changement d'échelle des axes, c'est-à-dire la division de chaque coordonnée de $\mathbb{R}^p$ par $s_j\sqrt{n}$, devient ici une déformation du nuage qui ramène chacun des points-variables à la distance 1 de l'origine. On a en effet dans cet espace, pour chaque point-variable j et par définition de s_j :

$$d^2(j,0) = \frac{1}{n}\sum_{i=1}^{n}(t_{ij} - \overline{t_j})^2/s_j^2 = 1$$

Les p points-variables sont donc sur une hypersphère de rayon 1 centrée à l'origine qui est le centre de gravité du nuage. La distance entre deux points-variables j et j' s'écrit :

$$d^2(j,j') = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_{ij} - \overline{t_j}}{s_j} - \frac{t_{ij'} - \overline{t'_j}}{s'_j}\right)^2$$

Après développement du carré et sommation sur l'indice i, on trouve :

$$d^2(j,j') = 2(1 - \rho_{jj'})$$

où $\rho_{jj'}$ est le coefficient de corrélation [2] entre les variables j et j'. Ainsi, les proximités entre points-variables pourront s'interpréter en termes de corrélations : les points sont très proches si leur corrélation est fortement positive ($\rho_{jj'} \simeq 1$) et éloignés si elle est fortement négative ($\rho_{jj'} \simeq -1$).

2. Pour deux vecteurs t_j et $t_{j'}$, de moyennes respectives $\overline{t_j}$ et $\overline{t'_j}$ et de variances s_j^2 et $s_{j'}^2$, le coefficient de corrélation linéaire entre t_j et $t_{j'}$ est défini par $\rho_{jj'} = \mathrm{cov}_{jj'}/(s_j s_{j'})$ avec $\mathrm{cov}_{jj'} = \frac{1}{n}\sum_{i=1}^{p}(t_{ij} - \overline{t_j})(t_{ij'} - \overline{t_{j'}})$

3.2 Interprétation d'une analyse en composantes principales

Les composantes principales peuvent être considérées comme de nouvelles variables, combinaisons linéaires des variables initiales, non corrélées entre elles et d'inertie (ou de variance) maximale.

On peut éditer l'image des projections du nuage de points sur les plans formés par des couples d'axes factoriels. La disposition des projections des points-variables permet d'interpréter le nuage des points-observations.

Remarquons que l'image des points-observations est centrée sur l'origine comme les points-observations eux-même (on a pratiqué un centrage des données). Il n'en est pas de même de l'image des points-variables et il peut se produire que toutes les projections soient situées d'un même coté de l'origine.

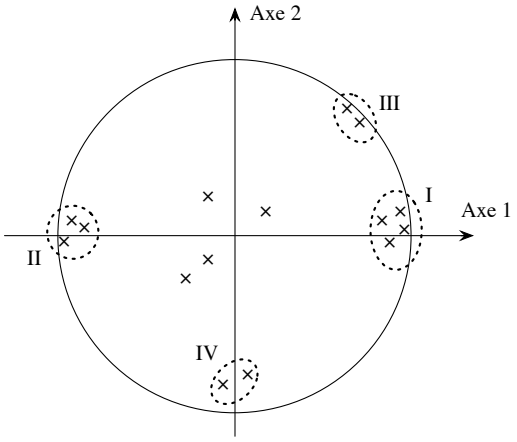
En règle générale, l'interprétation d'une ACP s'effectuera en analysant d'abord les images des projections du nuage des variables, on tentera alors d'interpréter la signification des axes factoriels. On considérera ensuite l'analyse des images des projections des observations en référence à l'interprétation des axes effectuée précédemment.

Interprétation du nuage des variables

Les projections des points-variables étant situé sur l'hypersphère (0,1), les images des points du nuage sur un plan factoriel se trouveront toutes à l'intérieur d'un cercle centré à l'origine et de rayon égal à 1. Il est utile de tracer ce cercle sur le graphe donnant l'image du nuage.

Les points-variables sont d'autant mieux représentées par le plan correspondant que leur image est proche du bord du cercle. On trouvera assez souvent une configuration semblable à celle dessinée ci-après [3] :

- sur l'axe 1, un groupe de variables figure avec des coordonnées proches de 1 (groupe I), un autre groupe avec des coordonnées proches de -1 (groupe II),
- un groupe se trouve près du bord du cercle, sans avoir de coordonnées fortes ni sur l'axe 1, ni sur l'axe 2 (groupe III),
- un groupe a des coordonnées relativement proches de -1 sur l'axe 2 (groupe IV),
- enfin, quelques variables se projettent à l'intérieur du cercle.

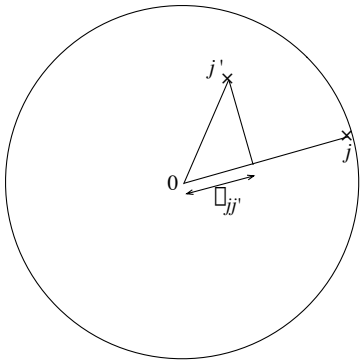


3. La situation décrite permet d'expliciter l'analyse, il ne s'agit que d'un exemple dont la portée n'est pas générale

On dira que l'axe 1 oppose les variables du groupe I à celles du groupe II. Chacun de ces groupes est formé de variables fortement corrélées entre elles ; deux variables appartenant chacune à l'un de ces groupes ont un coefficient de corrélation proche de -1. On peut interpréter la première composante principale comme une nouvelle variable qui serait approximativement fonction linéaire croissante de chacune des variables du groupe I et fonction linéaire décroissante de chacune des variables du groupe II. Ainsi, l'analyse factorielle apporte deux résultats : une description des corrélations qui distingue et oppose les groupes I et II et une nouvelle variable, la composante principale, qui peut être substituée à chacune des variables de ces groupes sans que l'on perde beaucoup d'information.

L'interprétation d'un nuage de points-variables peut se faire d'une façon purement visuelle en regardant les projections. La qualité de la représentation d'un point par un axe se lit en effet directement sur le graphique. Dans l'exemple précédent, les variables du groupe III sont bien représentées par le plan (1, 2), même si elles ne sont pas bien représentées ni par l'axe 1 ni par l'axe 2. Etant situées près du bord du cercle, les points de ce groupe ne sont pas éloignés du plan (1, 2).

Si la variable j est représentée par un point très proche du bord du cercle, remarquons que l'on peut lire directement sur le graphique le coefficient de corrélation $\rho_{jj'}$ avec une autre variable j' quelconque. Il suffit d'abaisser sur $0j$ la perpendiculaire issue de la projection de la variable j'.

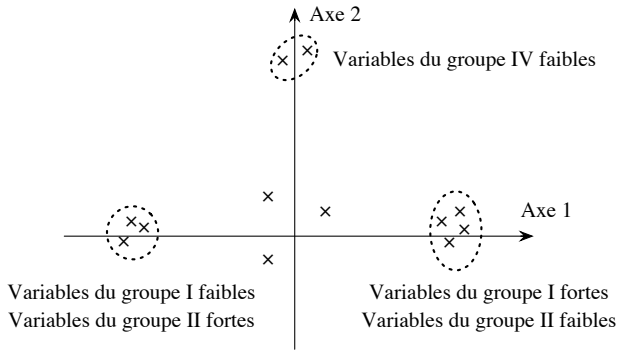


Le groupe IV est formé de variables dont la corrélation avec les variables des groupes I et II est nulle et qui ont une corrélation négative avec les variables du groupe III.

Le groupe IV est proche du bord du cercle (mais moins que les groupes I et II) dans la direction négative de l'axe 2. La deuxième composante principale est donc corrélée négativement avec les variables du groupe IV et cette corrélation est moins forte que celle qui lie la première composante principale aux variables des groupes I et II.

Interprétation du nuage des observations

Après l'interprétation du nuage des variables, on analyse le nuage des observations. Toujours pour l'exemple considéré, on sait, pour l'essentiel, que l'axe 1 oppose les observations pour lesquelles les variables du groupe I sont fortes et celles du groupe II sont faibles (projections situées à droite) aux observations pour lesquelles c'est l'inverse (projections situées à gauche).



Sur l'axe 2, on trouvera, en haut, les observations pour lesquelles les variables du groupe IV sont faibles et en bas, les observations pour lesquelles elles sont fortes.

4 Elaboration d'un "modèle ACP"

En règle générale, l'élaboration d'un modèle passe par deux étapes. Au cours de la première, on choisit une classe de modèles imposant ainsi la *structure* de celui-ci. On utilise ensuite les données expérimentales pour *identifier les paramètres* décrivant la structure choisie.

Lorsqu'on effectue une analyse en composantes principales, par construction, on se restreint à la classe des modèles linéaires. La structure du modèle dépend du nombre de composantes principales retenues pour l'estimation de la matrice initiale. L'estimation des paramètres du "modèle ACP" se résume au choix des valeurs et vecteurs propres de la matrice de corrélation des données.

4.1 Choix du nombre d'axes

L'analyse en composantes principales cherche donc une approximation de la matrice initiale des données X au moyen d'une matrice de rang inférieur issue d'une décomposition en valeurs singulières. La question qui se pose alors concerne le choix du nombre de composantes principales (nombre d'axes) qui doit être retenu pour "capter" l'information pertinente contenue dans X. De nombreuses règles ont été proposées dans la littérature pour cette détermination. Nous ne présenterons ici que les deux règles heuristiques les plus élémentaires.

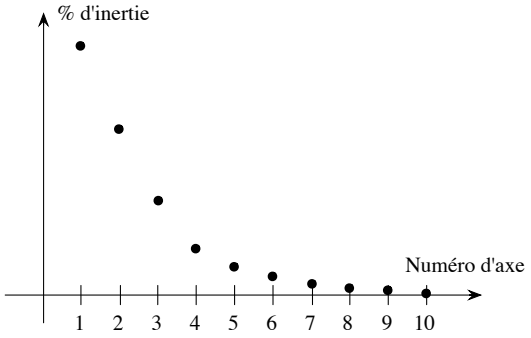
Nous avons vu précédemment que la part d'inertie du nuage de points, expliquée par l'axe F_i, est proportionnelle à la valeur propre correspondante (cf. section 2.5). Le nombre d'axes retenu peut alors être obtenu en fixant un seuil correspondant au pourcentage minimum d'inertie que l'on veut restituer. Pour un seuil I_{min}, le nombre de composantes à retenir est le plus petit entier q vérifiant la relation :

$$\tau_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{r} \lambda_i} \leq I_{min}$$

Si les données initiales sont bruitées, la décision basée uniquement sur le pourcentage d'inertie (ou de variance) expliquée est un peu arbitraire. Sa capacité à fournir le nombre "correct" de composantes principales dépendra fortement du rapport signal sur bruit.

Comme l'inertie expliquée par chaque axe va en décroissant, on peut construire la courbe de décroissance des axes en portant en abscisse les numéros des axes (ou des valeurs propres

correspondantes) et en ordonnée les pourcentages d'inertie restitués par chaque axe (l'amplitude des valeurs propres).



Si l'on observe un "coude" dans cette courbe, on ne retiendra que les axes dont le numéro d'ordre est situé avant ce coude. Pour l'exemple de la figure précédente, on retiendrait 5 composantes.

4.2 Modèles ACP

Revenons sur la décomposition en valeurs singulières d'une matrice de données X de dimension $n \times p$. Dans le cas général, on peut écrire, comme on l'a montré à la section 2.5 :

$$X = VSU^T$$

où S est la matrice diagonale des valeurs singulières correspondant aux racines carrées des valeurs propres de la matrice X^TX. Dans cette décomposition, on a $V \in \mathbb{R}^{n\times p}$, $S \in \mathbb{R}^{p\times p}$ et $U \in \mathbb{R}^{p\times p}$. On fera l'hypothèse que la décomposition est telle que les valeurs singulières sont ordonnées par ordre décroissant d'amplitude. Lorsqu'on effectue une estimation de X, en ne s'appuyant que sur les q premières composantes principales, on n'utilise que les q premières colonnes de V et de U. La décomposition peut alors être partitionnée de la manière suivante :

$$X = \begin{pmatrix} V_1 & V_2 \end{pmatrix} \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} = V_1 S_1 U_1^T + V_2 S_2 U_2^T$$

où $V_1 \in \mathbb{R}^{n\times q}, S_1 \in \mathbb{R}^{q\times q}$ et $U_1^T \in \mathbb{R}^{q\times p}$ et l'estimation de X s'écrit alors :

$$\hat{X} = V_1 S_1 U_1^T$$

Cette estimation peut également s'exprimer en fonction de X. En effet en post-multipliant l'expression de X par $U_1U_1^T$, on obtient :

$$XU_1U_1^T = V_1S_1U_1^TU_1U_1^T + V_2S_2U_2^TU_1U_1^T$$

En notant que les vecteurs propres de la matrice U forment une base orthogonale, on a $U_1^TU_1 = I$ et $U_2^TU_1 = 0$ d'où :

$$\hat{X} = XU_1U_1^T$$

La matrice C représente alors le "modèle ACP" des données initiales. Différents modèles ACP peuvent être engendrés selon le nombre de composantes principales retenues. On remarquera cependant, compte tenu de l'analyse précédente, que ces modèles sont implicitement tous déterminés une fois que l'on a calculé les valeurs propres et vecteurs propres de la matrice de corrélation des données. Contrairement à des techniques de régression linéaire multivariables

où l'incorporation d'une variable dans un modèle remet en cause l'ensemble de la procédure d'identification paramétrique, il suffit ici simplement de sélectionner des vecteurs propres.

Remarque : Considérons une matrice X, de dimension $n \times p$ (avec $n > p$), de données non bruitées, dont le rang est égal à r. Cela signifie qu'il existe implicement $p - r$ relations linéaires indépendantes entre les colonnes de cette matrice. Il y a alors $p - r$ valeurs singulières nulles dans la décomposition en valeurs singulières de X (ou $p - r$ valeurs propres nulles dans la décomposition en valeurs propres de $X^T X$). Lorsque les données considérées sont bruitées, ou lorsque les relations de dépendance ne sont pas rigoureusement linéaires, les valeurs propres précédentes ne sont plus rigoureusement nulles. On s'intéresse alors aux seules valeurs propres "dominantes" au sens de celles qui expliquent le plus l'inertie ou la variance des points-observations, d'où l'intérêt des critères de choix du nombre d'axes présentés sommairement à la section 4.1.

5 ACP et détection de défaut

5.1 Notion de reconstruction

Le principe de reconstruction consiste à estimer une des variables d'un vecteur d'observations x (de dimension $p \times 1$) [4], notée x_i en utilisant les autres variables x_j, $j = 1, \ldots, p, j \neq i$, à partir du modèle ACP.

Cette reconstruction peut s'effectuer de manière itérative. On utilise tout d'abord le modèle ACP pour obtenir une première estimation $\hat{x}$ du vecteur d'observation x.

$$\hat{x} = Cx$$

ou sous forme développée :

$$\begin{pmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_i \\ \vdots \\ \hat{x}_p \end{pmatrix} = \begin{pmatrix} c_1^T \\ \vdots \\ c_i^T \\ \vdots \\ c_p^T \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_p \end{pmatrix}$$

où les c_i^T représentent les lignes de la matrice C. Soit $\hat{x}_i$, la $i^{\text{ème}}$ composante de cette estimation, on remplace alors la $i^{\text{ème}}$ composante du vecteur d'observation x par cette estimation et on reconduit le processus d'estimation. Cette opération est répétée jusqu'à constater la convergence vers une valeur $\hat{z}_i$ de la $i^{\text{ème}}$ composante de l'estimation. Chaque itération du calcul, utilisant le modèle ACP, correspond à une projection orthogonale dans le sous-espace des composantes principales. Si l'on note k le numéro d'itération de ce calcul, on a, avec $\hat{z}_i^{(0)} = x_i$:

$$\hat{z}_i^{(k)} = c_i^T \begin{pmatrix} x_1 \\ \vdots \\ \hat{z}_i^{(k-1)} \\ \vdots \\ x_p \end{pmatrix}$$

4. Dans cette partie, le vecteur d'observation considéré est une colonne, contrairement à l'organisation initiale de la matrice de données où les observations constituaient les lignes de la matrice.

En explicitant la $i^{\text{ème}}$ ligne c_i^T, on a :

$$\hat{z}_i^{(k)} = \begin{pmatrix} c_{i1} & \cdots & c_{i(i-1)} & c_{ii} & c_{i(i+1)} & \cdots & c_{ip} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ \hat{z}_i^{(k-1)} \\ x_{i+1} \\ \vdots \\ x_p \end{pmatrix}$$

c'est-à-dire :

$$\hat{z}_i^{(k)} = \begin{pmatrix} c_{i1} & \cdots & c_{i(i-1)} & 0 & c_{i(i+1)} & \cdots & c_{ip} \end{pmatrix} x + c_{ii}\hat{z}_i^{(k-1)}$$

Une fois la convergence du calcul itératif obtenue, la solution $\hat{z}_i$ vérifie :

$$\hat{z}_i(1 - c_{ii}) = \begin{pmatrix} c_{i1} & \cdots & c_{i(i-1)} & 0 & c_{i(i+1)} & \cdots & c_{ip} \end{pmatrix} x$$

On obtient donc l'estimation $\hat{z}_i$ sous la forme :

$$\hat{z}_i = \frac{\begin{pmatrix} c_{i1} & \cdots & c_{i(i-1)} & 0 & c_{i(i+1)} & \cdots & c_{ip} \end{pmatrix}}{(1 - c_{ii})} x$$

Cette estimation n'est possible que si $c_{ii} \neq 1$. En effet, si $c_{ii} = 1$, cela signifie que la variable x_i n'est pas corrélée avec les autres variables et ne peut donc pas être reconstruite à partir des autres variables.

5.2 Détection de défauts

La méthode de reconstruction présentée précédemment peut être utilisée pour détecter des capteurs défaillants. Supposons en effet que l'on ait établi un modèle ACP satisfaisant à partir de données supposées saines. Ce modèle traduit les liaisons linéaires existant entre les variables. Pour surveiller le bon fonctionnement d'un capteur particulier, on peut reconstruire la valeur de sa mesure, supposée indisponible, à l'aide du modèle ACP et des mesures issues des autres capteurs. Si la mesure fournie par le capteur considéré est saine (exempte de défauts), la différence entre cette mesure et sa reconstruction doit être "faible". La reconstruction constitue ici une "prédiction" de la mesure de la variable correspondante. Si un capteur est en défaut, la mesure correspondante n'est plus cohérente avec les autres mesures et le modèle ACP. L'écart entre la mesure et sa reconstruction devient important. L'amplitude de l'écart entre une mesure et sa reconstruction constitue donc un indicateur du bon fonctionnement d'un capteur.

Références

[1] Dunia R., Qin S. A subspace approach to multidimensional indentification and reconstruction. *AIChE Journal*, vol. 44, pp. 1813-1831.

[2] Harkat M.F. Détection et localisation de défauts par analyse en composantes principales. Doctorat de l'Institut National Polytechnique de Lorraine, 30 juin 2003.

[3] Jolliffe I.T. *Principal component analysis*. Springer Series in Statistics. Springer-Verlag New York, 1986.

[4] Lebart L., Morineau A., Fénelon J.P. *Traitement des données statistiques - méthodes et programmes*. Bordas, Paris, 1979.

[5] Volle M. *Analyse des données*. Collection "Economie et statistiques avancées", Economica, 1985.