–Texte–
Analyse en composantes principales

1 Introduction

En archéologie, l'analyse de la composition de matériaux est devenue un outil essentiel pour l'étude des échanges dans les économies antiques. Des objets d'origines distinctes ont généralement des signatures chimiques différentes qui permettent d'identifier leur origine.

Pour identifier ces signatures il faut être capable de regrouper entre eux des objets de composition similaire. Si la composition était un vecteur de dimension 2 (teneur en deux composants), chaque objet pourrait être représenté dans le plan, des groupes distincts pourraient être identifiés, et un nouvel objet pourrait être attribué à un de ces groupes sans grande difficulté. Les problèmes commencent lorsque la dimension augmente.

Des chercheurs ont mesuré la composition de 45 poteries trouvées en Grande Bretagne datant de l'époque romaine[1], obtenue par spectrophotométrie. La composition est la teneur en 9 oxydes. Ces poteries proviennent de 5 fours différents (les 21 premières mesures proviennent du four 1, les 12 suivantes du four 2, les 2 suivantes du four 3, les 5 suivantes du four 4 et les 5 dernières du four 5).

PAl_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	MnO	BaO	Four
18.8	9.52	2	0.79	0. 4	3.2	1.01	0.077	0.015	1
16.9	7.33	1.65	0.84	0. 4	3.05	0.99	0.067	0.018	1
18.2	7.64	1.82	0.77	0. 4	3.07	0.98	0.087	0.014	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14.8	2.74	0.67	0.03	0.05	2.15	1.34	0.003	0.015	5
19.1	1.64	0.6	0.1	0.03	1.75	1.04	0.007	0.018	5

Soit plus généralement un tableau de n individus et p variables, $n > p$

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}.$$

1. D'après J.Holland Jones et I.G.Robertson, `www.stanford.edu/class/anthsci192`

n est typiquement grand devant p. Chaque ligne x_i est un individu et chaque colonne x^k représente une variable. Chaque individu est un point de l'espace $\mathbb{R}^p$.

S'il y a deux variables, la représentation des individus ne pose pas de problème : on les positionne dans le plan (chaque axe représente une variable). On peut voir apparaître ainsi typiquement des liens entre les variables (si les individus sont sur une (ou plusieurs) droites), ou bien des groupes d'individus.

L'objectif de l'ACP est de réaliser cette démarche dans le cas de plus de deux variables. L'idée est la suivante : supposons dans un premier temps que les individus soient en fait concentrés dans un plan de $\mathbb{R}^p$, le bon sens veut alors que l'on représente directement les individus dans ce plan. Mathématiquement cela signifie que l'on fait un changement de base où les deux premiers axes sont dans le plan et les autres leurs sont orthogonaux ; les coordonnées de 3 à p seront donc nulles pour tous les individus.

Supposons maintenant que les individus sont presque dans un plan (le ■presque■ devra être précisé et pose justement problème, mais on se peut se faire intuitivement une idée raisonnable de ce que cela peut signifier, de même lorsque l'on parle de points presque alignés dans le plan). L'idée est alors de trouver le plan le meilleur (en gros au sens où la somme des distances des points au plan est la plus petite possible) et de représenter les données dans ce plan. On va voir que ceci peut se faire mathématiquement sans grande difficulté.

Si ce ■meilleur plan■ n'est en réalité pas très bon, on peut alors chercher le sous-espace de dimension trois le meilleur, mais la représentation des données sera plus difficile : en pratique on représentera plusieurs projection en dimension deux et l'interprétation commence à être difficile.

Un changement de base sur les variables produit un tableau

$$C = XU$$

où U est la matrice de changement de base, de dimensions $p \times p$. La question posée est de trouver un changement de base tel que les premières nouvelles variables (premières colonnes de C)

- concentrent l'information (variables significatives)
- possèdent de bonnes propriétés descriptives : absence de corrélation entre variables.

Le but de L'ACP est d'extraire de nouvelles variables classées par ordre d'importance décroissante en un sens que l'on définira.

2 Inertie des espaces

Notons $\tilde{x}_i, \tilde{X}$ les individus recentrés :

$$\tilde{x}_i = x_i - \bar{x}, \quad \tilde{X} = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}.$$

La moyenne empirique $\bar{x}$ est aussi appelée centre de gravité. On définit l'inertie des individus par la quantité

$$I = \frac{1}{n} \sum_i \|\tilde{x}_i\|^2.$$

L'inertie est donc également la somme des variances empiriques des p variables ; c'est encore n^{-1} fois le carré de la norme de Frobenius de $\tilde{X}$ (somme des carrées des coefficients). Cette quantité réelle mesure la **dispersion** des individus dans l'espace à p dimensions. Une autre mesure de dispersion, cette fois-ci matricielle, est la matrice de covariance empirique des individus

$$R = \frac{1}{n} \sum_i \tilde{x}_i^T \tilde{x}_i = \frac{1}{n} \tilde{X}^T \tilde{X}, \quad R_{jk} = \frac{1}{n} \sum_i \tilde{x}_i^j \tilde{x}_i^k.$$

L'inertie est simplement la trace de R.

Soit P un projecteur de $\mathbb{R}^p$; par abus, on désignera ici également par P la matrice de P dans la base canonique[2] ; le vecteur x_i étant un vecteur ligne, la projection de x_i sera la projection du vecteur colonne associé, remise en ligne, soit $P(x_i) = (Px_i^T)^T$:

$$P(x_i) = x_i P^T, \quad XP^T = \begin{pmatrix} x_1 P^T \\ \vdots \\ x_n P^T \end{pmatrix}.$$

Soit E un sous-espace de $\mathbb{R}^p$ et P_E le projecteur orthogonal sur E, on note I_E l'inertie des individus projetés :

$$I_E = \frac{1}{n} \sum_i \|P_E(\tilde{x}_i)\|^2 = \frac{1}{n} \sum_i \|\tilde{x}_i\|^2 - \frac{1}{n} \sum_i \|\tilde{x}_i - P_E(\tilde{x}_i)\|^2$$

(noter que les individus projetés puis recentrés sont aussi les projetés des individus recentrés). L'inertie de E est donc une mesure de la proximité entre les individus et E. On appellera aussi I_E l'inertie de E. On vérifie immédiatement que l'inertie de E est

$$I_E = \mathrm{Tr}(P_E R P_E).$$

2. On rappelle que, comme la base canonique est orthogonale, P est un projecteur orthogonal si et seulement si la matrice P est idempotente et symétrique.

3 Propriétés fondamentales de l'ACP

Soient E et F deux espaces orthogonaux, on a alors $P_{E\oplus F} = P_E + P_F$ puis par le théorème de Pythagore :

$$I_{E\oplus F} = I_E + I_F.$$

Theorème 3.1. *Désignons par I_k l'inertie maximale des espaces de dimension k ; alors il existe une suite croissante F_k d'espaces de dimension k tels que*

$$I_{F_k} = I_k, \quad F_k \subset F_{k+1}, \quad k = 1, ...p-1.$$

Démonstration. Raisonnons par récurrence et construisons F_{k+1} à partir de F_k. Soit E un espace de dimension $k+1$ dont l'inertie vaut I_{k+1}. Comme

$$\dim(E) + \dim(F_k^\perp) = k + 1 + p - k > p,$$

on peut trouver un vecteur u dans $E \cap F_k^\perp$; soit G l'orthogonal de u dans E, alors

$$I_{k+1} = I_E = I_G + I_u \leq I_{F_k} + I_u = I_{F_k\oplus u} \leq I_{k+1}$$

Donc l'espace $F_{k+1} = F_k \oplus \mathbb{R}u$ est d'inertie maximale. □

Lemme 3.2. *Le vecteur unitaire u_k (unique au signe près) tel que*

$$F_k = F_{k-1} \oplus \mathbb{R}u_k$$

est d'inertie maximale parmi tous les vecteurs orthogonaux à F_{k-1}. Tout vecteur possédant cette propriété définit un espace F_k de dimension k d'inertie maximale.

Démonstration. Soit F un espace de dimension k contenant F_{k-1} et u le vecteur unitaire de F orthogonal à F_{k-1} ; alors $I_F = I_{F_{k-1}} + I_u$, donc F est d'inertie maximale si et seulement si u est d'inertie maximale parmi tous les vecteurs orthogonaux à F_{k-1}. □

Par ailleurs on a, comme la matrice de projection sur $\mathbb{R}u_k$ est $u_k u_k^T$:

$$I_{u_k} = u_k^T R u_k.$$

Il s'ensuit que u_k est solution du problème :

Problème :

Maximiser $u^T Ru$

Sous $\|u\| = 1$ *et* $u \perp u_j, \quad j = 1, ...k-1$

qui se résout à l'aide du

Theorème 3.3. *Toute suite u_k satisfaisant les conditions ci-dessus est une suite de vecteurs propres de R associé chacun à la k-ième valeur propre (par ordre décroissant) λ_k de R.*

Il s'ensuit que l'espace de dimension k d'inertie maximale F_k est engendré par les k vecteurs propres associées aux k plus grandes valeurs propres (si des valeurs propres sont égales il n'y a pas unicité).

Démonstration. Si R est diagonale, la démonstration est directe. Sinon, il existe une matrice orthogonale O et une matrice diagonale D telles que

$$R = O^T DO.$$

Si l'on pose $u'_k = O_k$, le problème se réécrit, du fait que O est orthogonale

$$\textit{Maximiser } {u'_k}^T D'_k \qquad \textit{sous} \quad \|u'_k\| = 1 \textit{ et } u'_k \perp u'_j, \quad j = 1, ...k-1.$$

On s'est donc ramené à la situation diagonale et on sait que u'_k est la suite des vecteurs de la base canonique pris par ordre décroissant de valeur de ${u'_k}^T Du'_k$. Les $u_k = O^T u'_k$ sont donc bien les vecteurs propres de r choisis par valeur décroissante de $u_k^T r u_k$, c'est-à-dire de la valeur propre correspondante. □

Calcul pratique de l'ACP. Diagonaliser $R = \tilde{X}^T \tilde{X}$ sous la forme $R = UDU^T$, D diagonale décroissante, U orthogonale. Les u_i sont dans les colonnes de U. Les composantes principales sont dans la matrice $C = XU$. La i-ème composante principale est la combinaison linéaire des variables avec les poids contenus dans la i-ième colonne de U.

Définition 3.4.

Les u_k sont les axes principaux.

Le vecteur $c^k = Xu_k$ est la k-ième composante principale.

$\frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p} = \frac{I_{F_k}}{I}$ est la fraction d'inertie expliquée par F_k.

Il résulte de ce qui précède que la matrice de covariance des c_i est la diagonale des λ_k.

Plus la fraction d'inertie expliquée par F_k est proche de 1, plus la projection des variables x_j sur F_k est proche de x_j, c'est-à-dire que les $c_1, ...c_k$ permettent de bien représenter les individus.

Bilan. L'analyse en composantes principales propose donc un changement de base $C = XU$ sur les variables explicatives tel que les nouvelles variables (composantes principales) soient empiriquement décorrélées et que les variables d'origine soient le plus proche possible des sous espaces emboîtés F_k. Ce changement de base est orthogonal.

4 Normalisation : ACP sur données réduites

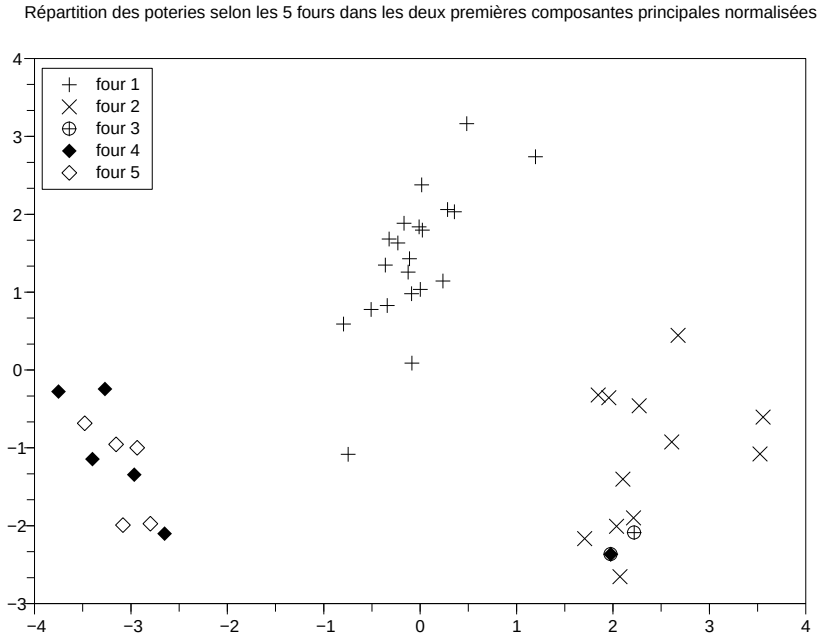
L'ACP n'est pas invariante par changement d'échelle sur les variables. Il est donc important, si les colonnes de X continent des données non comparables (i.e. des mètres et des kilogrammes) de les normaliser, afin d'avoir un résultat **indépendant des unités utilisées**. Si en revanche les colonnes sont comparables, on peut préférer de ne pas faire de normalisation : on considère ainsi que l'information de niveau relatif entre les différentes variables est importante, et les variables de faible amplitude seront pénalisées au sens où elles interviendront moins dans les premières composantes principales

5 Représentations dans les plans principaux

Il s'agit de la représentation des individus comme un nuage de points dans un des plans (c_i, c_j).

Le tracé des individus dans le plan (c_1, c_2) peut faire apparaître des individus marginaux ou des classes bien séparées. Cette représentation est d'autant meilleure que la fraction d'inertie expliquée par F_2 est grande.

Reprenons le problème initial des poteries antiques. La matrice X est la matrice 45×9 des compositions. On a tracé les individus dans le plan (c_1, c_2) en marquant le numéro du four. On voit nettement des regroupements qui permettent de classifier différents types de poterie (l'ACP est ici normalisée mais une ACP non normalisée donne des résultat similaires) :



6 Approfondissements

6.1 Liens avec la décomposition en valeurs singulières

Theorème 6.1. *Soit $X \in \mathbb{R}^{n\times p}$ une matrice, avec $n \geq p$. Il existe deux matrices à colonnes orthonormées U et V (i.e. $U^TU = V^TV = Id$) et une matrice diagonale $D \in \mathbb{R}^{p\times p}$ à entrées positives telles que*

$$X = UDV^T$$

Par conséquent, en appelant u_i et v_i les vecteurs colonne de U et V et d_i les éléments diagonaux de D, on a la décomposition en somme de matrices de rang 1 :

$$X = \sum_{i=1}^{p} d_i u_i v_i^T$$

La matrice D contient nécessairement les racines carrées des valeurs propres de X^TX, et si ces dernières sont distinctes cette décomposition est unique.

Démonstration. On ne traite que le cas où X^TX a toutes ses valeurs propres non nulles. On peut diagonaliser X^TX :

$$X^TX = V\Delta V^T.$$

On vérifie alors que $D = \sqrt{\Delta}$ et $U = XVD^{-1}$ convient.

Pour l'unicité, noter que u_i (resp. de v_i) est le vecteur propre de XX^T (resp. X^TX) associé à d_i^2. □

6.2 Approximation de matrices

Notons $\|M\|_F$ la norme de Frobenius de M : $\|M\|_F^2 = Tr(M^TM) = \sum M_{ij}^2$. Le théorème 3.1 s'énonce également comme suit :

> *Pour toute matrice de projection P orthogonale sur un espace de dimension k, on a*
>
> $$\|X - XP_k\|_F \leq \|X - XP\|_F.$$

Mais plus généralement :

Theorème 6.2. *Pour toute matrice A de rang k on a*

$$\|X - XP_k\|_F \leq \|X - A\|_F.$$

Démonstration. En effet, si P désigne le projecteur orthogonal sur l'orthogonal du noyau de A, on a $A = AP$ et donc :

$$\|X - A\|_F^2 = \|X(Id - P) + (X - A)P\|_F^2 = \|X(Id - P)\|_F^2 + \|(X - A)P\|_F^2 \geq \|X - XP_k\|_F.$$

□

7 Suggestions

1. On pourra démontrer certains résultats présentés dans le texte.
2. On pourra expliquer le principe de l'analyse en composantes principales.
3. On pourra faire le lien entre ACP et décomposition d'une matrice en valeurs singulières.
4. On pourra utiliser le fichier `poterie.dat` pour illustrer la méthode générale (et retrouver en particulier la figure du texte.
5. On pourra simuler des points presque alignés sur une droite puis calculer la droite correspondant à la première composante principale.
6. On pourra aussi simuler des points en dimension 3 proches d'un plan donné (passant par l'origine pour simplifier) puis essayer de voir si le plan défini par les deux premières composantes principales est proche du plan de départ.

