Introduction à la Science des données
Analyse en composantes principales

V. Monbet - fortement inspiré du cours de B. Delyon

[1] Université de Rennes/UFR Mathématiques

Outline

Tableaux individus/variables : Premier exemple.

Tour de hanche et poids de 500 individus [1] :

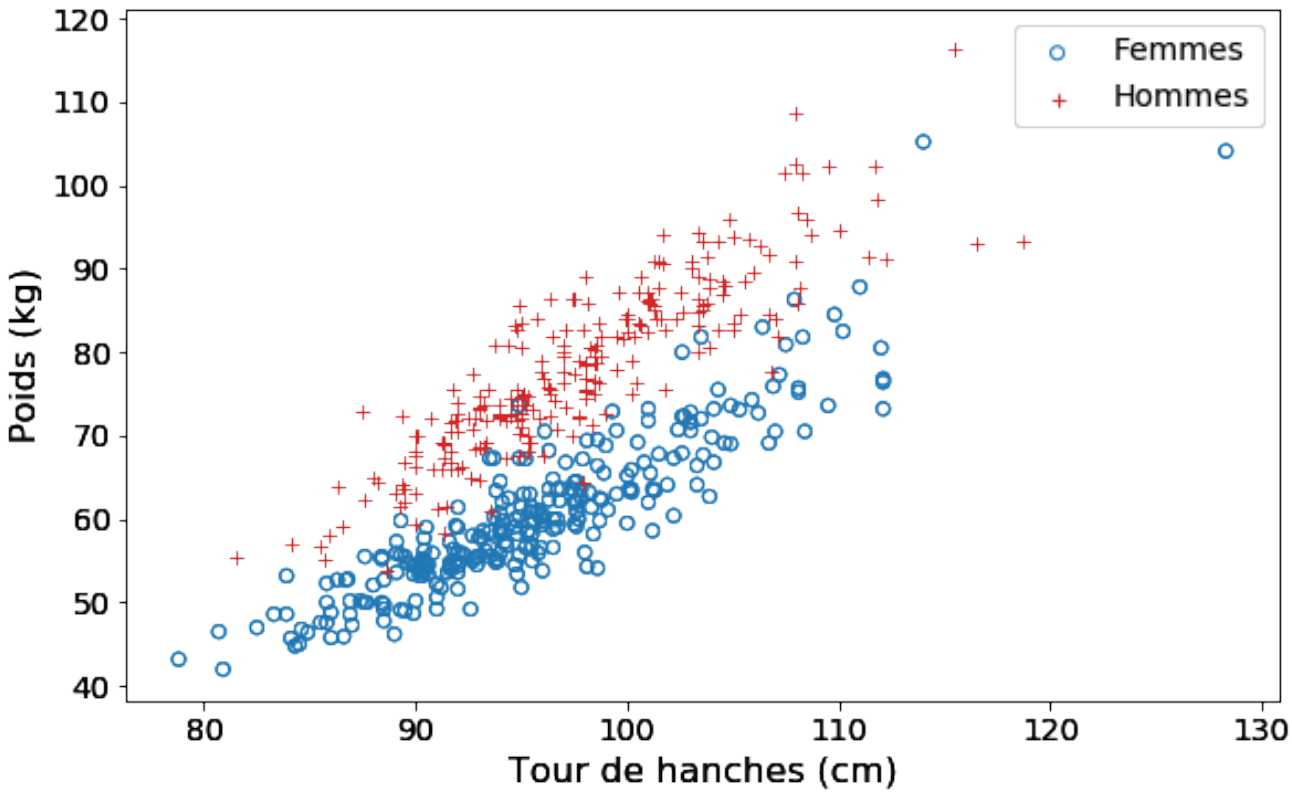
Tour de hanches (cm)	Poids (kg)
93.5	65.6
94.8	71.8
⋮	⋮
95.0	80.7

1. G. Heinz, L.J. Peterson, R.W. Johnson et C.J. Kerk, «Exploring Relationships in Body Dimensions»' *Journal of Statistics Education Volume 11, Number 2 (2003)*. www.amstat.org/publications/jse/v11n2/datasets.heinz.html

Deux variables : représentation dans le plan

$\longrightarrow$ liens entre variables
$\longrightarrow$ groupes d'individus.

Tour de hanches et poids :



- Lien entre les deux variables (alignement approximatif des points)
- Deux groupes distincts femmes/hommes.

Individus/variables : Deuxième exemple.

Composition de 45 poteries trouvées en Grande Bretagne datant de l'époque romaine[2]. 5 fours différents.

$X \in \mathbb{R}^{45\times 9}$:

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

Ou $X \in \mathbb{R}^{45\times 10}$ si l'on ajoute le four.

L'analyse de la composition de matériaux est un outil important 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.

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

Notations

On introduit des notations qui serviront aussi dans la suite du cours.

- La table X est composée de p variables observées pour n individus.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

- $\mathbf{X} \in \mathbb{R}^{n,p}$
- Une ligne x_i représente un individu (ou une observation) ; c'est un point dans $\mathbb{R}^p$
- Une colonne $x_{.j}$ représente une variable.

Analyse en composantes principales

Remarque
Si les individus sont concentrés dans un plan de $\mathbb{R}^p$, représenter dans ce plan.
Mathématiquement : changement de base (rotation) puis représentation des deux premières variables, les autres (de 3 à p) étant nulles.

Idée générale
Si les individus sont tous "proches" d'un plan,
- **trouver le plan le meilleur au sens où la somme des distances des points au plan est la plus petite possible**
- représenter les données par la projection sur ce plan.

Qualité recherchée
Les distances entre individus projetés doivent refléter au mieux leurs distances dans $\mathbb{R}^p$.
Respecter la géométrie des données.

Remarquesimportantes

- Projeter en dimension deux peut être trop grossier. On peut chercher le meilleur sous-espace de dimension trois (ou plus)
  ✗représentation des données plus difficile : en pratique on représentera 2 ou 3 projections en dimension 2.
  ✗interprétation délicate.

- Difficulté liée à la **normalisation des variables**.
  Si par exemple p = 3, multiplier la 3ème variable par un petit facteur concentrera les individus sur le plan contenant les deux premiers.
  Dire que les individus sont presque dans un plan dépend des échelles !

- **L'Analyse en Composantes Principales (ACP) consiste à créer de nouvelles variables appelées composantes principales.**

Outline

Problème mathématique

Trouver le plan Π qui minimise

$$\sum_i d(x_i, \Pi)^2 = \sum_i \|x_i - P(x_i)\|^2$$

P : projection orthogonale sur le plan Π.

Recentrage : Soit $\bar{x}$ l'individu moyen, si l'on recentre les variables d'abord

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \in \mathbb{R}^p,$$

$$X \longleftarrow \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}.$$

le problème reste le même sauf que tout est décalé du vecteur $\bar{x}$, et Π devient un plan vectoriel (i.e. $0 \in \Pi$), ce qui simplifie énormément les écritures.
Dans la suite du cours, on supposera ce recentrage préalable fait.

Attention : en pratique, pour des données, il faut penser à faire le recentrage !

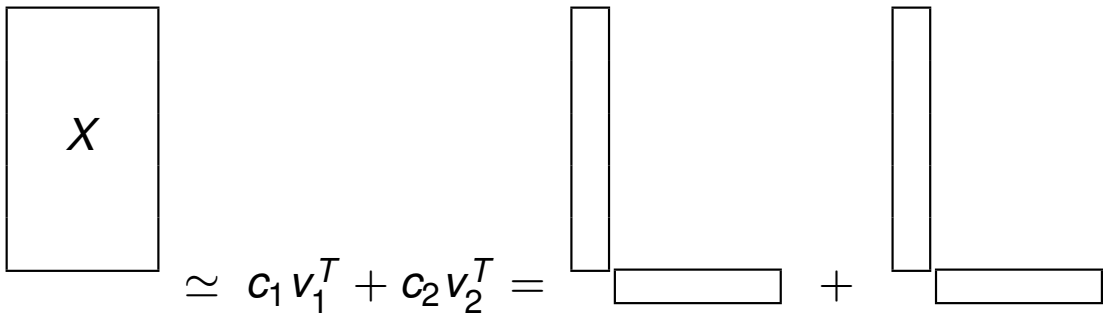
Rappels de calcul matriciel & notations

- Produit scalaire : $\langle x, y \rangle = x^T y = \sum x_i y_i$
- Transposée : $(AB)^T = B^T A^T$
- Trace : $Tr(A) = \sum A_{ii}$. $Tr(BA) = Tr(BA)$
  Exemple : $\langle Mx, Ny \rangle = x^T M^T N y = Tr(x^T M^T N y) = Tr(N y x^T M^T)$
- P projection orthogonale ssi $P = P^2 = P^T$ (idempotente et symétrique).
  Rq : On vérifie facilement qu'alors $(x - Px) \perp Px$
- Matrice orthogonale : $QQ^T = Q^T Q = Id$
- Toute matrice symétrique est diagonalisable dans une base orthogonale :
  $S = QDQ^{-1} = QDQ^T$
- Rang d'une matrice : dimension de l'espace image. $Rg(M) = Rg(M^T)$
- Si $x \in \mathbb{R}^n$ et $y \in \mathbb{R}^p$, $M = xy^T$ est de rang 1. $M_{ij} = x_i y_j$.
- $\|.\|_F$ désigne la norme de Frobenius :

$$\|M\|_F^2 = \sum_{ij} M_{ij}^2 = Tr(M^T M).$$

Point de vue matriciel

Le résultat du problème mathématique donnera une approximation de X par une matrice de rang faible (ici de rang 2) (formée des individus projetés) :



Ligne i : $x_i \simeq c_1(i)v_1^T + c_2(i)v_2^T$

Chaque individu est approché dans le plan Π engendré par les **axes principaux v_1 et v_2, qui seront normés à 1.**

Les variables. L'ACP permet aussi de comprendre les liens entre variables. En effet, l'identité $X \simeq c_1 v_1^T + c_2 v_2^T$ exprime que chaque variable, colonne de X, est approximativement combinaison linéaire de c_1 et c_2.

Réécriture du problème mathématique

En raisonnant sur les vecteurs colonne x_i^T, le critère devient

$$\begin{aligned}
\sum_i \|x_i - P(x_i)\|^2 &= \sum_i (x_i^T - Px_i^T)^T (x_i^T - Px_i^T) \\
&= \sum_i x_i (I-P)(I-P) x_i^T \\
&= \sum_i Tr\big((I-P) x_i^T x_i (I-P)\big) \\
&= Tr\Big(\sum_i (I-P) x_i^T x_i (I-P)\Big) \\
&= Tr\Big((I-P) X^T X (I-P)\Big) \\
&= \|(I-P)X^T\|_F^2
\end{aligned}$$

Le problème en dimension générale

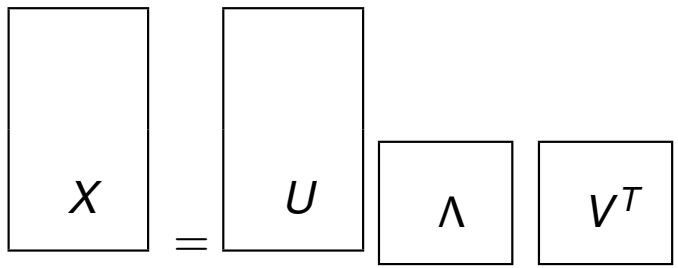
Il s'agit donc de trouver la matrice P de projection orthogonale sur un espace E de dimension k (p.ex. $k = 2$) qui réalise

$$P = \arg \min_{\text{Rang}(P) \leq k} \|X - XP\|_F^2, \tag{1}$$

La décomposition en valeurs singulières sera l'instrument central pour résoudre ce problème.

La SVD (Singular value decomposition)

La décomposition en valeurs singulières permet d'écrire une matrice X sous la forme



où

- U est à colonnes orthonormées : $U^T U = I$
- Λ est diagonale
- V est orthogonale.

La transposition de l'identité informelle ci-dessus donne la forme dans le cas où X est horizontalement allongée.

✓Algorithmes numériquement très efficaces pour réaliser cette décomposition.
✓Transformation de base extensivement utilisée en pratique.

Plus formellement :

Théorème

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. $U^T U = V^T V = Id$) et une matrice diagonale $\Lambda \in \mathbb{R}^{p \times p}$ à entrées positives telles que

$$X = U \Lambda V^T$$

Par conséquent, en appelant u_i et v_i les vecteurs colonne de U et V (vecteurs singuliers à droite et à gauche), on a la décomposition en somme de matrices de rang 1 :

$$X = \sum_{i=1}^{p} \lambda_i u_i v_i^T, \quad \lambda_i = \Lambda_{ii}.$$

La matrice Λ contient nécessairement les racines carrées des valeurs propres de $X^T X$, appelées valeurs singulières, et si ces dernières sont distinctes et rangées par ordre décroissant, cette décomposition est unique.

Démonstration basée sur la diagonalisation de $X^T X$ (noter que $X^T X = V \Lambda^2 V^T$).

Solution du problème

Théorème

On suppose les valeurs singulières distinctes et ordonnées : $\lambda_1 > \lambda_2 > \ldots \lambda_n$. Soit P_k

$$P_k = \sum_{i \leq k} v_i v_i^T$$

la projection orthogonale sur l'espace engendré par les k premiers vecteurs singuliers à droite. Pour toute matrice de projection P orthogonale sur un espace de dimension $\leq k$, on a

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

Remarque. Le lemme reste vrai même si des valeurs singulières sont égales, mais P_k risque de ne plus être unique.
La matrice approchée vaut donc

$$XP = \sum_{i=1}^{k} \lambda_i u_i v_i^T.$$

On a simplement mis à zéro les $p - k$ plus petites valeurs singulières.

On a même plus généralement :

Théorème

Sous les hypothèses du théorème précédent, pour toute matrice A de rang k on a

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

Principe de la démonstration : Vérifier que

$$\begin{aligned}\|X - A\|_F^2 &= \|X(Id - P) + (X - A)P\|_F^2 \\ &= \|X(Id - P)\|_F^2 + \|(X - A)P\|_F^2 \\ &\geq \|X - XP_k\|_F.\end{aligned}$$

Inertie des espaces

On définit l'inertie des individus par la quantité

$$I = \frac{1}{n} \sum_i \|x_i\|^2.$$

Comme les individus sont centrés, cette quantité réelle mesure la **dispersion** des individus dans l'espace à p dimensions. C'est la somme des variances des p variables.

Soit P_E le projecteur orthogonal sur un e.v. $E \subset \mathbb{R}^p$; on définit l'inertie de E :

$$I_E = \frac{1}{n} \sum_i \|P_E x_i\|^2 = \frac{1}{n} \sum_i \|x_i\|^2 - \frac{1}{n} \sum_i \|x_i - P_E x_i\|^2$$

en vertu du théorème de Pythagore. Le problème (1) est donc équivalent à trouver l'espace E_k de dimension k d'inertie maximale.

Minimiser les erreurs $x_i - Px_i$ revient à maximiser la variabilité des Px_i

Outline

Calcul pratique de l'acp.

La matrice X est supposée déjà recentrée.

- Faire une SVD de X : on obtient $X = U\Lambda V^T$ avec $U = [u_1, \dots u_p]$ et $V = [v_1, \dots v_p]$
- **Composantes principales** : $c_k = \lambda_k u_k$ avec $C = [c_1 \dots c_p] = U\Lambda$
- **Axes principaux** : $v_1, \dots v_p$

On a

$$X = CV^T = \sum_i c_i v_i^T$$

Comme $C = XV$, la i-ème composante principale est la combinaison linéaire des variables avec les poids contenus dans la i-ième colonne de V.

La représentation des individus dans un plan se borne alors à ne considérer que c_1 et c_2.

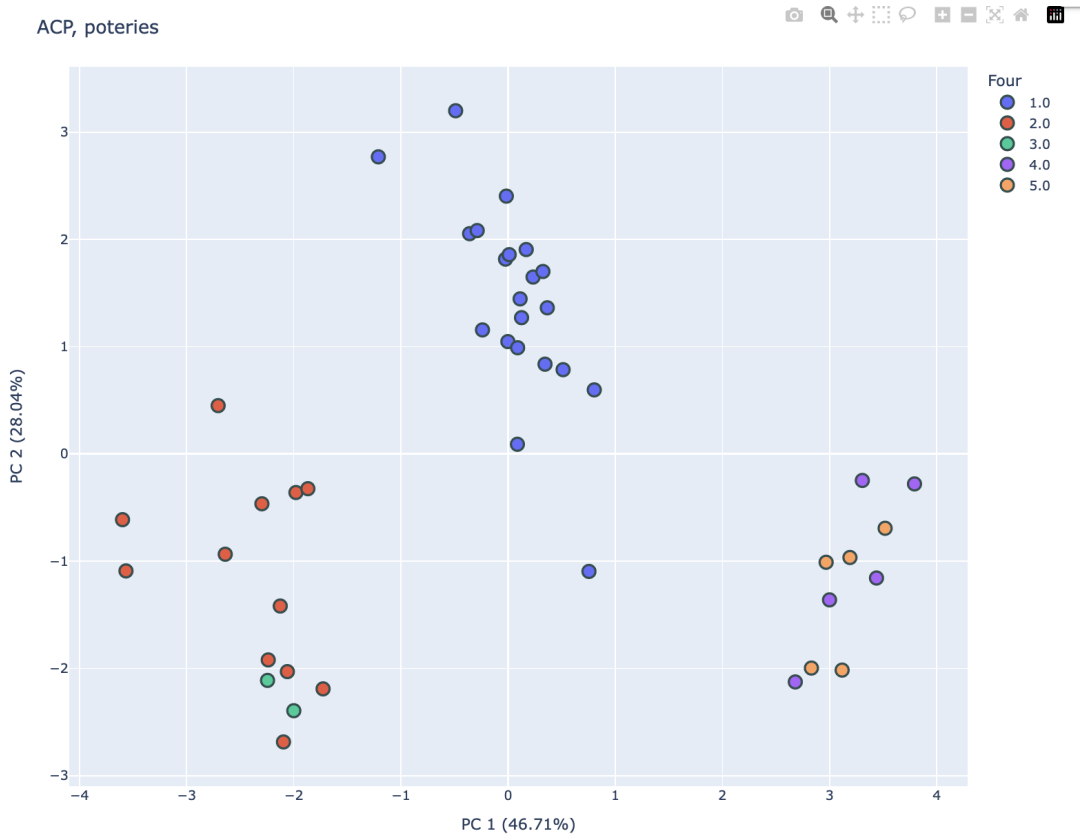
Représentation des poteries, interprétation des axes principaux

Pour interpréter les axes de l'ACP, on peut regarder les coordonnées des variables sur les axes principaux (ie les vecteurs propres de la SVD). On peut aussi les représenter par un diagramme en barres.

	PAL2O3	Fe2O3	MgO	CaO	Na2O	K2O	TiO2	MnO	BaO
PC1	0.35	-0.33	-0.43	-0.06	-0.22	-0.46	0.34	-0.46	-0.02
PC2	0.33	0.40	-0.19	0.50	0.46	-0.02	0.30	0.09	0.38

code

Représentation des poteries



Sur les légendes d'axe, on reporte le pourcentage d'inertie expliqué par chaque composante principale (voir ci-dessous).

La variable «four» n'a pas servi à l'ACP : apprentissage non-supervisé.

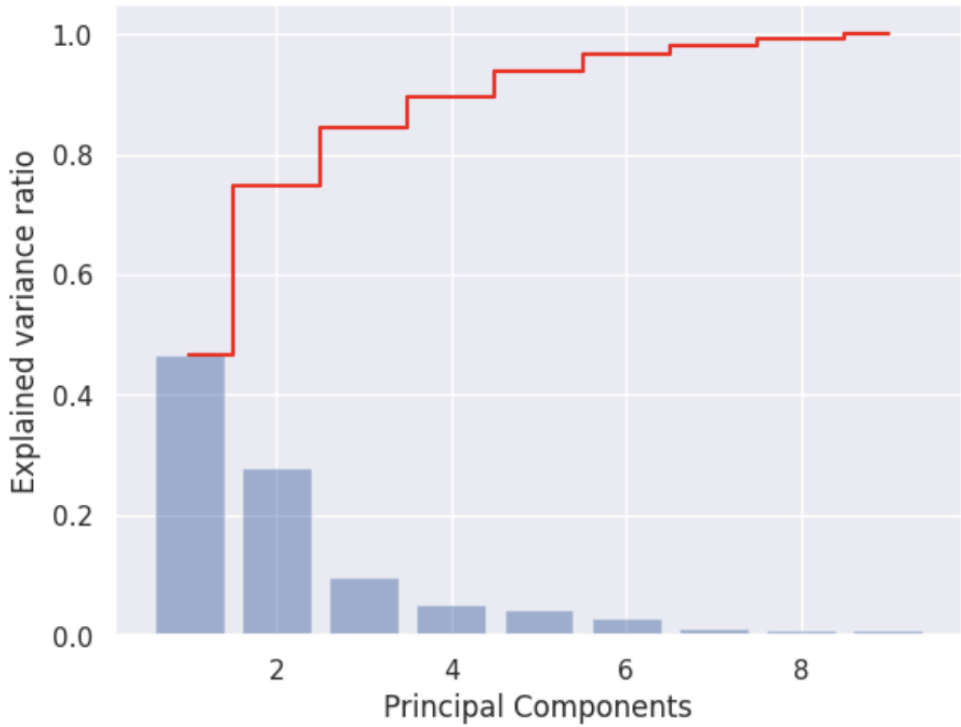
code

Définition

Les v_k sont les **facteurs principaux**, ou axes pricipaux.
Le vecteur $c_k = Xv_k$ est la k-ième **composante principale**. $\|c_k\| = \lambda_k$.
$\frac{\lambda_1^2+\ldots+\lambda_k^2}{\lambda_1^2+\ldots+\lambda_p^2} = \frac{I_{E_k}}{I}$ est la **fraction d'inertie expliquée** par E_k.

Plus la fraction d'inertie expliquée par E_k est proche de 1,
plus la projection de chaque variable $x_{.j}$ sur E_k est proche de $x_{.j}$,
plus les $c_1, \ldots c_k$ représentent bien les individus.

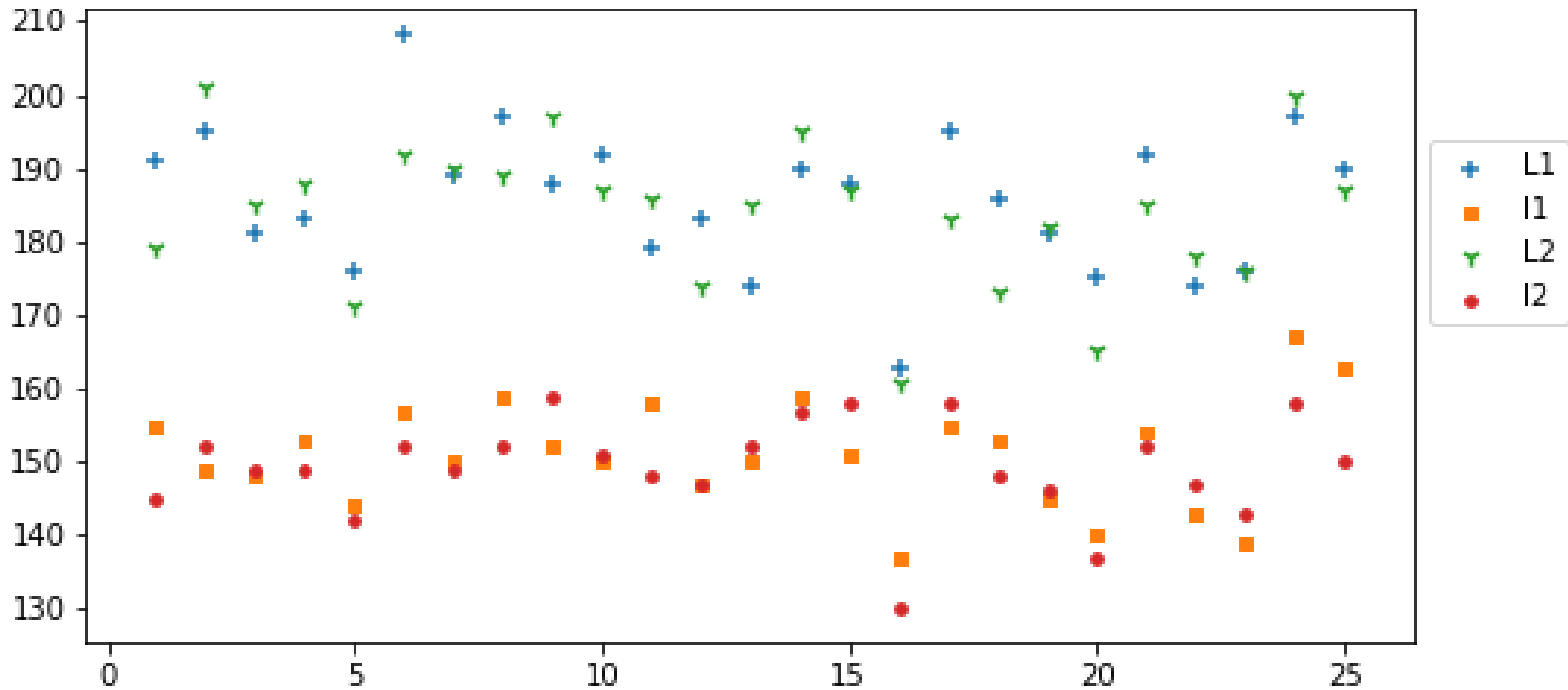
Sur les poteries :



Il faut cependant se méfier : p. ex., sur des données non normalisées, on aura typiquement une distribution plus concentrée de l'inertie.

Un exemple.

On a mesuré la largeur et la longueur de la tête chez 25×2 frères [3].
Ce qui donne 25 individus, et 4 variables : longueur et largeur pour l'un, longueur et largeur pour l'autre.



3. données `frets{boot}` de R.

On obtient sur ces exemples

$$V = \begin{pmatrix} 0,57 & -0,69 & -0,44 & -0,01 \\ 0,41 & -0,22 & 0,87 & -0,17 \\ 0,60 & 0,63 & -0,21 & -0,44 \\ 0,39 & 0,27 & 0,06 & 0,88 \end{pmatrix}, \quad D = \Lambda^2 = \begin{pmatrix} 228 & 0 & 0 & 0 \\ 0 & 29,4 & 0 & 0 \\ 0 & 0 & 17 & 0 \\ 0 & 0 & 0 & 9 \end{pmatrix}$$

Les composantes principales sont en gros

- somme des périmètres des têtes
- différence entre les deux frères
- allongement de la tête du premier
- allongement de celle du second.

Le fait que la première valeur propre sorte du lot (80% d'inertie expliquée) indique que les individus se distinguent surtout par la somme des périmètres des crânes des deux frères.

Normalisation

L'ACP doit être faite sur avec des variables centrées, sinon, si les individus sont dans un nuage écarté de l'origine, la première composante sera essentiellement la distance à l'origine et pourra porter quasiment toute l'inertie si le centre du nuage est loin.

L'ACP n'est pas invariante par changement d'échelle sur les variables.

Si les variables sont non comparables (i.e. m et kg), on normalise :

$$x_{ik} \longleftarrow x_{ik} / ||x_{.k}||$$

afin d'avoir un résultat **indépendant des unités utilisées**.

Si les variables sont comparables on peut préférer de ne pas faire de normalisation. Par exemple si les individus sont les pays et les variables le nombre de personnes ayant telle ou telle profession, on peut vouloir ne pas donner trop de poids dans l'analyse aux professions d'effectif très faible, ce qui sera compromis si l'on normalise. La normalisation serait une forme d'amplification du bruit.

Solution par défaut : normalisation.

Normalisation : l'exemple des têtes

On obtient un résultat un peu différent :

$$V = \begin{pmatrix} 0,49 & -0,48 & -0,72 & 0,07 \\ 0,49 & -0,54 & 0,68 & -0,08 \\ 0,51 & 0,50 & -0,05 & -0,70 \\ 0,51 & 0,48 & 0,09 & 0,71 \end{pmatrix}, \quad D = \Lambda^2 = \begin{pmatrix} 3,2 & 0 & 0 & 0 \\ 0 & 0,38 & 0 & 0 \\ 0 & 0 & 0,27 & 0 \\ 0 & 0 & 0 & 0,16 \end{pmatrix}.$$

V est cette fois très proche de la matrice

$$V = \frac{1}{2}\begin{pmatrix} 1 & -1 & -\sqrt{2} & 0 \\ 1 & -1 & \sqrt{2} & 0 \\ 1 & 1 & 0 & -\sqrt{2} \\ 1 & 1 & 0 & \sqrt{2} \end{pmatrix}.$$

Interprétation de Λ^2. Chaque terme correspond à l'inertie du sous-espace de dimension 1 correspondant, et leur somme fait l'inertie totale (théorème de Pythagore). Sa valeur mesure la contribution de l'axe aux données. Ici, le premier axe contribue à 80% de l'inertie car $3,2/(3,2+0,38+0,27+0,16) = 80\%$. Λ^2 contient les valeurs propres de la matrice de corrélation des données.

Outline

4

Take home messages

Take home messages

Nouvelles variables explicatives $c_j = Xv_j$ (composantes principales) obtenues par combinaison des anciennes. C'est un prétraitement, pas une méthode de prédiction.

Choisies en sorte que la structure géométrique des individus restreints aux premières composantes soit la plus fidèle possible à la géométrie des individus complets.

Ce qui signifie que les premières variables concentrent l'information. Fidélité mesurée par le poids relatif des valeurs propres (inerties).

Ces nouvelles variables sont décorrélées : $c_i \perp c_j$, $i \neq j$.

Il est généralement préférable de travailler sur données standardisées, au minimum centrées.

Le premier axe fait souvent apparaître un effet « taille ».

Dans la SVD de X : $X = U\Lambda V^T$, la matrice V contient les v_i en colonne, et $C = U\Lambda = XV$ contient les c_i en colonne.

$(c_{i1}, \ldots c_{ik})$ exprime, dans la base orthonormée $(v_1, \ldots v_p)$, l'individu $(x_{i1}, \ldots x_{ip})$ projeté sur Vect$(v_1 \ldots v_k)$.