Analyse de Données – ID
L'Analyse en Composantes Principales

Philippe LERAY
philippe.leray@univ-nantes.fr

Equipe COnnaissances et Décision
Laboratoire d'Informatique de Nantes Atlantique – FRE 2729
Site de l'Ecole Polytechnique de l'université de Nantes

Polytech'Nantes
École d'ingénieurs de l'université de Nantes

UNIVERSITÉ DE NANTES

lina LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

Tableau de données

- échantillon de n individus $i_1, \ldots, i_n$
- p variables **quantitatives** $v_1, \ldots, v_p$

	v_1	$\ldots$	v_k	$\ldots$	v_p
i_1	x_{11}	$\ldots$	x_{1k}	$\ldots$	x_{1p}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i_j	x_{j1}	$\ldots$	x_{jk}	$\ldots$	x_{jp}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i_n	x_{n1}	$\ldots$	x_{nk}	$\ldots$	x_{np}

- x_{jk} = valeur de la variable v_k observée sur l'individu i_j
- individu i_j assimilé au vecteur $i_j = (x_{j1}, \ldots, x_{jp})$ de $\mathbb{R}^p$
- variable v_k assimilée au vecteur $v_j = (x_{1k}, \ldots, x_{nk})$ de $\mathbb{R}^n$
- on note X la matrice $(v_1 | \ldots | v_p)$

La matrice des poids

Définition

A chaque individu i_j est associé un poids p_j représentant l'importance relative de l'individu dans l'échantillon

- $p_j \geq 0$, pour tout $j \in \{1, \ldots, n\}$
- $\sum_{j=1}^{n} p_j = 1$

- sauf exception (échantillons redressés, données regroupées) $p_j = 1/n$

- matrice de poids = matrice diagonale D d'ordre n : $D = diag(p_1, \ldots, p_n)$
- $(D = \frac{1}{n} I)$

Centre de gravité

Individu "moyen"

$g = (\bar{v}_1, \ldots, \bar{v}_p)'$, vecteur des moyennes arithmétiques de chaque variable

$\mathbf{1}$ = le vecteur de $\mathbb{R}^n$ dont les composantes sont égales à 1

Proposition

$$g = X'D\mathbf{1}$$

Matrice des données centrées

$$Y = X - \mathbf{1}g' = (I - \mathbf{1}\mathbf{1}'D)X$$

Covariance - corrélation

Proposition : matrice de covariance des variables

$$V = X'DX - gg' = Y'DY$$

$D_{1/s} = diag(1/s_1, \ldots, 1/s_p)$ = matrice diagonale des inverses des écarts-types

Proposition : matrice des données centrées-réduites

$$Z = YD_{1/s}$$

Proposition : matrice R de corrélation linéaire des variables

$$R = D_{1/s}VD_{1/s} = Z'DZ$$

NB: R est aussi la matrice de covariance de Z

Espace des individus

- un individu = point défini par p coordonnées = vecteur de l'e.v. $F = \mathbb{R}^p$ muni de sa base canonique.
- F = *espace des individus*.
- ensemble des n individus = nuage de points dans F (g en est le centre de gravité)

- choix d'une métrique M pour mesurer la distance entre deux individus i_j et i_k :

$$d^2_M(i_j, i_k) = < i_j - i_k, i_j - i_k >_M = (i_j - i_k)' M (i_j - i_k).$$

  - $M = I$: privilégier les variables les plus dispersées
  - $M = D_{1/s^2} = diag(1/s_1^2, \ldots, 1/s_p^2)$: la plus courante = donner à chaque variable la même importance

Inertie

Définition : inertie totale du nuage de points

$$I_F = \sum_{j=1}^{n} p_j \| i_j - g \|_M^2$$

Proposition

$$I_F = Trace(MV) = Trace(VM)$$

Proposition

Si $M = I$, inertie = somme des variances des p variables

Proposition

Si $M = D_{1/s^2}$, inertie = nombre de variables

Espace des variables

- variable v_j = vecteur de l'e.v. $E = \mathbb{R}^n$ muni de sa base canonique
- E est appelé *espace des variables*

- choix d'une métrique pour mesurer la proximité entre deux variables = matrice D des poids

Propositions

- $< v_j, v_k >_D = \text{cov}(v_j, v_k)$ si les deux variables sont centrées
- $\|v_j\|_D =$ écart-type s_j de la variable
- l'angle $\theta(v_j, v_k)$ entre deux variables centrées v_j et v_k est donné par :

$$\cos\theta(v_j, v_k) = r(v_j, v_k).$$

Projections des individus

- but de l'ACP = trouver une représentation approchée du nuage des n individus dans F_k (s.e.v. de F de dimension k faible)
- choix de F_k = déformer le moins possible la réalité, i.e. maximiser l'inertie du nuage projeté

- $P \in \mathcal{M}_p(\mathbb{R})$ = matrice de projection M-orthogonale sur F_k

Propositions

- le nuage projeté est associé à la matrice de données YP'
- la matrice de covariance du tableau YP' est $(YP')'D(YP') = PVP'$
- l'inertie du nuage de projeté vaut $Trace(PVP'M)$
- $Trace(PVP'M) = Trace(VMP)$

Projections des individus

Proposition

Soit G un $\mathbb{R}$-e.v. et soient G_1 et G_2 deux s.e.v. orthogonaux de G. Notons $I_{G_1 \oplus G_2}$ (resp. I_{G_1}, resp. I_{G_2}) l'inertie du nuage de points projeté sur $G_1 \oplus G_2$ (resp. G_1, resp. G_2). Alors, $I_{G_1 \oplus G_2} = I_{G_1} + I_{G_2}$.

Théorème "des solutions emboîtées"

Soit F_k le s.e.v. de F de dimension k portant l'inertie maximale. Alors, le s.e.v. de dimension $k+1$ portant l'inertie maximale est la somme directe de F_k et du s.e.v. de dimension 1 M-orthogonal à F_k portant l'inertie maximale.

$\Rightarrow$ ACP = trouver l'axe portant le plus d'inertie, et ainsi de suite

... mais comment le trouver ?

Axes et vecteurs principaux

- soit a vecteur caractéristique d'un axe,
- matrice du projecteur M-orthogonal sur $a = P = \frac{1}{\|a\|_M^2} aa'M$.
- inertie du nuage projeté
  $Trace(VMP) = Trace(VM\frac{1}{\|a\|_M^2} aa'M) =$
  $\frac{1}{\|a\|_M^2} Trace(VMaa'M) = \frac{1}{\|a\|_M^2} Trace(a'MVMa) = \frac{a'MVMa}{a'Ma}$
- le vecteur a maximisant $\frac{a'MVMa}{a'Ma}$ est le vecteur propre de VM de plus grande valeur propre

Théorème

F_k est engendré par les k vecteurs propres de VM associés aux k plus grandes valeurs propres

vecteurs principaux = vecteurs propres de VM, M-normés à 1.

Facteurs principaux

Définition

Aux vecteurs principaux $a_1, \ldots, a_k$ (M-normés à 1), on peut associer les vecteurs $u_1 = Ma_1, \ldots, u_k = Ma_k$ de $\mathbb{R}^p$ appelés *facteurs principaux*.

Proposition

Les u_i sont les vecteurs propres de MV, M^{-1} normés à 1.

Composantes principales

Définition

Les *composantes principales* $c_1, \ldots, c_p$ sont les nouvelles variables (vecteurs de $E = \mathbb{R}^n$) définies par les facteurs principaux : $c_i = Xu_i$
Chaque c_i est le vecteur de $E = \mathbb{R}^n$ refermant les coordonnées des projections M-orthogonales des individus sur l'axe engendré par a_i (M-normé à 1).

Propositions

- la variance d'une composante principale c_i est égale à la valeur propre λ_i du vecteur principal a_i correspondant
- les composantes principales sont D-orthogonales 2 à 2, i.e., elles sont non corrélées linéairement

En pratique, on calcule les u_i par diagonalisation de MV, puis on obtient les $c_i = Xu_i$.