

Variance, covariance, corrélation dans $\mathbb{R}^p$

Cours 8 – Modèles en biologie

Christophe Ambroise

L3 GBI – Université d'Évry

Centre de gravité et inertie

$$\mathbf{g} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i, \quad I = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{g}\|^2 = \sum_{j=1}^p s_j^2$$

Huygens : $I = W + B$, inertie intra-classe plus inertie inter-classe.

Acquis des cours précédents

k-means et classification hiérarchique

k-means : affectation au centre le plus proche, recentrage ; W décroît jusqu'à stabilité.

Classification ascendante hiérarchique : fusions successives, dendrogramme ; UPGMA pour les arbres phylogénétiques.

Entropie et information

$$H(\mathbf{p}) = - \sum_{i=1}^K p_i \log_2 p_i, \quad 0 \leq H \leq \log_2 K$$

Contenu en information d'une position : $R = 2 - H$. Information mutuelle : $I(X; Y) = H(X) + H(Y) - H(X, Y)$.

Le tableau de données

Loi jointe comme matrice

Loi jointe de deux positions : une matrice J ; marges $J \mathbf{1}$ et $J^T \mathbf{1}$.

Indépendance : $J = \mathbf{p} \mathbf{q}^T$, une colonne fois une ligne, matrice de rang 1.

Matrice des données

n individus, p variables : une matrice X de taille $n \times p$.

Lecture par lignes : l'individu i , un point $\mathbf{x}_i \in \mathbb{R}^p$. Un nuage de n points.

Lecture par colonnes : la variable j , un vecteur $\mathbf{x}^j \in \mathbb{R}^n$. Un nuage de p vecteurs.

Centre de gravité

Le vecteur des moyennes des variables.

$$\mathbf{g} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} X^T \mathbf{1}, \quad g_j = \bar{x}_j$$

Un produit matrice-vecteur : la somme des lignes de X , divisée par n .

Un petit tableau

Quatre individus, deux variables :

$$X = \begin{pmatrix} 2 & 1 \\ 4 & 3 \\ 6 & 2 \\ 8 & 6 \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \quad X_c = \begin{pmatrix} -3 & -2 \\ -1 & 0 \\ 1 & -1 \\ 3 & 3 \end{pmatrix}$$

Tableau illustratif.

Matrice centrée

Retirer à chaque individu le centre de gravité.

$$X_c = X - \mathbf{1} \mathbf{g}^T$$

$\mathbf{1} \mathbf{g}^T$: n lignes toutes égales à $\mathbf{g}^T$, une matrice de rang 1 (C7).

$X_c^T \mathbf{1} = X^T \mathbf{1} - \mathbf{g} (\mathbf{1}^T \mathbf{1}) = n \mathbf{g} - n \mathbf{g} = \mathbf{0}$: colonnes centrées, orthogonales à $\mathbf{1}$.

9

10

Variance et covariance

11

12

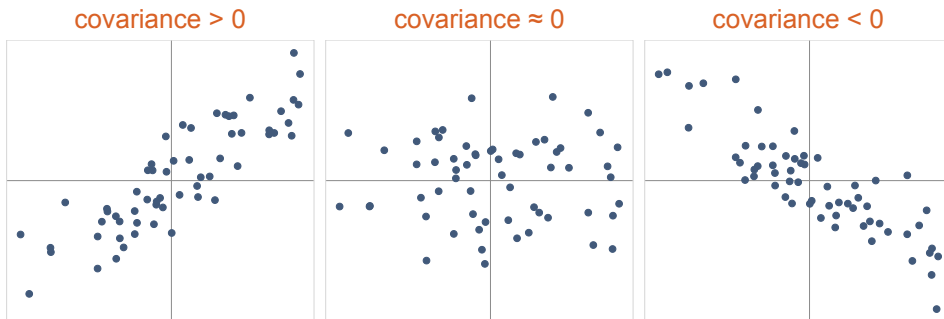
Variance d'une variable

$$s_j^2 = \frac{1}{n} \|\mathbf{x}_c^j\|^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

Carré de l'écart-type du C5 ; somme des variances = inertie (C6).

Petit tableau : $s_1^2 = \frac{1}{4} (9 + 1 + 1 + 9) = 5$, $s_2^2 = \frac{1}{4} (4 + 0 + 1 + 9) = 3,5$.

Signe de la covariance



Positive : les deux variables s'écartent de leur moyenne dans le même sens.
Négative : en sens contraires.

Données simulées, centrées sur les axes.

Covariance

La moyenne des produits des écarts à la moyenne de deux variables : un produit scalaire de vecteurs centrés.

$$s_{jk} = \frac{1}{n} \mathbf{x}_c^j \cdot \mathbf{x}_c^k = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

$s_{jj} = s_j^2$: la covariance d'une variable avec elle-même est sa variance.

Petit tableau : $s_{12} = \frac{1}{4} (6 + 0 - 1 + 9) = 3,5$.

Covariance et unités

Kleiber, unités brutes : masse et métabolisme.

Masse en kg : $s_{12} \approx 394\,493$ kg·kcal/jour. Masse en g : 1 000 fois plus.

Unités

Covariance : l'unité du produit des deux variables ; sa valeur dépend des unités choisies.

Corrélation

Invariance par changement d'unité

Multiplier une variable par $a > 0$: covariance et écart-type multipliés par a , corrélation inchangée.

Kleiber, unités brutes : $r \approx 0,984$, masse en kg comme en g.

Kleiber, logarithmes : $r \approx 0,998$, un nuage presque aligné.

Corrélation

La covariance divisée par le produit des écarts-types : le cosinus des vecteurs centrés (C5).

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\mathbf{x}_c^j \cdot \mathbf{x}_c^k}{\|\mathbf{x}_c^j\| \|\mathbf{x}_c^k\|} \in [-1, 1]$$

Variables non constantes : $s_j > 0, s_k > 0$.

Petit tableau : $r_{12} = \frac{3,5}{\sqrt{5} \times 3,5} \approx 0,84$.

Limites de la corrélation

$\mathbf{x} = (-2, -1, 0, 1, 2), \mathbf{y} = (4, 1, 0, 1, 4) : y_i = x_i^2$.

$\bar{y} = 2, \mathbf{y}_c = (2, -1, -2, -1, 2) : \mathbf{x}_c \cdot \mathbf{y}_c = -4 + 1 + 0 - 1 + 4 = 0$, donc $r = 0$.

Corrélation nulle

Pas de lien linéaire, et non pas absence de lien : ici $\mathbf{y}$ est entièrement déterminé par $\mathbf{x}$. L'information mutuelle (C7) détecte cette dépendance.

Matrice de covariance

Symétrie et diagonale

$$S^T = \frac{1}{n} (X_c^T X_c)^T = \frac{1}{n} X_c^T X_c = S$$

Transposée d'un produit : $(AB)^T = B^T A^T$ (C3).

Diagonale : les variances. Petit tableau :

$$S = \begin{pmatrix} 5 & 3,5 \\ 3,5 & 3,5 \end{pmatrix}$$

Matrice de covariance

La matrice des covariances de toutes les paires de variables.

$$S = \frac{1}{n} X_c^T X_c, \quad S_{jk} = s_{jk}$$

Coefficients : produits scalaires des colonnes de X_c , comme $X^T X$ dans les équations normales du C4.

Trace et inertie

La trace d'une matrice carrée : la somme de ses coefficients diagonaux. Pour S , la somme des variances, c'est-à-dire l'inertie du nuage.

$$\text{tr}(S) = \sum_{j=1}^p s_{jj} = \sum_{j=1}^p s_j^2 = I$$

Petit tableau : $\text{tr}(S) = 8,5$. Kleiber, logarithmes :

$$S = \begin{pmatrix} 1,071 & 0,810 \\ 0,810 & 0,614 \end{pmatrix}, \quad \text{tr}(S) = 1,685$$

Matrice des corrélations

Les corrélations de toutes les paires de variables : la matrice de covariance des variables réduites.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad R = \frac{1}{n} Z^T Z, \quad R_{jk} = r_{jk}$$

Diagonale de 1 ; aucune unité (réduction du C5).

Projection du nuage sur un axe

Corrélations des iris

	Long. sépale	Larg. sépale	Long. pétale	Larg. pétale
Long. sépale	1	-0,12	0,87	0,82
Larg. sépale	-0,12	1	-0,43	-0,37
Long. pétale	0,87	-0,43	1	0,96
Larg. pétale	0,82	-0,37	0,96	1

Longueur et largeur des pétales : $r = 0,96$. Largeur des sépales : corrélée négativement aux trois autres mesures.

150 iris, quatre mesures en cm ; jeu [iris](#) de R (Anderson, 1935).

Coordonnées sur un axe

Axe dirigé par $\mathbf{u}$ unitaire : coordonnées des n individus centrés (C4).

$$\mathbf{c} = X_c \mathbf{u} \in \mathbb{R}^n$$

Moyenne nulle : $\frac{1}{n} \mathbf{1}^T X_c \mathbf{u} = \frac{1}{n} (X_c^T \mathbf{1})^T \mathbf{u} = 0$.

Variance projetée

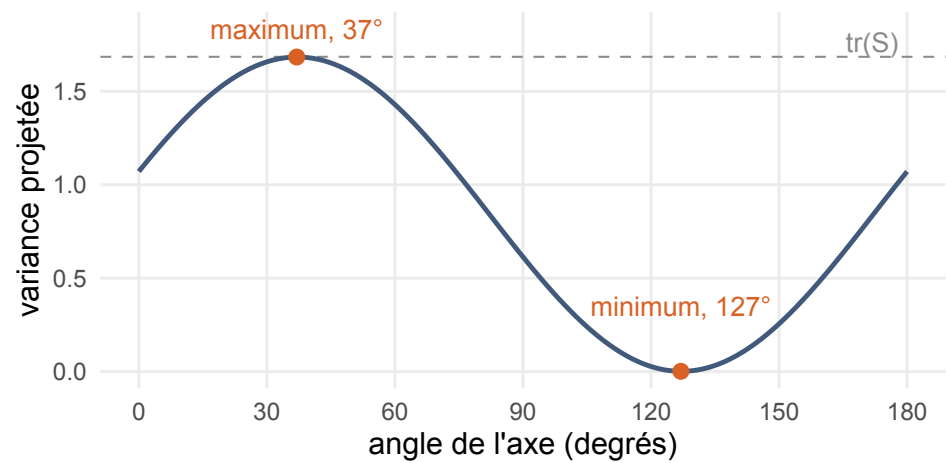
La variance des coordonnées sur un axe unitaire s'écrit avec la matrice de covariance.

$$\frac{1}{n} \|X_c \mathbf{u}\|^2 = \mathbf{u}^T S \mathbf{u}, \quad \|\mathbf{u}\| = 1$$

Démonstration : $\frac{1}{n} (X_c \mathbf{u})^T (X_c \mathbf{u}) = \mathbf{u}^T (\frac{1}{n} X_c^T X_c) \mathbf{u} = \mathbf{u}^T S \mathbf{u}.$ $\square$

$\mathbf{u} = \mathbf{e}_j$: $\mathbf{u}^T S \mathbf{u} = s_j^2$. Toujours $\mathbf{u}^T S \mathbf{u} \geq 0$: une variance.

Courbe de la variance projetée



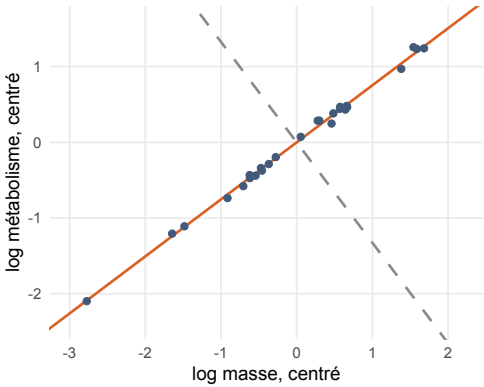
Variance projetée selon l'angle

Kleiber, logarithmes : $\mathbf{u} = (\cos \theta, \sin \theta)$.

θ	0°	45°	90°	37°	127°
$\mathbf{u}^T S \mathbf{u}$	1,071	1,652	0,614	1,684	0,001

$\theta = 0^\circ$: variance du log de la masse. $\theta = 90^\circ$: variance du log du métabolisme.

Axes de plus grande et de plus petite variance



Orange : l'axe vers 37°, le long du nuage. Pointillés : l'axe perpendiculaire.

Deux axes orthogonaux

Sur deux axes orthogonaux du plan, les variances projetées s'additionnent et redonnent l'inertie.

$$\mathbf{u} \perp \mathbf{v}, \|\mathbf{u}\| = \|\mathbf{v}\| = 1 \implies \mathbf{u}^T S \mathbf{u} + \mathbf{v}^T S \mathbf{v} = \text{tr}(S)$$

$$\mathbf{u} = (c, s), \mathbf{v} = (-s, c), S = \begin{pmatrix} a & b \\ b & d \end{pmatrix} : \\ (ac^2 + 2bcs + ds^2) + (as^2 - 2bcs + dc^2) = a + d. \quad \square$$

Kleiber : 30° et 120° donnent $1,658 + 0,027 = 1,685$. En dimension p , pour p axes orthogonaux : admis.

33

Question ouverte

Axe de variance maximale

Problème d'optimisation

Trouver l'axe unitaire $\mathbf{u}$ qui maximise $\mathbf{u}^T S \mathbf{u}$.

Kleiber : un axe vers 37° porte **1,684** sur **1,685**, soit **99,9 %** de l'inertie. Un nuage de $\mathbb{R}^2$ presque réduit à une droite.

Pour un nuage de $\mathbb{R}^p$: quel axe, et quelle part d'inertie ?

35