

Analyse en composantes principales

Cours 9 – Modèles en biologie

Christophe Ambroise

L3 GBI – Université d'Évry

Entropie et information

$H(\mathbf{p}) = -\sum_i p_i \log_2 p_i$: incertitude d'une loi, maximale pour la loi uniforme.

Loi normale : à variance fixée, entropie maximale ; garder de la variance, c'est garder de l'information.

Acquis des cours précédents

Loi jointe comme matrice

Loi jointe J ; marges $J\mathbf{1}$ et $J^T\mathbf{1}$; indépendance $J = \mathbf{p}\mathbf{q}^T$, de rang 1.

Information mutuelle : $I(X;Y) = H(X) + H(Y) - H(X,Y)$, nulle si et seulement si indépendance.

Matrice de covariance

$$S = \frac{1}{n} X_c^T X_c, \quad S^T = S, \quad \text{tr}(S) = \sum_j s_j^2 = I$$

Matrice des corrélations : $R = \frac{1}{n} Z^T Z$, variables réduites.

Axe de variance maximale

Variance projetée

Axe unitaire $\mathbf{u}$: coordonnées $X_c \mathbf{u}$, variance projetée $\mathbf{u}^T S \mathbf{u}$.

Kleiber : maximum **1,684** vers 37° , sur une inertie de **1,685**.

Question laissée ouverte : quel axe maximise $\mathbf{u}^T S \mathbf{u}$?

Axe de variance maximale

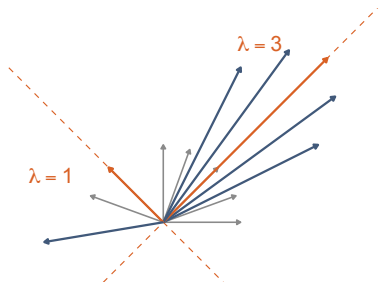
Parmi les axes unitaires, celui de variance projetée maximale est dirigé par un vecteur propre de S ; la variance maximale est la plus grande valeur propre.

$$\max_{\|\mathbf{u}\|=1} \mathbf{u}^T S \mathbf{u} = \lambda_1, \quad \text{atteint pour } S \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

Démontré ici en dimension 2 ; admis en dimension p .

Directions conservées par une matrice

Vecteur propre (C3) : $A\mathbf{v} = \lambda\mathbf{v}$, $\mathbf{v} \neq \mathbf{0}$. Canal ionique : π et $(1, -1)$, vérifiés, jamais cherchés.



$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

$$A(1, 1) = 3(1, 1)$$

$$A(1, -1) = 1(1, -1)$$

Vecteur propre : une direction que la matrice ne fait pas tourner.

Gris : vecteurs unitaires $\mathbf{u}$. Bleu : leurs images $A\mathbf{u}$. Orange : les deux directions conservées.

Pivot et lignes proportionnelles

$$M = \begin{pmatrix} p & q \\ r & s \end{pmatrix}, p \neq 0 : L_2 \leftarrow L_2 - \frac{r}{p} L_1.$$

$$\begin{pmatrix} p & q \\ r & s \end{pmatrix} \longrightarrow \begin{pmatrix} p & q \\ 0 & \frac{ps - qr}{p} \end{pmatrix}$$

$p = 0$: échanger les deux lignes, même critère.

Critère

Ligne nulle, donc variable libre, si et seulement si $ps - qr = 0$.

Vecteur propre et système homogène

$$A\mathbf{v} = \lambda\mathbf{v} \iff (A - \lambda I)\mathbf{v} = \mathbf{0}$$

Un système homogène (C3) : toujours la solution $\mathbf{v} = \mathbf{0}$. Un vecteur propre existe si et seulement s'il reste une variable libre (C2).

$$\lambda = 3 : A - 3I = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, \text{ lignes proportionnelles, } v_2 \text{ libre : } \mathbf{v} = (1, 1).$$

$$\lambda = 2 : A - 2I = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \text{ seule solution } \mathbf{0} : 2 \text{ n'est pas valeur propre.}$$

Déterminant 2×2

Le nombre qui décide si le pivot laisse une ligne nulle.

$$\det \begin{pmatrix} p & q \\ r & s \end{pmatrix} = ps - qr$$

$\det M \neq 0$: solution unique ; pour un système homogène, la seule solution $\mathbf{0}$.

$\det M = 0$: colonnes colinéaires, une variable libre.

$$\det \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 3, \quad \det \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = 0.$$

Équation des valeurs propres

λ est une valeur propre quand $A - \lambda I$ laisse une variable libre.

$$\det(A - \lambda I) = (a - \lambda)(d - \lambda) - bc = \lambda^2 - \text{tr}(A)\lambda + \det(A) = 0$$

$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$: $\lambda^2 - 4\lambda + 3 = 0$, $\lambda = 1$ ou $\lambda = 3$: les deux directions de la figure.

Méthode : les valeurs propres par cette équation, puis chaque vecteur propre par le pivot de Gauss.

Plus grande valeur propre, variance maximale

$$S = \begin{pmatrix} a & b \\ b & d \end{pmatrix} : \lambda^2 - (a + d)\lambda + (ad - b^2) = 0.$$

$$\lambda_1 = \frac{a + d}{2} + \sqrt{\left(\frac{a + d}{2}\right)^2 - (ad - b^2)} = \frac{a + d}{2} + \sqrt{\left(\frac{a - d}{2}\right)^2 + b^2}$$

Exactement le maximum obtenu par Cauchy-Schwarz.

Axe de variance maximale

Le vecteur propre de la plus grande valeur propre.

Variance projetée en dimension 2

$$\mathbf{u} = (\cos \theta, \sin \theta), S = \begin{pmatrix} a & b \\ b & d \end{pmatrix} :$$

$$\begin{aligned} \mathbf{u}^T S \mathbf{u} &= a \cos^2 \theta + 2b \cos \theta \sin \theta + d \sin^2 \theta \\ &= \frac{a+d}{2} + \frac{a-d}{2} \cos 2\theta + b \sin 2\theta \end{aligned}$$

Cauchy-Schwarz (C4) sur $\left(\frac{a-d}{2}, b\right) \cdot (\cos 2\theta, \sin 2\theta)$:

$$\mathbf{u}^T S \mathbf{u} \leq \frac{a+d}{2} + \sqrt{\left(\frac{a-d}{2}\right)^2 + b^2}$$

avec égalité pour une direction θ bien choisie.

Valeurs propres de Kleiber

$$S = \begin{pmatrix} 1,071 & 0,810 \\ 0,810 & 0,614 \end{pmatrix}, \quad \text{tr}(S) = 1,685, \quad \det(S) \approx 0,0021$$

$$\lambda_1 \approx 1,684, \quad \lambda_2 \approx 0,0013$$

Le maximum et le minimum de la courbe du C8.

Vecteur propre par le pivot de Gauss

$$S - \lambda_1 I = \begin{pmatrix} -0,613 & 0,810 \\ 0,810 & -1,070 \end{pmatrix}$$

$L_2 \leftarrow L_2 + \frac{0,810}{0,613} L_1$: ligne nulle, u_2 libre (C2, C3).

$-0,613 u_1 + 0,810 u_2 = 0$: $u_2 = 0,757 u_1$. Normé : $\mathbf{u}_1 = (0,797; 0,603)$, un axe à 37° .

Valeurs propres et inertie

La somme des valeurs propres est la trace : l'inertie se répartit entre les axes principaux.

$$\lambda_1 + \dots + \lambda_p = \text{tr}(S) = I$$

Kleiber : $1,684 + 0,001 = 1,685$. Part d'inertie de l'axe k : $\lambda_k / \text{tr}(S)$.

Axes principaux orthogonaux

Pour une matrice symétrique : valeurs propres réelles, vecteurs propres de valeurs propres distinctes orthogonaux.

$$S^T = S, \quad S\mathbf{u}_k = \lambda_k \mathbf{u}_k, \quad S\mathbf{u}_l = \lambda_l \mathbf{u}_l, \quad \lambda_k \neq \lambda_l \implies \mathbf{u}_k \perp \mathbf{u}_l$$

Admis. Covariance : valeurs propres positives ou nulles.

Kleiber : $\mathbf{u}_2 = (-0,603; 0,797) \perp \mathbf{u}_1$.

17

18

ACP normée

19

20

ACP normée

Variables réduites (C5) : matrice des corrélations au lieu de la matrice de covariance.

$$R = \frac{1}{n} Z^T Z, \quad \text{tr}(R) = p$$

Changer d'unité ne change plus rien. Chaque variable pèse une inertie de 1.

Les crabes

200 crabes *Leptograpsus* : deux formes (bleue, orange), deux sexes, 50 par groupe.

Mesure (mm)	FL	RW	CL	CW	BD
	lobe frontal	largeur arrière	longueur carapace	largeur carapace	épaisseur

ACP normée sur les 5 mesures ; forme et sexe gardés à part.

Campbell et Mahon, *Australian Journal of Zoology*, 1974 ; jeu `crabs` du paquet R `MASS`.

Axes et composantes principales

Les axes principaux sont les vecteurs propres de R ; les composantes principales, les coordonnées des individus sur ces axes.

$$R \mathbf{u}_k = \lambda_k \mathbf{u}_k \quad (\lambda_1 \geq \dots \geq \lambda_p), \quad \mathbf{c}_k = Z \mathbf{u}_k \in \mathbb{R}^n$$

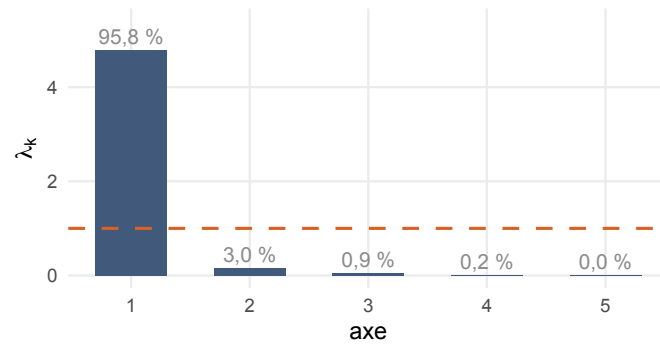
Variance de $\mathbf{c}_k$: λ_k . Composantes non corrélées : $\mathbf{u}_k^T R \mathbf{u}_l = \lambda_l \mathbf{u}_k \cdot \mathbf{u}_l = 0$.

Parts d'inertie

Axe	1	2	3	4	5
λ_k	4,79	0,15	0,05	0,01	0,002
Part (%)	95,8	3,0	0,9	0,2	0,03
Cumul (%)	95,8	98,8	99,7	99,97	100

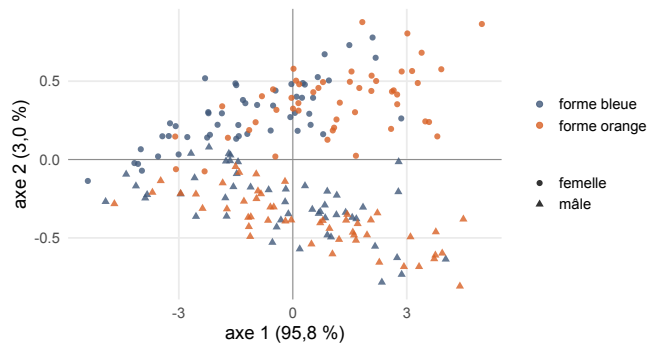
Somme des λ_k : $p = 5$. Un premier axe écrasant.

Éboulis des valeurs propres



Choisir le nombre d'axes : coude de l'éboulis, ou $\lambda_k > 1$ (au-dessus d'une variable). Ici, un seul axe retenu : un piège.

Plan principal des crabes



Axe 1 : des petits aux grands crabes. Axe 2 : femelles en haut, mâles en bas.

Échelles différentes sur les deux axes. Forme et sexe : non utilisés dans le calcul.

Nuage des individus

Coordonnées d'un individu

Coordonnée de l'individu i sur l'axe k : le produit scalaire de son vecteur réduit avec l'axe (C4).

$$c_{ik} = \mathbf{z}_i \cdot \mathbf{u}_k$$

Crabe 51 (forme bleue, femelle) : $c_{i1} = -5,39$, le plus à gauche : le plus petit crabe.

Qualité de représentation

La part du carré de la distance au centre conservée par la projection : le carré du cosinus de l'angle entre l'individu et son projeté (C4).

$$\cos_{ik}^2 = \frac{c_{ik}^2}{\|\mathbf{z}_i\|^2}, \quad \cos_{i, \text{plan}}^2 = \cos_{i1}^2 + \cos_{i2}^2 \in [0, 1]$$

Crabe 51 : $\cos^2 = 1,00$ sur le plan. Crabe 123 (forme orange, mâle) : **0,47**, avec $\|\mathbf{z}_i\|^2 = 0,13$, tout près du centre.

Contribution d'un individu

La part de l'inertie d'un axe apportée par un individu.

$$\text{ctr}_{ik} = \frac{c_{ik}^2}{n \lambda_k}, \quad \sum_{i=1}^n \text{ctr}_{ik} = 1$$

Contribution moyenne : $1/n = 0,5 \%$. Crabe 51 : **3,0 %** de l'axe 1. Crabe 183 (forme orange, femelle) : **2,5 %** de l'axe 2.

Une contribution forte et isolée : un axe construit par un seul individu, à surveiller.

Piège de lecture

Deux individus proches sur le plan : proches dans $\mathbb{R}^p$ seulement s'ils sont bien représentés.

Crabes : $\cos^2$ médian **0,98** sur le plan ; 5 % des crabes sous **0,76**, tous près du centre.

Règle

Vérifier le $\cos^2$ avant d'interpréter une proximité sur le plan.

Éléments supplémentaires

Variables supplémentaires : forme et sexe, non utilisés dans le calcul, révélés par les couleurs du plan.

Individus supplémentaires : réduits avec les moyennes et écarts-types du jeu, puis projetés, $\mathbf{z} \cdot \mathbf{u}_k$, sans modifier les axes.

Nuage des variables

Coordonnée d'une variable

La coordonnée d'une variable sur un axe : sa corrélation avec la composante principale.

$$r(\mathbf{z}^j, \mathbf{c}_k) = \sqrt{\lambda_k} u_{jk}$$

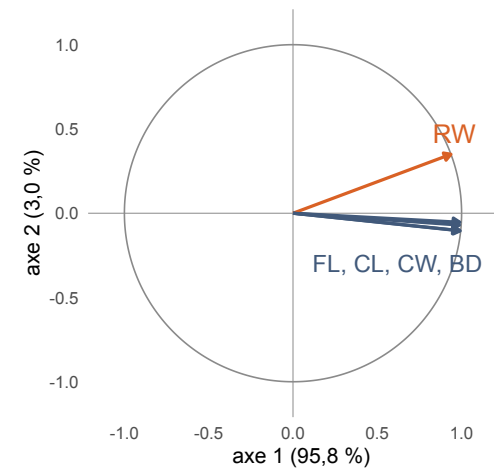
Démonstration : $\text{cov}(\mathbf{z}^j, \mathbf{c}_k) = \frac{1}{n} (Z^T Z \mathbf{u}_k)_j = (R \mathbf{u}_k)_j = \lambda_k u_{jk}$;
variances 1 et λ_k . $\square$

Variables, vecteurs de $\mathbb{R}^n$

Lecture par colonnes (C8) : la variable réduite $\mathbf{z}^j \in \mathbb{R}^n$, de norme $\sqrt{n}$.

Corrélation de deux variables : le cosinus de leurs vecteurs (C5).

Cercle des corrélations



Lecture du cercle

Flèche proche du cercle : variable bien représentée sur le plan, $\cos^2 = r_{j1}^2 + r_{j2}^2$.
Crabes : de **0,98** à **1**.

Deux flèches bien représentées : l'angle entre elles traduit leur corrélation (cosinus, C4 et C5). Flèches groupées : variables corrélées.

Direction d'une flèche : le sens de l'axe. Axe 1 : toutes les mesures vers la droite.

Individus et variables ensemble

Axe 2 : RW du côté positif, femelles du côté positif.

À taille égale, les femelles ont une largeur arrière plus grande : l'abdomen qui porte les œufs.

Interpréter un axe : croiser le cercle des corrélations et le plan des individus.

Contributions des variables

Contribution de la variable j à l'axe k : u_{jk}^2 , de somme 1 sur les variables.

	FL	RW	CL	CW	BD
Axe 1 (%)	20,4	18,3	20,5	20,4	20,4
Axe 2 (%)	1,9	80,7	7,2	3,3	7,0

Lecture des axes

Axe 1 : toutes les mesures, à parts presque égales. Axe 2 : la largeur arrière.

Effet taille

Corrélations toutes fortes

	FL	RW	CL	CW	BD
FL	1	0,91	0,98	0,96	0,99
RW	0,91	1	0,89	0,90	0,89
CL	0,98	0,89	1	0,99	0,98
CW	0,96	0,90	0,99	1	0,97
BD	0,99	0,89	0,98	0,97	1

Un grand crabe est grand sur toutes les mesures.

Axe 2, le sexe

Moyennes des groupes sur l'axe 2 :

	femelles	mâles
forme bleue	0,29	-0,28
forme orange	0,38	-0,39

3 % de l'inertie, une séparation nette des sexes : la forme de l'abdomen.

Premier axe, facteur de taille

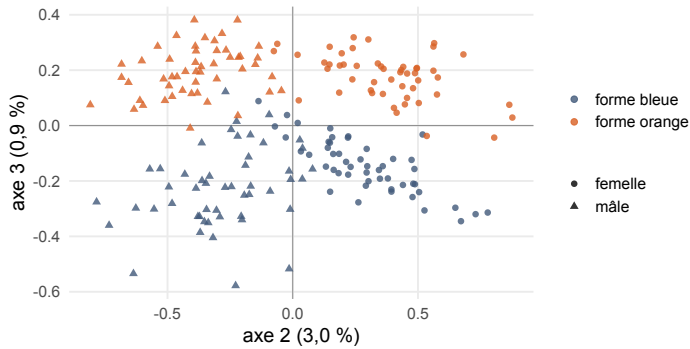
Coefficients de u_1 : **0,43 à 0,45**, tous positifs. La composante c_1 : presque la somme des mesures réduites.

Facteur commun

L'axe 1 mesure la taille. Il porte 95,8 % de l'inertie sans rien dire de la forme.

Moyennes sur l'axe 1 : forme orange plus grande que la forme bleue ; la taille mélange espèce et croissance.

Axe 3, la forme



0,9 % de l'inertie : les deux formes se séparent. Coefficients : CW **-0,65**, FL **0,53**, BD **0,44**, une forme de carapace.

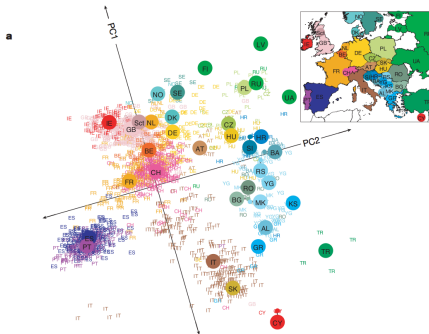
Remèdes à l'effet taille

Face à un facteur commun évident

Ne pas conclure à une seule dimension. Interpréter les axes suivants. Ou retirer la taille avant l'ACP : logarithmes des mesures, moins leur moyenne par individu.

Critère $\lambda_k > 1$: un seul axe retenu, et l'information biologique perdue.

Gènes et géographie



ACP de 500 000 SNP sur 3 000 Européens : le plan des axes 1 et 2 dessine la carte.

Novembre *et al.*, *Nature*, 2008.

Retour au C1

Faible part d'inertie, structure nette

500 000 variables, beaucoup de variations individuelles : chaque axe ne porte qu'une faible part de l'inertie.

Même leçon que l'axe 3 des crabes

Une part d'inertie faible n'empêche pas une structure biologique nette.

Grille de lecture

Grille de lecture d'une ACP

Cinq étapes

1. Valeurs propres, parts d'inertie, éboulis.
2. Cercle des corrélations : variables bien représentées, sens des axes.
3. Individus : $\cos^2$ avant toute proximité, contributions.
4. Facteur commun évident ? Effet taille, axes suivants.
5. Éléments supplémentaires : ce que l'ACP retrouve sans l'avoir reçu.

Limites de l'ACP

Des liens linéaires seulement : la parabole du C8, $r = 0$, reste invisible.

Une grande variance n'est pas forcément la question biologique : la taille des crabes.

ACP normée ou non : un choix, qui décide du poids de chaque variable.