

Modèles en biologie — Livret de TD

Version étudiante

Christophe Ambroise

Table des matières

Présentation	4
I Fiches de TD	5
1 TD 1 — Vecteurs et systèmes linéaires	6
1.1 Objectifs	6
1.2 Pour s'entraîner	8
2 TD 2 — Matrices et chaînes de Markov	9
2.1 Objectifs	9
2.2 Pour s'entraîner	11
3 TD 3 — Géométrie de $\mathbb{R}^p$	12
3.1 Objectifs	12
3.2 Pour s'entraîner	13
4 TD 4 — Distances et similarités	15
4.1 Objectifs	15
4.2 Pour s'entraîner	17
5 TD 5 — Clustering	18
5.1 Objectifs	18
5.2 Pour s'entraîner	19
6 TD 6 — Information et entropie	21
6.1 Objectifs	21
6.2 Pour s'entraîner	23
7 TD 7 — Variance, covariance, corrélation	24
7.1 Objectifs	24
7.2 Pour s'entraîner	26
8 TD 8 — Analyse en composantes principales	27
8.1 Objectifs	27
8.2 Pour s'entraîner	30
9 TD 9 — Synthèse : les crabes <i>Leptograpsus</i>	32
9.1 Objectifs	32
9.2 Pour s'entraîner	35

Annexes	37
A Notions d'algèbre linéaire à connaître	37
A.1 C1 — Vecteurs	37
A.2 C2 — Systèmes linéaires et pivot de Gauss	37
A.3 C3 — Matrices et chaînes de Markov	37
A.4 C4 — Géométrie de $\mathbb{R}^p$	38
A.5 C5 — Distances et similarités	38
A.6 C6 — Clustering	38
A.7 C7 — Information et entropie	38
A.8 C8 — Variance, covariance, corrélation	39
A.9 C9 — Analyse en composantes principales	39
A.10 Autres outils mathématiques	39
B Compétences à maîtriser	40
B.1 Résoudre un système par le pivot de Gauss	40
B.2 Calculer produit scalaire, norme, cosinus, projection	40
B.3 Centrer et réduire un tableau de données	40
B.4 Calculer une corrélation	41
B.5 Calculer une distance	41
B.6 Utiliser et interpréter les k-means	42
B.7 Lancer et interpréter une ACP	42
C Vérifier avec sympy	44
C.1 Résoudre un système : linsolve	44
C.2 Lire la forme échelonnée : rref	44
C.3 Reproduire le pivot : elementary_row_op	45
C.4 Un pivot détaillé automatiquement	45
C.5 Coefficients symboliques	47

Présentation

Ce livret accompagne les neuf séances de travaux dirigés du cours de modèles en biologie de L3 GBI.

Chaque fiche contient plus d'exercices qu'une séance ne permet d'en traiter. Les exercices restants servent à l'entraînement personnel.

Trois annexes terminent le livret : les notions d'algèbre linéaire à connaître, les compétences à maîtriser sous forme de méthodes, et la vérification des calculs avec sympy.

Cette version contient uniquement les énoncés.

Première partie

Fiches de TD

1 TD 1 — Vecteurs et systèmes linéaires

1.1 Objectifs

Cette fiche mobilise les notions des cours C1 et C2 : calcul dans $\mathbb{R}^n$, combinaisons linéaires, mise en équations, résolution de systèmes par le pivot de Gauss et lecture du nombre de solutions.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1 à 3), démonstration (4), applications biologiques (5 et 6). Les exercices 7 à 9 prolongent le travail et permettent de s'entraîner.

Exercice 1.1. Calcul vectoriel et combinaisons linéaires.

1. Dans $\mathbb{R}^2$, soient $\mathbf{u} = (2, -1)$, $\mathbf{v} = (-1, 3)$ et $\mathbf{w} = (4, 1)$. Calculer $\mathbf{u} + \mathbf{v}$, $2\mathbf{u} - \mathbf{v}$ et $-3\mathbf{v}$.
2. Déterminer deux réels α et β tels que $\mathbf{w} = \alpha\mathbf{u} + \beta\mathbf{v}$.
3. Dans $\mathbb{R}^3$, soient $\mathbf{a} = (1, 0, 2)$ et $\mathbf{b} = (0, 1, -1)$. Les vecteurs $\mathbf{c} = (2, 1, 3)$ et $\mathbf{e} = (1, 1, 2)$ sont-ils des combinaisons linéaires de $\mathbf{a}$ et $\mathbf{b}$? Donner les coefficients lorsqu'ils existent.
4. Caractériser par une équation les vecteurs (x, y, z) qui sont combinaisons linéaires de $\mathbf{a}$ et $\mathbf{b}$.

Exercice 1.2. Pivot de Gauss. Résoudre le système suivant par élimination de Gauss. Les opérations élémentaires effectuées sur les lignes doivent apparaître.

$$\begin{cases} x - 2y + 3z = 9, \\ 2x - y + 2z = 7, \\ -3x - 4y + z = 3. \end{cases}$$

Vérifier ensuite la solution dans le système initial.

Exercice 1.3. Combien de solutions ? Pour chacun des systèmes suivants :

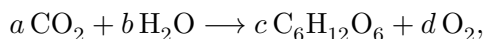
1. donner une forme échelonnée de la matrice augmentée ;
2. déterminer s'il possède une solution unique, aucune solution ou une infinité de solutions ;
3. donner l'ensemble des solutions lorsqu'il est non vide, sous forme vectorielle s'il est infini.

$$(S_1) \quad \begin{cases} x + y + z = 6, \\ 2x - y + z = 3, \\ x + 2y - z = 2, \end{cases} \quad (S_2) \quad \begin{cases} x + y - z = 1, \\ 2x + 2y - 2z = 2, \\ x - y + z = 3, \end{cases} \quad (S_3) \quad \begin{cases} x + y + z = 1, \\ 2x + 2y + 2z = 3. \end{cases}$$

Exercice 1.4. Deux résultats du cours. Soit un système S de m équations à n inconnues, d'équations $\sum_{j=1}^n a_{ij}x_j = b_i$ pour $i = 1, \dots, m$.

1. Soient $\mathbf{u}$ et $\mathbf{v}$ deux solutions et $t \in \mathbb{R}$. Montrer que $(1 - t)\mathbf{u} + t\mathbf{v}$ est encore une solution.
2. En déduire qu'un système qui possède deux solutions distinctes en possède une infinité.
3. Soit S' le système obtenu à partir de S par la transvection $L_i \leftarrow L_i + \lambda L_j$, avec $i \neq j$. Montrer que toute solution de S est solution de S' , puis que S s'obtient à partir de S' par une opération du même type. Conclure.
4. Résoudre le système $x + y = 2$, $x - y = 0$, puis le système obtenu en remplaçant sa seconde ligne par $0 \cdot L_2$. Pourquoi la dilatation exige-t-elle $\lambda \neq 0$?

Exercice 1.5. Équilibrer la photosynthèse. L'équation bilan de la photosynthèse s'écrit



où a, b, c, d sont des coefficients positifs. Chaque élément (C, H, O) compte autant d'atomes de part et d'autre de la flèche.

1. Écrire l'équation de conservation de chaque élément, puis le système linéaire en (a, b, c, d) , tous seconds membres nuls.
2. Résoudre ce système par le pivot de Gauss. Combien de variables libres?
3. Donner la solution en nombres entiers positifs les plus petits.
4. Même travail pour la respiration : $a \text{C}_6\text{H}_{12}\text{O}_6 + b \text{O}_2 \longrightarrow c \text{CO}_2 + d \text{H}_2\text{O}$. Comparer les deux bilans.

Exercice 1.6. Mélange de profils d'expression. Deux types cellulaires possèdent les profils d'expression moyens suivants pour trois gènes :

$$\mathbf{p}_A = (8, 2, 1), \quad \mathbf{p}_B = (2, 6, 5).$$

Le profil d'un échantillon mélangé est modélisé par $\mathbf{x} = \alpha \mathbf{p}_A + \beta \mathbf{p}_B$, avec $\alpha + \beta = 1$ et $\alpha, \beta \geq 0$.

1. Pour $\mathbf{x} = (3, 5; 5; 4)$, déterminer α et β , puis vérifier le résultat sur les trois coordonnées.
2. Même travail pour $\mathbf{y} = (5, 4, 3)$.
3. Le profil $\mathbf{z} = (6, 3, 4)$ est-il un mélange des deux types? Proposer une interprétation biologique.
4. Interpréter les coefficients α et β .

1.2 Pour s'entraîner

Exercice 1.7. Discussion suivant un paramètre. Pour $m \in \mathbb{R}$, étudier le nombre de solutions du système

$$\begin{cases} mx + y + z = 1, \\ x + my + z = 1, \\ x + y + mz = 1. \end{cases}$$

Donner l'ensemble des solutions dans chacun des cas rencontrés.

Exercice 1.8. Flux métaboliques à l'état stationnaire. Dans une cellule, un métabolite B est importé au flux $v_1 = 10$ (en mmol/h). Quatre réactions irréversibles le transforment :

$$v_2 : B \rightarrow C, \quad v_3 : B \rightarrow D, \quad v_4 : C \rightarrow D, \quad v_5 : D \rightarrow \text{excrétion}.$$

À l'état stationnaire, chaque métabolite interne (B , C , D) est produit autant qu'il est consommé.

1. Écrire les trois équations de conservation, d'inconnues v_2, v_3, v_4, v_5 .
2. Résoudre le système par le pivot de Gauss et donner l'ensemble des solutions sous forme vectorielle.
3. Les flux sont positifs. Quelles valeurs la variable libre peut-elle prendre ?
4. Une mesure donne $v_4 = 3$. Calculer tous les flux.
5. Quel flux est déterminé sans aucune mesure ? Interpréter.

Exercice 1.9. Système à quatre inconnues. Résoudre par élimination de Gauss le système suivant, en indiquant les principales opérations sur les lignes :

$$\begin{cases} x - 2y + 4z - 3t = -1, \\ -2x + y - 2z + 3t = 5, \\ -x + 3y - 6z - 3t = 14, \\ 2x - 3y + 5z - 5t = -6. \end{cases}$$

Vérifier la solution obtenue dans les quatre équations initiales.

2 TD 2 — Matrices et chaînes de Markov

2.1 Objectifs

Cette fiche mobilise les notions du cours C3 : produit matrice-vecteur, produit matriciel, puissances, matrice de transition, distribution d'équilibre par le pivot de Gauss, vecteur propre et valeur propre, transposée.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1, 2), applications (3, 4, 6), démonstration (5). Les exercices 7 à 9 prolongent le travail et permettent de s'entraîner.

Exercice 2.1. Dimensions et produits. Soient

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 3 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 1 \\ 1 & 0 \\ -1 & 2 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}.$$

1. Donner les dimensions de A , B et $\mathbf{u}$.
2. Parmi $A + B$, $A\mathbf{u}$, $B\mathbf{u}$, AB et BA , indiquer les opérations définies et la dimension de leur résultat.
3. Calculer $A\mathbf{u}$ de deux façons : ligne par ligne, puis comme combinaison des colonnes de A .
4. Calculer AB et BA . Pourquoi leur égalité était-elle impossible dès le départ ?

Exercice 2.2. Ordre des produits. Soient

$$C = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

1. Calculer CD et DC . Que montre ce calcul ?
2. Calculer $D\mathbf{x}$, puis $C(D\mathbf{x})$.
3. Calculer $(CD)\mathbf{x}$ et comparer avec la question précédente.
4. Calculer $(DC)\mathbf{x}$. Dans quel ordre les deux matrices agissent-elles sur $\mathbf{x}$?

Exercice 2.3. Cellule quiescente ou en division. Une cellule est soit quiescente (Q), soit en division (D). D'un jour à l'autre, une cellule quiescente entre en division avec la probabilité 0,2 ; une cellule en division redevient quiescente avec la probabilité 0,3. L'état est $\mathbf{x}_t = (q_t, d_t)$, et $\mathbf{x}_{t+1} = P\mathbf{x}_t$.

1. Dessiner le diagramme des transitions, écrire la matrice P (colonne j : départ de l'état j) et vérifier qu'il s'agit d'une matrice de transition.
2. Pour $\mathbf{x}_0 = (1, 0)$, calculer $\mathbf{x}_1$ et $\mathbf{x}_2$.
3. Calculer P^2 . Interpréter le coefficient $(P^2)_{21}$ en énumérant les chemins possibles.
4. Déterminer la distribution d'équilibre π par le pivot de Gauss.
5. Vérifier que $\mathbf{v} = (1, -1)$ est un vecteur propre de P et donner sa valeur propre λ . Écrire $\mathbf{x}_0 = \pi + c\mathbf{v}$, puis en déduire $\mathbf{x}_t$ par linéarité. Contrôler avec $\mathbf{x}_2$.
6. Quelle proportion de cellules est en division à long terme ? Dépend-elle de l'état initial ?

Exercice 2.4. Récepteur à trois états. Un récepteur membranaire est inactif (I), actif (A) ou désensibilisé (D). À chaque pas de temps : un récepteur inactif s'active avec la probabilité 0,2 ; un récepteur actif se désensibilise avec la probabilité 0,5 ; un récepteur désensibilisé redevient inactif avec la probabilité 0,4. Dans les autres cas, il reste dans son état.

1. Écrire la matrice de transition P , états dans l'ordre I, A, D.
2. Pour $\mathbf{x}_0 = (1, 0, 0)$, calculer $\mathbf{x}_1$, $\mathbf{x}_2$ et $\mathbf{x}_3$.
3. Déterminer la distribution d'équilibre par le pivot de Gauss appliqué à $P - I$, avec la contrainte $\pi_1 + \pi_2 + \pi_3 = 1$.
4. Quelle part des récepteurs est active à l'équilibre ? Les hypothèses du théorème de Perron-Frobenius sont-elles satisfaites ?

Exercice 2.5. Propriétés des matrices de transition. Soit P une matrice carrée d'ordre n à coefficients positifs ou nuls, et $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$.

1. Montrer que les colonnes de P ont pour somme 1 si et seulement si $\mathbf{1}^T P = \mathbf{1}^T$.
2. Si de plus P est une matrice de transition et $\mathbf{x}$ une loi de probabilité ($x_i \geq 0$, $\sum_i x_i = 1$). Montrer que $P\mathbf{x}$ est une loi de probabilité.
3. Montrer que le produit de deux matrices de transition est une matrice de transition. En déduire que P^t en est une pour tout $t \geq 1$.
4. Montrer que si $P\mathbf{v} = \lambda\mathbf{v}$, alors $P^t\mathbf{v} = \lambda^t\mathbf{v}$ pour tout $t \geq 1$.

Exercice 2.6. Matrice de Leslie. Une population est divisée en trois classes d'âge (0–1 an, 1–2 ans, 2–3 ans). Chaque année, un individu de la classe 1, 2 ou 3 produit en moyenne 0,5, 1,5 ou 4,5 jeunes, qui entrent dans la classe 1. La moitié des individus de la classe 1 survit et passe en classe 2 ; la moitié de ceux de la classe 2 passe en classe 3 ; aucun individu ne survit au-delà de la classe 3. Le vecteur $\mathbf{x}_t$ contient les effectifs des trois classes, et $\mathbf{x}_{t+1} = L\mathbf{x}_t$.

1. Écrire la matrice L et justifier chaque coefficient non nul.
2. Pour $\mathbf{x}_0 = (100, 0, 0)$, calculer $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ et l'effectif total chaque année.
3. Vérifier que $\mathbf{v} = (9, 3, 1)$ est un vecteur propre de L et donner sa valeur propre λ .

- Un calcul sur ordinateur donne des effectifs totaux de 2 783,2 à $t = 9$ et 4 158,4 à $t = 10$, avec $\mathbf{x}_{10} \approx (2871; 966; 321)$. Calculer le rapport des deux totaux et les proportions des trois classes à $t = 10$. Comparer avec λ et avec $\mathbf{v}$ normalisé. Interpréter.
- L est-elle une matrice de transition ? Pourquoi l'équilibre $L\pi = \pi$ n'est-il pas la bonne question pour ce modèle ?

2.2 Pour s'entraîner

Exercice 2.7. Surfeur aléatoire. Quatre pages web sont reliées par les liens $1 \rightarrow 2$, $1 \rightarrow 3$, $2 \rightarrow 3$, $3 \rightarrow 1$, $4 \rightarrow 1$ et $4 \rightarrow 3$. Un internaute suit, à chaque étape, un des liens de sa page au hasard, de façon équiprobable.

- Écrire la matrice de transition P (colonne j : départ de la page j).
- Déterminer l'équilibre π par le pivot de Gauss. Pourquoi la page 4 reçoit-elle une probabilité nulle ? Quelle hypothèse de Perron-Frobenius n'est pas satisfaite ?
- Pour y remédier, l'internaute suit un lien avec la probabilité 0,85, ou saute sur l'une des quatre pages au hasard avec la probabilité 0,15 : $G = 0,85P + \frac{0,15}{4}J$, où J est la matrice 4×4 dont tous les coefficients valent 1. Montrer que G est une matrice de transition à coefficients tous strictement positifs.
- Un calcul par itérations donne l'équilibre de G : (0,380; 0,199; 0,384; 0,038). Classer les pages et comparer avec la question 2.

Exercice 2.8. Transposée et matrice symétrique. Soit

$$E = \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 0 & 0 \\ 3 & 0 \end{pmatrix}.$$

- Donner les dimensions de E^T et de $E^T E$.
- Calculer $E^T E$ et vérifier qu'elle est symétrique.
- Soit M une matrice quelconque. À l'aide de la transposée d'un produit, montrer que $M^T M$ est toujours symétrique.

Exercice 2.9. Hypothèses de Perron-Frobenius. Pour chacune des matrices de transition suivantes, étudier l'existence et l'unicité d'une distribution d'équilibre, puis le comportement de $P^t \mathbf{x}_0$ à partir de $\mathbf{x}_0 = (0, 1)$. Dire quelle hypothèse du théorème n'est pas satisfaite.

$$P_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P_3 = \begin{pmatrix} 1 & 0,3 \\ 0 & 0,7 \end{pmatrix}.$$

Pour P_3 , l'état 1 est absorbant : proposer un exemple biologique.

3 TD 3 — Géométrie de $\mathbb{R}^p$

3.1 Objectifs

Cette fiche mobilise les notions du cours C4 : norme, distance, produit scalaire, similarité cosinus, angle, orthogonalité, projection sur une droite, centrage et droite des moindres carrés.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1, 3, 5), démonstration (4), applications (2, 6). Les exercices 7 et 8 prolongent le travail et permettent de s'entraîner.

Exercice 3.1. Calculs euclidiens. Soient $\mathbf{u} = (1, -2, 2)$ et $\mathbf{v} = (2, 1, 0)$.

1. Calculer $\|\mathbf{u}\|$, $\|\mathbf{v}\|$ et $\mathbf{u} \cdot \mathbf{v}$.
2. Calculer la distance $d(\mathbf{u}, \mathbf{v})$.
3. Les deux vecteurs sont-ils orthogonaux ? Vérifier sur cet exemple la relation $\|\mathbf{u} - \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\mathbf{u} \cdot \mathbf{v}$.
4. Donner le vecteur unitaire de même direction et de même sens que $\mathbf{u}$.

Exercice 3.2. Profils d'expression : forme et amplitude. Trois gènes sont mesurés à quatre instants successifs :

$$\mathbf{g}_1 = (1, 2, 3, 4), \quad \mathbf{g}_2 = (2, 4, 6, 8), \quad \mathbf{g}_3 = (4, 3, 2, 1).$$

1. Calculer les trois distances euclidiennes entre ces profils.
2. Calculer les trois similarités cosinus, puis les angles correspondants.
3. Quel gène est le plus proche de $\mathbf{g}_1$ selon la distance ? Selon le cosinus ? Lequel de ces deux critères repère des gènes régulés de la même façon ?
4. Centrer les trois profils, puis recalculer le cosinus entre $\mathbf{g}_1$ et $\mathbf{g}_3$. Interpréter.

Exercice 3.3. Orthogonalité et vecteur 1. Dans $\mathbb{R}^4$, soient

$$\mathbf{a} = (1, 1, 0, 0), \quad \mathbf{b} = (1, -1, 2, 0), \quad \mathbf{c} = (0, 0, 1, 1), \quad \mathbf{d} = (1, -1, -1, 1).$$

1. Déterminer les couples de vecteurs orthogonaux.
2. Vérifier le théorème de Pythagore pour $\mathbf{a} + \mathbf{d}$.
3. Lesquels de ces vecteurs sont orthogonaux à $\mathbf{1} = (1, 1, 1, 1)$? Traduire cette propriété en termes de moyenne.
4. Déterminer le réel α tel que $\mathbf{a} + \alpha \mathbf{1}$ soit orthogonal à $\mathbf{1}$. Reconnaître le vecteur obtenu.

Exercice 3.4. Identités du produit scalaire. Soient $\mathbf{x}$ et $\mathbf{y}$ deux vecteurs de $\mathbb{R}^p$.

1. Montrer l'identité du parallélogramme : $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$.
2. Montrer que $\mathbf{x} \cdot \mathbf{y} = \frac{1}{4} (\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2)$.
3. Réciproque du théorème de Pythagore : montrer que si $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, alors $\mathbf{x} \perp \mathbf{y}$.
4. Montrer que $\|\mathbf{x}\| = \|\mathbf{y}\|$ si et seulement si $\mathbf{x} + \mathbf{y}$ et $\mathbf{x} - \mathbf{y}$ sont orthogonaux. Interpréter dans le plan.

Exercice 3.5. Projection et centrage.

1. Calculer la projection $\mathbf{p}$ de $\mathbf{x} = (2, 1, 3)$ sur la droite dirigée par $\mathbf{1} = (1, 1, 1)$, le résidu $\mathbf{x} - \mathbf{p}$, puis la distance de $\mathbf{x}$ à cette droite. Vérifier l'orthogonalité du résidu.
2. Même travail avec la droite dirigée par $\mathbf{w} = (1, 0, 1)$.
3. Montrer que, pour tout $\mathbf{x} \in \mathbb{R}^n$, la projection de $\mathbf{x}$ sur la droite dirigée par $\mathbf{1}$ est $\bar{x}\mathbf{1}$.
4. Centrer le vecteur $\mathbf{y} = (3, 1, 4, 2)$ et vérifier que la somme de ses coordonnées centrées est nulle.

Exercice 3.6. Nuage projeté sur deux axes. Quatre individus sont décrits par deux mesures morphologiques. Le tableau est déjà centré :

$$X_c = \begin{pmatrix} -2 & -1 \\ -1 & -1 \\ 1 & 0 \\ 2 & 2 \end{pmatrix}, \quad \mathbf{u}_1 = \begin{pmatrix} 0,6 \\ 0,8 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0,8 \\ -0,6 \end{pmatrix}.$$

1. Vérifier que les deux colonnes de X_c sont orthogonales à $\mathbf{1}$.
2. Vérifier que $\mathbf{u}_1$ et $\mathbf{u}_2$ sont unitaires et orthogonaux.
3. Calculer les coordonnées des individus sur chaque axe, $X_c\mathbf{u}_1$ et $X_c\mathbf{u}_2$.
4. Pour le premier individu, vérifier que $\mathbf{x}_1 = c_{11}\mathbf{u}_1 + c_{12}\mathbf{u}_2$, où c_{11} et c_{12} sont ses deux coordonnées. En déduire, par Pythagore, $\|\mathbf{x}_1\|^2 = c_{11}^2 + c_{12}^2$.
5. Calculer la somme des carrés des coordonnées sur chaque axe et la comparer à $\sum_i \|\mathbf{x}_i\|^2$. Quel axe résume le mieux le nuage ?

3.2 Pour s'entraîner

Exercice 3.7. Loi d'échelle du métabolisme. Pour quatre espèces de mammifères, de masses 1, 10, 100 et 1 000 kg, le métabolisme de repos M (en kcal/jour) vérifie

$$x = \log_{10}(\text{masse}) = (0, 1, 2, 3), \quad y = \log_{10} M = (1,85; 2,50; 3,25; 4,10).$$

1. Écrire le système $\beta_0 + \beta_1 x_i = y_i$ ($i = 1, \dots, 4$) sous la forme $X\beta = \mathbf{y}$ et montrer qu'il est incompatible.
2. Calculer $X^T X$ et $X^T \mathbf{y}$.

3. Résoudre les équations normales $X^T X \hat{\beta} = X^T \mathbf{y}$ par le pivot de Gauss.
4. Calculer le résidu $\mathbf{r} = \mathbf{y} - X \hat{\beta}$ et vérifier qu'il est orthogonal aux deux colonnes de X .
5. Prédire le métabolisme d'un animal de 50 kg. Interpréter la pente.

Exercice 3.8. Projection par minimisation. Soient $\mathbf{x} \in \mathbb{R}^p$ et $\mathbf{u} \neq \mathbf{0}$. Pour $c \in \mathbb{R}$, soit $f(c) = \|\mathbf{x} - c\mathbf{u}\|^2$.

1. Développer $f(c)$ en un polynôme du second degré en c .
2. Déterminer la valeur c^* qui minimise f et retrouver la formule de la projection du cours.
3. Calculer $f(c^*)$. En utilisant $f(c^*) \geq 0$, retrouver l'inégalité de Cauchy-Schwarz.

4 TD 4 — Distances et similarités

4.1 Objectifs

Cette fiche mobilise les notions du cours C5 : axiomes d'une distance, distances d_1 , d_2 , d_∞ , effet des unités et réduction, distances de Hamming et de Jaccard, similarité cosinus, corrélation, matrice des distances.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1), démonstration (2), applications (3 à 6). Les exercices 7 et 8 prolongent le travail et permettent de s'entraîner.

Exercice 4.1. Trois distances sur $\mathbb{R}^p$.

1. Pour $\mathbf{x} = (1, 2, 0, 4)$ et $\mathbf{y} = (3, -1, 1, 4)$, calculer $d_1(\mathbf{x}, \mathbf{y})$, $d_2(\mathbf{x}, \mathbf{y})$ et $d_\infty(\mathbf{x}, \mathbf{y})$. Vérifier $d_\infty \leq d_2 \leq d_1$.
2. Dans le plan, dessiner la boule unité de centre $\mathbf{0}$ pour chacune des trois distances.
3. Pour chacun des points $(0,5; 0,5)$, $(0,8; 0,8)$ et $(1,0)$, dire s'il appartient à chacune des trois boules unités.

Exercice 4.2. Axiomes d'une distance.

1. Montrer que $d(x, y) = |x - y|$ est une distance sur $\mathbb{R}$.
2. Trouver trois réels x, y, z , différents de ceux du cours, pour lesquels $\delta(x, y) = (x - y)^2$ viole l'inégalité triangulaire. Montrer que $\sqrt{\delta}$ est une distance.
3. Montrer que d_∞ est une distance sur $\mathbb{R}^p$.
4. Soit $\mathbf{a} \in \mathbb{R}^p$. Montrer que $\max_i |a_i| \leq \|\mathbf{a}\| \leq \sum_i |a_i|$. En déduire $d_\infty \leq d_2 \leq d_1$.

Exercice 4.3. Séquences et génotypes.

1. Quatre séquences alignées d'un même gène :

S1 ACGTTGCAAC
S2 ACGTTGCATC
S3 ACCTTGGATC
S4 TCGATGCAAG

Calculer la matrice des distances de Hamming. Quelle paire est la plus proche ?
Donner la proportion de sites différents entre S3 et S4.

2. Vérifier l'inégalité triangulaire pour le triplet (S2, S1, S4).

3. Trois individus sont génotypés en cinq SNP, codés par le nombre d'allèles mineurs (0, 1 ou 2) : $I_1 = (0, 1, 2, 1, 0)$, $I_2 = (0, 2, 2, 0, 0)$, $I_3 = (2, 1, 0, 1, 1)$. Calculer, pour chaque paire, la distance d_1 et le nombre de SNP où les génotypes diffèrent. Interpréter la différence entre les deux mesures.

Exercice 4.4. Jaccard et doubles absences. Dix espèces sont recherchées sur trois sites (1 : présente, 0 : absente).

Site	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}
1	1	1	1	0	1	0	0	0	0	0
2	1	1	0	0	1	1	0	0	0	0
3	0	0	1	1	0	0	1	0	0	0

1. Pour chaque paire de sites, calculer l'indice de Jaccard J , la distance $d_J = 1 - J$ et l'accord simple (part des espèces de même statut).
2. Quatre-vingt-dix autres espèces, absentes des trois sites, sont ajoutées à l'inventaire. Recalculer les deux indices.
3. Quel indice décrit la ressemblance écologique des sites ? Pourquoi ?

Exercice 4.5. Unités et réduction. Quatre rats sont décrits par la longueur du corps et la masse.

Rat	A	B	C	D
Longueur (cm)	20	22	26	28
Masse (kg)	0,3	0,5	0,3	0,5

1. Avec les unités du tableau, quel rat est le plus proche de A au sens de d_2 ?
2. Même question avec la masse en grammes.
3. Calculer l'écart-type s_j de chaque variable (en $1/n$), puis les données réduites $x_{ij}^* = x_{ij}/s_j$.
4. Calculer les distances de A aux trois autres rats après réduction. Pourquoi la réponse ne dépend-elle plus de l'unité de masse ?

Exercice 4.6. Profils : euclidienne, cosinus, corrélation. Quatre gènes sont mesurés à cinq instants :

$$\mathbf{g} = (1, 3, 5, 3, 1), \quad \mathbf{g} + 10\mathbf{1}, \quad 2\mathbf{g}, \quad \mathbf{h} = (5, 3, 1, 3, 5).$$

1. Calculer la distance euclidienne de $\mathbf{g}$ à chacun des trois autres profils.
2. Calculer les similarités cosinus correspondantes.
3. Centrer $\mathbf{g}$, puis calculer les corrélations de $\mathbf{g}$ avec les trois autres profils.
4. Quelle mesure regroupe les gènes de même dynamique ? Quelle mesure distingue les gènes de dynamique opposée ?
5. Montrer que pour $a > 0$ et $b \in \mathbb{R}$, $r(\mathbf{x}, a\mathbf{y} + b\mathbf{1}) = r(\mathbf{x}, \mathbf{y})$. Que devient ce résultat pour $a < 0$? Retrouver ainsi $r(\mathbf{g}, \mathbf{h})$.

4.2 Pour s'entraîner

Exercice 4.7. Insertions et délétions. La distance d'édition entre deux séquences est le plus petit nombre d'opérations élémentaires (substitution, insertion ou délétion d'une base) qui transforme l'une en l'autre.

1. Pour $u = \text{ACGTACGT}$ et $v = \text{CGTACGTA}$, calculer $d_H(u, v)$.
2. Transformer u en v en deux opérations. Montrer qu'une seule opération ne suffit pas, et conclure.
3. Montrer que, pour deux séquences de même longueur, la distance d'édition est au plus égale à la distance de Hamming.
4. Pourquoi la distance de Hamming surestime-t-elle la divergence après une insertion ou une délétion ?

Exercice 4.8. Dissimilarités.

1. Sur des profils non constants, soit $\delta(\mathbf{x}, \mathbf{y}) = 1 - r(\mathbf{x}, \mathbf{y})$. Montrer que δ n'est pas une distance.
2. Une matrice est proposée comme matrice des distances de quatre objets :

$$M = \begin{pmatrix} 0 & 2 & 5 & 3 \\ 2 & 0 & 2 & 3 \\ 5 & 2 & 0 & 1 \\ 3 & 3 & 1 & 0 \end{pmatrix}.$$

Vérifier la symétrie et la diagonale, puis rechercher les triplets qui violent l'inégalité triangulaire.

3. Quelles valeurs de $m_{13} = m_{31}$ rendent M compatible avec les axiomes ?

5 TD 5 — Clustering

5.1 Objectifs

Cette fiche mobilise les notions du cours C6 : centre de gravité, inertie, inerties intra-classe et inter-classe, théorème de Huygens, algorithme des k-means, classification ascendante hiérarchique, dendrogramme, UPGMA.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1 à 3), démonstration (4), applications (5, 6). Les exercices 7 et 8 prolongent le travail et permettent de s'entraîner.

Exercice 5.1. Inertie et Huygens en dimension 1. Cinq larves mesurent 1, 2, 3, 9 et 10 mm.

1. Calculer le centre de gravité g et l'inertie totale I .
2. Pour la partition $\{1, 2, 3\}, \{9, 10\}$, calculer les centres de classe, W et B , puis vérifier $I = W + B$.
3. Même travail pour la partition $\{1\}, \{2, 3, 9, 10\}$.
4. Calculer la part d'inertie expliquée B/I des deux partitions. Laquelle décrit le mieux les données ?

Exercice 5.2. k-means à la main en dimension 1. Mêmes données, $K = 2$. À chaque étape, noter le critère $\frac{1}{n} \sum_i (x_i - c_{k(i)})^2$, où $c_{k(i)}$ est le centre auquel l'individu i est affecté.

1. Partir des centres $c_1 = 1$ et $c_2 = 2$. Enchaîner affectations et recentrages jusqu'à stabilité.
2. Recommencer à partir des centres $c_1 = 9$ et $c_2 = 10$.
3. Comparer les deux partitions finales et l'évolution du critère.

Exercice 5.3. k-means en dimension 2. Six cellules sont repérées par deux marqueurs :

$$(0, 0), (1, 0), (0, 1), (4, 4), (5, 4), (4, 5).$$

1. Avec $K = 2$ et les centres de départ $\mathbf{c}_1 = (0, 0)$, $\mathbf{c}_2 = (1, 0)$, effectuer la première affectation, puis le recentrage. Calculer le critère après chacune de ces étapes.
2. Poursuivre jusqu'à stabilité.
3. Calculer l'inertie totale I et la part B/I de la partition finale.

Exercice 5.4. Trois résultats du cours.

1. Énumérer les partitions de $\{1, 2, 3\}$ en deux classes. Montrer qu'un ensemble de n individus possède $2^{n-1} - 1$ partitions en deux classes.
2. Soient $x_1, \dots, x_n$ réels et $f(c) = \sum_i (x_i - c)^2$. Montrer que $f(c) = \sum_i (x_i - \bar{x})^2 + n(\bar{x} - c)^2$. En déduire que f est minimale en $c = \bar{x}$, et donner sa valeur minimale à l'aide de l'écart-type.
3. Montrer que $W = 0$ si et seulement si chaque classe ne contient que des points confondus. Que vaut alors B/I ?

Exercice 5.5. Classification hiérarchique. Pour cinq espèces P, Q, R, S, T, la distance entre deux espèces est le nombre de différences pour 100 sites d'un gène :

$$D = \begin{array}{c|ccccc} & P & Q & R & S & T \\ \hline P & 0 & 9 & 12 & 16 & 18 \\ Q & 9 & 0 & 10 & 14 & 16 \\ R & 12 & 10 & 0 & 6 & 10 \\ S & 16 & 14 & 6 & 0 & 4 \\ T & 18 & 16 & 10 & 4 & 0 \end{array}$$

1. Effectuer la classification ascendante hiérarchique au saut minimal : donner chaque fusion, sa hauteur et la matrice mise à jour. Dessiner le dendrogramme.
2. Même travail au saut maximal.
3. Comparer les partitions en trois classes, puis en deux classes, obtenues par les deux méthodes. Expliquer la différence.

Exercice 5.6. UPGMA et horloge moléculaire. Même matrice D .

1. Construire l'arbre UPGMA (saut moyen) : fusions et hauteurs.
2. Sous l'hypothèse d'horloge moléculaire, chaque lignée accumule une différence pour 100 sites par million d'années : deux espèces séparées depuis t millions d'années sont à la distance $2t$. Dater chaque nœud de l'arbre.
3. Dessiner l'arbre enraciné, feuilles alignées sur l'axe du temps présent.
4. Quelle observation sur les données remettrait en cause l'hypothèse d'horloge moléculaire ?

5.2 Pour s'entraîner

Exercice 5.7. k-means sur les iris avec R. Le jeu `iris` décrit 150 fleurs de trois espèces par quatre mesures des sépales et des pétales.

```
x <- scale(iris[, 1:4])
set.seed(1)
km <- kmeans(x, centers = 3, nstart = 25)
table(km$cluster, iris$Species)
km$betweenss / km$totss
```

```
W <- sapply(1:6, function(k) kmeans(x, k, nstart = 25)$tot.withinss)
plot(1:6, W, type = "b")
```

1. Pourquoi réduire les variables ? Pourquoi `nstart = 25` ?
2. Combien de fleurs sont regroupées avec leur espèce ? Quelle espèce est isolée sans erreur ?
3. Lire la part d'inertie expliquée et la comparer à celle du cours (pétales seuls, $B/I \approx 94\%$).
4. Où se situe le coude de la courbe $W(K)$?

Exercice 5.8. Minimum local en dimension 1. Quatre valeurs 0, 2, 10, 11, $K = 3$.

1. Montrer que la partition $\{0, 2\}$, $\{10\}$, $\{11\}$ est stable pour les k-means, et calculer son critère W .
2. Même question pour $\{0\}$, $\{2\}$, $\{10, 11\}$.
3. Énumérer toutes les partitions en trois classes et identifier celle de plus petit W . Conclure sur l'algorithme.

6 TD 6 — Information et entropie

6.1 Objectifs

Cette fiche mobilise les notions du cours C7 : surprise, bit, entropie et ses bornes, matrice des fréquences d'un alignement, contenu en information et logo, entropie comparée à la variance, loi jointe de deux variables, marges, indépendance, lois conditionnelles et information mutuelle.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1, 2), applications (3, 5, 6), démonstration (4). Les exercices 7 et 8 prolongent le travail et permettent de s'entraîner.

Valeurs utiles : $\log_2 3 \approx 1,585$, $\log_2 0,75 \approx -0,415$, $\log_2 0,7 \approx -0,515$, $\log_2 0,4 \approx -1,322$, $\log_2 0,1 \approx -3,322$, $\log_2 \frac{5}{7} \approx -0,485$, $\log_2 \frac{2}{7} \approx -1,807$.

Exercice 6.1. Surprise et entropie.

1. Calculer la surprise, en bits, d'événements de probabilités $1/8$, $1/2$ et $3/4$.
2. Calculer l'entropie des lois suivantes sur A, C, G, T : $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $(0,7; 0,1; 0,1; 0,1)$, $(1,0,0,0)$. Les ranger par incertitude croissante.
3. Sous l'effet d'un médicament, l'équilibre du canal ionique du C3 devient $\pi' = (\frac{5}{7}, \frac{2}{7})$ (fermé, ouvert). Calculer son entropie et la comparer à celle de l'équilibre sans médicament, 0,65 bit. Interpréter.

Exercice 6.2. Entropie et nombre de questions.

1. Une espèce est tirée au hasard parmi 8 espèces équiprobables. Combien de questions oui/non suffisent pour l'identifier ? Comparer à l'entropie.
2. Cinq allèles a, b, c, d, e ont les fréquences $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16})$. Calculer l'entropie, puis le nombre moyen de questions de la stratégie « $a?$ », « $b?$ », « $c?$ », « $d?$ ».
3. Pour trois modalités équiprobables, calculer l'entropie et le nombre moyen de questions de la meilleure stratégie. Comparer.

Exercice 6.3. Information d'un site de fixation. Huit sites de fixation d'un facteur de transcription sont alignés :

ATAACG
CTAACG
GTGATG
TTGATG

ATACTG
CTGGTG
GTACTG
TTGTTG

1. Écrire la matrice des effectifs n_{bj} (lignes A, C, G, T ; colonnes : les six positions), puis la matrice des fréquences.
2. Calculer l'entropie H_j de chaque position.
3. Calculer le contenu en information $R_j = 2 - H_j$ de chaque position et l'information totale du site.
4. Esquisser le logo de séquence. Quelles positions la protéine reconnaît-elle ? Donner la séquence consensus.

Exercice 6.4. Propriétés de l'entropie.

1. Montrer que la surprise est additive : $h(pq) = h(p) + h(q)$.
2. Montrer que $H(\mathbf{p}) \geq 0$, et que $H(\mathbf{p}) = 0$ si et seulement si une modalité est certaine.
3. Montrer que la loi uniforme sur K modalités a pour entropie $\log_2 K$.
4. Soit $J = \mathbf{p}\mathbf{q}^T$ la loi jointe de deux variables indépendantes. Montrer que $H(X, Y) = H(X) + H(Y)$ et en déduire $I(X; Y) = 0$.

Exercice 6.5. Entropie et variance.

1. Sur les valeurs $\{1, 2, 3, 4\}$, comparer la loi uniforme et la loi qui donne $\frac{1}{2}$ à 1 et $\frac{1}{2}$ à 4 : entropie, moyenne et variance $\sum_i p_i(x_i - \mu)^2$. Laquelle est la plus dispersée selon chaque critère ?
2. Une position d'ADN porte A ou T avec probabilité $\frac{1}{2}$ chacune. Calculer son entropie, puis la variance obtenue en codant A = 1, C = 2, G = 3, T = 4, et en codant A = 1, T = 2, C = 3, G = 4. Conclure.
3. L'entropie d'une loi normale d'écart-type σ vaut $\frac{1}{2} \log_2(2\pi e \sigma^2)$ (résultat admis). Combien de bits gagne-t-elle quand σ double ? Quand σ est multiplié par 10 ?

Exercice 6.6. Déséquilibre de liaison. Deux locus d'un chromosome portent chacun deux allèles, A ou a et B ou b. Les fréquences des quatre haplotypes forment la loi jointe (X : locus 1 en ligne ; Y : locus 2 en colonne) :

$$J = \begin{pmatrix} 0,4 & 0,1 \\ 0,1 & 0,4 \end{pmatrix} \quad \begin{array}{l} \text{lignes : A, a} \\ \text{colonnes : B, b} \end{array}$$

1. Calculer les marges $J\mathbf{1}$ et $J^T\mathbf{1}$: fréquences des allèles à chaque locus.
2. Calculer la loi jointe $\mathbf{p}\mathbf{q}^T$ attendue sous indépendance, et l'écart $D = J_{11} - p_1q_1$.
3. Calculer les lois conditionnelles du locus 1 sachant le locus 2 (colonnes de J normalisées). Que vaut $\Pr(A | B)$?
4. Calculer $H(X)$, $H(Y)$, $H(X, Y)$ et l'information mutuelle.
5. Même calcul, sans détail, pour $J = \begin{pmatrix} 0,5 & 0 \\ 0 & 0,5 \end{pmatrix}$ et pour $J = \begin{pmatrix} 0,25 & 0,25 \\ 0,25 & 0,25 \end{pmatrix}$. Interpréter les trois situations en termes de recombinaison entre les deux locus.

6.2 Pour s'entraîner

Exercice 6.7. Indépendance et rang 1. Deux positions d'une séquence prennent trois valeurs chacune. Deux lois jointes sont proposées :

$$J_1 = \begin{pmatrix} 0,02 & 0,08 & 0,10 \\ 0,03 & 0,12 & 0,15 \\ 0,05 & 0,20 & 0,25 \end{pmatrix}, \quad J_2 = \begin{pmatrix} 0,10 & 0,05 & 0,10 \\ 0,10 & 0,20 & 0,10 \\ 0,10 & 0,05 & 0,20 \end{pmatrix}.$$

1. Calculer les marges de chaque loi.
2. Laquelle s'écrit $\mathbf{p}\mathbf{q}^T$? Justifier par les colonnes.
3. Calculer les lois conditionnelles (colonnes normalisées) des deux matrices. Que remarquer dans le cas indépendant ?
4. Sans calcul, dire pour quelle loi l'information mutuelle est nulle.

Exercice 6.8. Dépendance sans corrélation. X prend les valeurs $-1, 0, 1$ avec la probabilité $\frac{1}{3}$ chacune, et $Y = X^2$.

1. Écrire la loi jointe de (X, Y) sous forme d'une matrice 3×2 .
2. Les trois couples équiprobables $(-1, 1), (0, 0), (1, 1)$ forment deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^3$. Calculer leur corrélation.
3. Calculer $H(X), H(Y), H(X, Y)$ et $I(X; Y)$.
4. Conclure sur ce que mesurent la corrélation et l'information mutuelle.

7 TD 7 — Variance, covariance, corrélation

7.1 Objectifs

Cette fiche mobilise les notions du cours C8 : matrice des données, centre de gravité, matrice centrée, variance, covariance, corrélation, matrice de covariance S , matrice des corrélations R , trace et inertie, coordonnées sur un axe et variance projetée $\mathbf{u}^T S \mathbf{u}$.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1, 2, 6), démonstration (3), applications (4, 5). Les exercices 7 et 8 prolongent le travail et permettent de s'entraîner.

Exercice 7.1. Variance, covariance et corrélation à la main. Cinq plants sont décrits par leur hauteur x (en dm) et leur nombre de feuilles y :

$$\mathbf{x} = (1, 2, 3, 4, 5), \quad \mathbf{y} = (2, 4, 5, 4, 5).$$

1. Calculer les moyennes et les vecteurs centrés.
2. Calculer les variances s_x^2 et s_y^2 .
3. Calculer la covariance s_{xy} et la corrélation r .
4. Vérifier la covariance par la formule $s_{xy} = \frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y}$ (démontrée à l'exercice 3).

Exercice 7.2. Matrice de covariance en dimension 3. Quatre individus sont décrits par trois variables :

$$X = \begin{pmatrix} 9 & 3 & 4 \\ 9 & 5 & 2 \\ 11 & 5 & 2 \\ 11 & 7 & 4 \end{pmatrix}.$$

1. Calculer le centre de gravité $\mathbf{g} = \frac{1}{n} X^T \mathbf{1}$.
2. Calculer $X_c = X - \mathbf{1} \mathbf{g}^T$ et vérifier $X_c^T \mathbf{1} = \mathbf{0}$.
3. Calculer $S = \frac{1}{n} X_c^T X_c$. Vérifier sa symétrie.
4. Calculer $\text{tr}(S)$ et l'inertie du nuage.
5. Calculer la matrice des corrélations R . Quelles variables sont non corrélées ?

Exercice 7.3. Trois propriétés de la covariance. Soient $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

1. Montrer que $s_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y}$.

2. Soient $a, b \in \mathbb{R}$ et $\mathbf{x}' = a\mathbf{x} + b\mathbf{1}$. Exprimer $s_{x'y}$ et $s_{x'}$ en fonction de s_{xy} et s_x . En déduire $r(\mathbf{x}', \mathbf{y})$ selon le signe de a .
3. Montrer que la matrice de covariance vérifie $\mathbf{u}^T S \mathbf{u} \geq 0$ pour tout $\mathbf{u} \in \mathbb{R}^p$.

Exercice 7.4. Température et vitesse de course. Chez quatre lézards, la température corporelle T (en °C) et la vitesse de course v (en m/s) sont :

$$T = (20, 24, 28, 32), \quad v = (1, 0; 1, 4; 1, 6; 2, 0).$$

1. Calculer s_T^2 , s_v^2 , s_{Tv} et r .
2. Convertir la température en degrés Fahrenheit, $F = 1,8T + 32$. Recalculer la covariance et la corrélation sans refaire tout le calcul.
3. Que deviennent la covariance et la corrélation si la vitesse est exprimée en cm/s ?
4. Laquelle des deux mesures permet de comparer des études réalisées dans des unités différentes ?

Exercice 7.5. Optimum thermique. Le taux de croissance y d'une bactérie est mesuré à cinq températures, repérées par leur écart x (en °C) à la température optimale :

$$\mathbf{x} = (-2, -1, 0, 1, 2), \quad \mathbf{y} = (1, 4, 5, 4, 1).$$

1. Représenter les cinq points.
2. Calculer la covariance et la corrélation de $\mathbf{x}$ et $\mathbf{y}$.
3. La croissance dépend-elle de la température ? Que mesure exactement la corrélation ?
4. Quelle grandeur du cours C7 détecterait cette dépendance ?

Exercice 7.6. Variance projetée. Le nuage centré du TD 3 (exercice 6) a pour matrice

$$X_c = \begin{pmatrix} -2 & -1 \\ -1 & -1 \\ 1 & 0 \\ 2 & 2 \end{pmatrix}.$$

1. Calculer sa matrice de covariance S et son inertie.
2. Pour $\mathbf{u} = (\cos \theta, \sin \theta)$, calculer la variance projetée $\mathbf{u}^T S \mathbf{u}$ pour $\theta = 0^\circ, 45^\circ, 90^\circ$ et 135° .
3. Calculer $\mathbf{u}^T S \mathbf{u}$ pour $\mathbf{u}_1 = (0, 6; 0, 8)$ et $\mathbf{u}_2 = (0, 8; -0, 6)$. Comparer avec les sommes de carrés du TD 3.
4. Vérifier que deux axes orthogonaux se partagent toujours l'inertie.
5. Parmi les axes testés, lequel porte le plus de variance ? L'axe $\mathbf{u}_1$ est-il le meilleur ?

7.2 Pour s'entraîner

Exercice 7.7. Covariances des iris avec R.

```
x <- iris[, 1:4]
n <- nrow(x)
cov(x)                                # divise par n - 1
S <- cov(x) * (n - 1) / n             # matrice S du cours, en 1/n
sum(diag(S))                          # trace
round(cor(x), 2)                      # matrice des corrélations
u <- rep(0.5, 4)
t(u) %*% S %*% u                     # variance projetée
coord <- as.matrix(x) %*% u
mean((coord - mean(coord))^2)         # variance des coordonnées
```

1. Quelle variable a la plus grande variance ? Quelle est l'inertie du nuage ?
2. Quelle paire de variables est la plus corrélée ? Quelle variable est corrélée négativement aux autres ?
3. Vérifier que les deux dernières commandes donnent le même nombre. Quelle formule du cours cette égalité illustre-t-elle ?
4. Pourquoi `cov(x)` et `S` diffèrent-elles, et pourquoi `cor(x)` n'est-elle pas concernée ?

Exercice 7.8. Variance d'une somme.

1. Soient $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, de matrice de covariance $S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$. Montrer que la variance de $\mathbf{x} + \mathbf{y}$ vaut $s_x^2 + s_y^2 + 2s_{xy}$, en l'écrivant $\mathbf{u}^T S \mathbf{u}$ pour un vecteur $\mathbf{u}$ à préciser. Même question pour $\mathbf{x} - \mathbf{y}$.
2. Chez quatre souris, la longueur du corps vaut (8, 9, 10, 11) cm et celle de la queue (7, 9, 8, 10) cm. Calculer les deux variances et la covariance, puis la variance de la longueur totale et de la différence, par la formule et directement.
3. Pourquoi la longueur totale varie-t-elle plus que chacune de ses deux parties ?

8 TD 8 — Analyse en composantes principales

8.1 Objectifs

Cette fiche mobilise les notions du cours C9 : déterminant 2×2 , équation des valeurs propres, vecteur propre par le pivot de Gauss, axe de variance maximale, axes principaux orthogonaux, valeurs propres et inertie, ACP normée, composantes principales, $\cos^2$ et contributions des individus, cercle des corrélations, éléments supplémentaires, effet taille.

Les exercices 1 à 6 constituent le cœur de la séance : calcul (1, 2), démonstration (4), applications (3, 5, 6). Les exercices 7 et 8 prolongent le travail et permettent de s'entraîner.

Les sorties R utiles sont reproduites dans les énoncés : les exercices 5 à 8 se traitent sur papier ou sur machine. Les signes des axes calculés par R sont arbitraires : changer $\mathbf{u}_k$ en $-\mathbf{u}_k$ ne change pas l'analyse.

Exercice 8.1. Valeurs et vecteurs propres 2×2 . Soit $A = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$.

1. Calculer $\text{tr}(A)$ et $\det(A)$. Écrire et résoudre l'équation des valeurs propres.
2. Pour chaque valeur propre λ , trouver un vecteur propre par le pivot de Gauss sur $(A - \lambda I)\mathbf{v} = \mathbf{0}$. Vérifier $A\mathbf{v} = \lambda\mathbf{v}$.
3. Vérifier que les deux vecteurs propres sont orthogonaux, puis les normer.
4. Même travail pour $C = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$. La matrice C est la matrice de covariance d'un nuage de points du plan : que dit sa valeur propre nulle de ce nuage ?

Exercice 8.2. Axe de variance maximale d'un nuage. Le nuage centré du TD 3 (exercice 6) et du TD 7 (exercice 6) a pour matrice et pour matrice de covariance

$$X_c = \begin{pmatrix} -2 & -1 \\ -1 & -1 \\ 1 & 0 \\ 2 & 2 \end{pmatrix}, \quad S = \begin{pmatrix} 2,5 & 1,75 \\ 1,75 & 1,5 \end{pmatrix}.$$

1. Calculer les valeurs propres de S . Vérifier que leur somme est l'inertie.
2. Calculer le vecteur propre unitaire $\mathbf{u}_1$ par le pivot de Gauss, et son angle avec le premier axe. En déduire $\mathbf{u}_2$ sans calcul.

3. Calculer la composante principale $\mathbf{c}_1 = X_c \mathbf{u}_1$ et vérifier que sa variance vaut λ_1 .
4. Calculer les parts d'inertie des deux axes. Comparer λ_1 aux variances projetées du TD 7 : 3,75 à 45°, 3,54 sur (0,6; 0,8).
5. Calculer le $\cos^2$ de chaque point sur l'axe 1 et sa contribution ctr_{i1} . Quel point est le moins bien représenté ? Lequel contribue le plus ?

Exercice 8.3. Effet taille sur trois mesures. Chez des moules, la longueur, la largeur et la hauteur de la coquille sont corrélées deux à deux avec le même coefficient 0,6 :

$$R = \begin{pmatrix} 1 & 0,6 & 0,6 \\ 0,6 & 1 & 0,6 \\ 0,6 & 0,6 & 1 \end{pmatrix}.$$

1. Vérifier que $(1, 1, 1)$ est un vecteur propre de R et donner sa valeur propre.
2. Vérifier que $(1, -1, 0)$ et $(1, 0, -1)$ sont des vecteurs propres, de même valeur propre. Sont-ils orthogonaux ? Vérifier que $(1, 1, -2)$ est aussi vecteur propre et qu'il est orthogonal à $(1, -1, 0)$.
3. En utilisant la trace, justifier que les valeurs propres sont exactement celles trouvées. Donner les parts d'inertie.
4. Écrire la première composante principale en fonction des variables réduites z_1, z_2, z_3 , puis la corrélation de chaque variable avec elle. Interpréter.
5. Même question pour des corrélations toutes égales à $\rho > 0$: valeur propre et part d'inertie du premier axe. Que se passe-t-il quand ρ tend vers 1 ?

Exercice 8.4. Propriétés des axes principaux. Soit $S = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$ symétrique, de valeurs propres $\lambda_1 \geq \lambda_2$.

1. En développant $(\lambda - \lambda_1)(\lambda - \lambda_2)$, montrer que $\lambda_1 + \lambda_2 = \text{tr}(S)$ et $\lambda_1 \lambda_2 = \det(S)$.
2. Montrer que les deux valeurs propres sont réelles : le discriminant de l'équation vaut $(a - d)^2 + 4b^2$.
3. Soient $\mathbf{u}_1, \mathbf{u}_2$ tels que $S\mathbf{u}_1 = \lambda_1 \mathbf{u}_1, S\mathbf{u}_2 = \lambda_2 \mathbf{u}_2$, avec $\lambda_1 \neq \lambda_2$. En calculant $(S\mathbf{u}_1) \cdot \mathbf{u}_2$ de deux façons, montrer que $\mathbf{u}_1 \perp \mathbf{u}_2$.
4. ACP normée de deux variables de corrélation $r > 0$: calculer les valeurs propres et les axes de $R = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$. Les axes dépendent-ils de r ? Quelle est la part d'inertie du premier axe ?
5. Soit Z la matrice réduite ($n \times p$) et $\mathbf{u}_k$ un vecteur propre unitaire de $R = \frac{1}{n} Z^T Z$, de valeur propre λ_k . Montrer que $\mathbf{c}_k = Z\mathbf{u}_k$ est centrée, de variance λ_k , et que $\mathbf{c}_k$ et $\mathbf{c}_l$ ($k \neq l$) ne sont pas corrélées.

Exercice 8.5. ACP normée des iris : valeurs propres et variables. Les 150 fleurs du jeu `iris` sont décrites par quatre mesures (en cm) des sépales et des pétales.

```
x <- iris[, 1:4]
pca <- prcomp(x, scale. = TRUE)      # ACP normée
round(pca$sdev^2, 3)                  # valeurs propres
round(100 * pca$sdev^2 / 4, 1)        # parts d'inertie (%)
round(cor(x, pca$x)[, 1:2], 2)       # corrélations variables-axes
round(100 * pca$rotation[, 1:2]^2, 1) # contributions des variables (%)
```

Axe	1	2	3	4
λ_k	2,918	0,914	0,147	0,021
Part (%)	73,0	22,9	3,7	0,5

Variable	r axe 1	r axe 2	ctr axe 1 (%)	ctr axe 2 (%)
longueur sépale	0,89	-0,36	27,2	14,2
largeur sépale	-0,46	-0,88	7,3	85,2
longueur pétale	0,99	-0,02	33,7	0,1
largeur pétale	0,96	-0,06	31,9	0,4

1. Vérifier la somme des valeurs propres. Combien d'axes retenir d'après l'éboulis ? D'après le critère $\lambda_k > 1$? Discuter ce second critère.
2. Tracer le cercle des corrélations des axes 1 et 2. Calculer le $\cos^2$ de chaque variable sur le plan. Les variables sont-elles bien représentées ?
3. Interpréter l'axe 1, puis l'axe 2. S'agit-il d'un effet taille ?
4. Moyennes des espèces sur l'axe 1 : setosa -2,22, versicolor 0,49, virginica 1,72. L'espèce n'a pas servi au calcul : que révèle l'axe 1 ?

Exercice 8.6. ACP normée des iris : individus. Suite de l'exercice 5. Pour quelques fleurs, les mesures, les coordonnées sur les axes 1 et 2, le $\cos^2$ sur le plan et les contributions sont :

```
cos2 <- pca$x^2 / rowSums(pca$x^2)
ctr <- sweep(pca$x^2, 2, colSums(pca$x^2), "/")
```

Fleur	Espèce	long. sépale	larg. sépale	long. pétale	larg. pétale
16	setosa	5,7	4,4	1,5	0,4
42	setosa	4,5	2,3	1,3	0,3
61	versicolor	5,0	2,0	3,5	1,0
85	versicolor	5,4	3,0	4,5	1,5
119	virginica	7,7	2,6	6,9	2,3
moyenne		5,84	3,06	3,76	1,20

Fleur	c_{i1}	c_{i2}	$\cos^2$ plan	ctr axe 1 (%)	ctr axe 2 (%)
16	-2,25	-2,68	1,00	1,17	5,26
42	-1,85	2,33	0,99	0,79	3,98
61	-0,11	2,65	1,00	0,00	5,14
85	0,22	0,29	0,21	0,01	0,06
119	3,30	-0,02	0,96	2,50	0,00

1. Placer les cinq fleurs sur le plan principal. Quelle est la contribution moyenne d'une fleur ? Les fleurs 119 et 16 ont les plus fortes contributions aux axes 1 et 2 : relier ces valeurs à leurs mesures.
2. Les fleurs 42 et 61 sont d'espèces différentes. Qu'ont-elles en commun ? Comment l'axe 2 les place-t-il ?
3. La fleur 85 a un $\cos^2$ de 0,21. Sa position sur le plan est-elle fiable ? Sachant que $\|\mathbf{z}_{85}\|^2 = 0,64$ et $c_{85,3} = -0,66$, où se trouve l'essentiel de son écart au centre ?
4. Une nouvelle fleur mesure $6,0 \cdot 2,8 \cdot 4,5 \cdot 1,4$. La commande `predict(pca, nouvelle)` donne les coordonnées (0,65; 0,45) sur le plan. Moyennes des espèces sur l'axe 2 : setosa -0,29, versicolor 0,55, virginica -0,26. De quelle espèce la rapprocher ?

8.2 Pour s'entraîner

Exercice 8.7. ACP normée ou non. Les quatre mesures des iris sont en centimètres. L'ACP sur les variables seulement centrées donne :

```
q <- prcomp(x)                # ACP non normée
round(apply(x, 2, var), 3)     # variances des mesures
round(q$sdev^2, 3)
round(q$rotation[, 1], 2)     # premier axe
```

	longueur sépale	largeur sépale	longueur pétale	largeur pétale
Variance	0,686	0,190	3,116	0,581
$\mathbf{u}_1$	0,36	-0,08	0,86	0,36

Valeurs propres : 4,228 ; 0,243 ; 0,078 ; 0,024.

1. Vérifier que la somme des valeurs propres est la somme des variances.
2. Calculer la part d'inertie du premier axe et la comparer à celle de l'ACP normée (73 %).
3. Quelle variable domine le premier axe ? Pourquoi ?
4. Laquelle des deux analyses choisir ici ? Argumenter.

Exercice 8.8. Retrouver l'ACP avec les formules du cours.

```
n <- nrow(x)
z <- scale(x) * sqrt(n / (n - 1))      # réduction en 1/n, comme le cours
R <- crossprod(z) / n                  # matrice des corrélations
e <- eigen(R)                          # valeurs et vecteurs propres
e$values; sum(e$values)
round(crossprod(e$vectors), 10)        # U transposée fois U
c1 <- z %*% e$vectors[, 1]             # première composante principale
mean(c1^2)
pca$x[1:3, 1] / c1[1:3]
```

1. Comparer `e$values` aux valeurs propres de l'exercice 5, et `sum(e$values)` à la trace de R .
2. La commande `crossprod(e$vectors)` renvoie la matrice identité. Quelle propriété des axes traduit-elle ?
3. `mean(c1^2)` renvoie 2,918. Quelle formule du cours ce résultat illustre-t-il ?
4. Le dernier rapport vaut $\pm 0,9967$ pour chaque fleur. Expliquer cette valeur, sachant que `scale()` divise par $n - 1$.

9 TD 9 — Synthèse : les crabes

Leptograpsus

9.1 Objectifs

Cette séance de synthèse enchaîne, sur un même jeu de données, les méthodes des cours C5 à C9 : réduction et distances, k-means et classification hiérarchique, ACP, effet taille, information mutuelle. Question biologique : les cinq mesures d'un crabe suffisent-elles à retrouver sa forme et son sexe ?

Les exercices 1 à 6 constituent le cœur de la séance et se suivent : distances (1), classification des mesures brutes (2), retrait de la taille (3), ACP (4), classification des formes (5), bilan (6). Les exercices 7 et 8 prolongent le travail et permettent de s'entraîner. Les sorties R utiles sont reproduites : la fiche se traite sur papier ou sur machine.

Les 200 crabes du C9 se répartissent en deux formes (bleue B, orange O) et deux sexes (femelle F, mâle M), 50 par groupe. Cinq mesures, en mm : FL, lobe frontal ; RW, largeur arrière ; CL, longueur de la carapace ; CW, largeur de la carapace ; BD, épaisseur.

```
data(crabs, package = "MASS")
x <- crabs[, c("FL", "RW", "CL", "CW", "BD")]
groupe <- paste(crabs$sp, crabs$sex) # "B F", "B M", "O F", "O M"
```

Exercice 9.1. Réduction et distances. Trois crabes mâles, deux de forme bleue et un de forme orange :

	FL	RW	CL	CW	BD
crabe 8 (B M)	11,6	9,1	24,5	28,4	10,4
crabe 47 (B M)	19,7	15,3	41,9	48,5	17,8
crabe 105 (O M)	12,5	9,4	23,2	26,0	10,8
moyenne	15,58	12,74	32,11	36,41	14,03
écart-type	3,49	2,57	7,10	7,85	3,42

1. Les cinq mesures sont en millimètres. Pourquoi les réduire malgré tout ? Calculer les variables réduites du crabe 8.
2. Les crabes 47 et 105 ont pour variables réduites (1,18; 1,00; 1,38; 1,54; 1,10) et (−0,88; −1,30; −1,25; −1,33; −0,95). Calculer les distances euclidiennes $d(8, 47)$ et $d(8, 105)$.

3. Le crabe 8 est-il plus proche d'un crabe de son groupe ou du crabe orange ? Qu'est-ce que cette distance mesure surtout ?

Exercice 9.2. k-means sur les mesures réduites.

```
z <- scale(x)
set.seed(1)
km <- kmeans(z, centers = 4, nstart = 25)
table(km$cluster, groupe)
round(km$centers, 2)
km$betweenss / km$totss # 0,874
```

Classe	B F	B M	O F	O M	FL	RW	CL	CW	BD
1	12	9	2	5	-1,55	-1,58	-1,60	-1,61	-1,57
2	16	19	21	15	0,33	0,35	0,35	0,36	0,31
3	3	7	18	13	1,38	1,28	1,36	1,34	1,39
4	19	15	9	17	-0,61	-0,55	-0,59	-0,59	-0,59

À gauche, le croisement des classes avec les groupes ; à droite, les centres des classes (variables réduites). Les numéros de classes sont arbitraires.

1. Les classes retrouvent-elles les quatre groupes ?
2. Lire les centres : qu'est-ce qui distingue les quatre classes ?
3. La part d'inertie expliquée $B/I = 0,874$ est élevée. Est-ce contradictoire ? Relier au premier axe de l'ACP des crabes au C9 (95,8 % de l'inertie).

Exercice 9.3. Retirer l'effet taille. Le C9 propose de remplacer chaque mesure par son logarithme népérien, diminué de la moyenne des cinq logarithmes du même crabe :

$$f_{ij} = \ln x_{ij} - \frac{1}{5} \sum_{k=1}^5 \ln x_{ik}.$$

```
f <- log(x) - rowMeans(log(x))
```

1. Calculer les cinq valeurs f_{8j} du crabe 8.
2. Montrer que f_{ij} ne change pas si toutes les mesures du crabe i sont multipliées par un même nombre $a > 0$. Montrer que $\sum_j f_{ij} = 0$ pour chaque crabe.
3. Pour les crabes 47 et 105, R donne $(-0,260; -0,513; 0,494; 0,641; -0,362)$ et $(-0,184; -0,469; 0,435; 0,548; -0,330)$. Comparer au crabe 8. Après réduction des colonnes de f , les distances deviennent $d(8, 47) = 0,31$, $d(8, 105) = 4,23$, $d(47, 105) = 4,43$. Conclure.

Exercice 9.4. ACP des variables de forme.

```
pca <- prcomp(f, scale. = TRUE)
round(pca$sdev^2, 3)
round(cor(f, pca$x)[, 1:2], 2)
aggregate(pca$x[, 1:2], list(groupe), mean)
```

Axe	1	2	3	4	5
λ_k	2,255	2,127	0,523	0,095	0,000
Part (%)	45,1	42,5	10,5	1,9	0,0

	FL	RW	CL	CW	BD
r axe 1	-0,36	1,00	-0,77	-0,28	-0,68
r axe 2	0,76	-0,06	-0,59	-0,93	0,57

Moyennes	B F	B M	O F	O M
axe 1	1,46	-1,00	1,01	-1,46
axe 2	-1,11	-1,58	1,48	1,20

1. Combien d'axes retenir ? Comparer aux parts d'inertie de l'ACP des mesures brutes du C9 (95,8 %, 3,0 %, 0,9 %). Qu'est devenu l'axe de la taille ?
2. Tracer le cercle des corrélations. Interpréter les axes 1 et 2.
3. Placer les moyennes des quatre groupes sur le plan. Quel axe sépare les sexes, lequel sépare les formes ? Rapprocher des axes 2 et 3 du C9.
4. D'après l'exercice 3, pourquoi la dernière valeur propre est-elle nulle ?

Exercice 9.5. Classification des formes.

```
z <- scale(f)
set.seed(1)
km <- kmeans(z, centers = 4, nstart = 25)
table(km$cluster, groupe)
km$betweenss / km$totss # 0,738
W <- sapply(1:6, function(k) kmeans(z, k, nstart = 25)$tot.withinss)
h <- hclust(dist(z), method = "average")
table(cutree(h, k = 4), groupe)
```

k-means	B F	B M	O F	O M	saut moyen	B F	B M	O F	O M
1	50	14	1	0	1	49	7	1	0
2	0	0	0	42	2	0	43	0	0
3	0	36	0	0	3	1	0	0	0
4	0	0	49	8	4	0	0	49	50

K	1	2	3	4	5	6
W, variables de forme	995	624	398	260	223	198
W, mesures brutes réduites	995	352	198	126	95	82

Longueur de carapace médiane des mâles classés avec les femelles : 24 mm (forme bleue) et 23 mm (forme orange), contre 36 mm pour les autres mâles.

1. Associer chaque classe des k-means à un groupe. Combien de crabes sont classés avec leur groupe ? Quelles erreurs subsistent ? Proposer une explication biologique.
2. B/I vaut 0,738, contre 0,874 à l'exercice 2. Quelle partition est la meilleure ? Que conclure sur ce critère ?
3. Où se situe le coude de chaque courbe $W(K)$? Interpréter la différence.
4. Comparer la classification hiérarchique au saut moyen aux k-means.

Exercice 9.6. Bilan.

1. Résumer la chaîne d'analyse : pour chaque représentation des crabes (mesures réduites, variables de forme), la distance obtenue, les classes, les axes.
2. Pourquoi comparer B/I entre les exercices 2 et 5 n'a-t-il pas de sens ? Quel critère permet ici de juger une partition ?
3. Rédiger en quelques lignes la réponse à la question de la séance, en citant un résultat chiffré par méthode.
4. Parmi les cinq étapes de la grille de lecture d'une ACP (annexe B), laquelle a été décisive ?

9.2 Pour s'entraîner

Exercice 9.7. Une valeur propre nulle. Soit F la matrice $n \times 5$ des variables de forme ; d'après l'exercice 3, $F\mathbf{1} = \mathbf{0}$. Soient $\mathbf{g}$ son centre de gravité, $F_c = F - \mathbf{1}\mathbf{g}^T$, $s_1, \dots, s_5$ les écarts-types des colonnes, $\mathbf{s} = (s_1, \dots, s_5)$ et Z la matrice réduite, de colonnes $\mathbf{z}^j = (\mathbf{f}^j - \bar{f}_j\mathbf{1})/s_j$.

1. Montrer que $\sum_j \bar{f}_j = 0$, puis que $F_c\mathbf{1} = \mathbf{0}$.
2. Montrer que $Z\mathbf{s} = \mathbf{0}$.
3. En déduire que $\mathbf{s}$ est un vecteur propre de $R = \frac{1}{n}Z^T Z$, de valeur propre 0.

4. R donne $\mathbf{s} = (0,032; 0,073; 0,026; 0,033; 0,046)$. Quelle combinaison des variables réduites a une variance nulle ? Relier au pivot de Gauss : que dire du système $R\mathbf{v} = \mathbf{0}$?

Exercice 9.8. Information apportée par une classification. Divisé par 200, un tableau croisant classes et groupes devient une loi jointe (C7).

```
H <- function(p) { p <- p[p > 0]; -sum(p * log2(p)) }  
J <- table(km$cluster, groupe) / 200  
H(rowSums(J)); H(colSums(J)); H(J)
```

1. Calculer l'entropie des groupes.
2. Pour les k-means sur les variables de forme (exercice 5), calculer les fréquences des classes, puis $H(\text{classe})$, $H(\text{classe}, \text{groupe})$ et l'information mutuelle.
3. Pour les k-means sur les mesures brutes (exercice 2), R donne $H(\text{classe}) = 1,917$ et $H(\text{classe}, \text{groupe}) = 3,816$. Calculer l'information mutuelle.
4. Quelle valeur prendrait l'information mutuelle si les classes coïncidaient exactement avec les groupes ? Comparer les deux classifications.

A Notions d'algèbre linéaire à connaître

A.1 C1 — Vecteurs

- Vecteur de $\mathbb{R}^n$: $\mathbf{x} = (x_1, \dots, x_n)$. Somme et produit par un réel : coordonnée par coordonnée.
- Combinaison linéaire : $\alpha_1 \mathbf{v}_1 + \dots + \alpha_p \mathbf{v}_p$, $\alpha_i \in \mathbb{R}$.
- Trois lectures d'un vecteur : un profil, un état, une loi de probabilité.

A.2 C2 — Systèmes linéaires et pivot de Gauss

- Système de m équations à n inconnues : $\sum_j a_{ij}x_j = b_i$, $i = 1, \dots, m$. Résoudre : décrire l'ensemble $\mathcal{S}$ des solutions.
- Systèmes équivalents : même ensemble de solutions. Opérations élémentaires, toutes réversibles : échange $L_i \leftrightarrow L_j$, dilatation $L_i \leftarrow \lambda L_i$ ($\lambda \neq 0$), transvection $L_i \leftarrow L_i + \lambda L_j$.
- Matrice augmentée $[A \mid \mathbf{b}]$, forme échelonnée, pivots. Variable liée : colonne avec pivot ; variable libre : les autres.
- Théorème : compatible $\iff$ aucun pivot dans la colonne du second membre ; solution unique $\iff$ compatible et aucune variable libre. Toujours 0, 1 ou une infinité de solutions.
- Infinité de solutions : forme vectorielle $\mathbf{x} = \mathbf{p} + t \mathbf{d}$, $t \in \mathbb{R}$ (une variable libre).

A.3 C3 — Matrices et chaînes de Markov

- Produit matrice-vecteur : $A\mathbf{x} = x_1 \mathbf{a}_1 + \dots + x_p \mathbf{a}_p$, $(A\mathbf{x})_i = \sum_j a_{ij}x_j$. Un système s'écrit $A\mathbf{x} = \mathbf{b}$.
- Produit matriciel $AB = (A\mathbf{b}_1 \dots A\mathbf{b}_q)$; en général $AB \neq BA$. Transposée $(A^T)_{ij} = a_{ji}$, $(AB)^T = B^T A^T$. Identité : $I\mathbf{x} = \mathbf{x}$.
- Chaîne de Markov : $\mathbf{x}_{t+1} = P \mathbf{x}_t$, P de colonnes de somme 1 ; $\mathbf{x}_t = P^t \mathbf{x}_0$.
- Système homogène : second membre nul, toujours solution $\mathbf{0}$. Équilibre : $P\pi = \pi$, soit $(P - I)\pi = \mathbf{0}$ avec $\sum_i \pi_i = 1$.
- Vecteur propre : $A\mathbf{v} = \lambda \mathbf{v}$, $\mathbf{v} \neq \mathbf{0}$. L'équilibre est un vecteur propre de valeur propre 1.
- Perron-Frobenius : nombre fini d'états, transitions fixes, tout état accessible depuis tout autre, pas de cycle fixe $\Rightarrow$ équilibre unique, limite de la chaîne pour toute loi initiale.

A.4 C4 — Géométrie de $\mathbb{R}^p$

- Norme $\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_p^2}$, distance $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.
- Produit scalaire $\mathbf{x} \cdot \mathbf{y} = \sum_i x_i y_i = \mathbf{x}^T \mathbf{y}$; $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2\mathbf{x} \cdot \mathbf{y} + \|\mathbf{y}\|^2$.
- Cauchy-Schwarz : $|\mathbf{x} \cdot \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|$, égalité si et seulement si les vecteurs sont colinéaires. Similarité cosinus $s = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \in [-1, 1]$, angle $\theta = \arccos s$.
- Orthogonalité : $\mathbf{x} \cdot \mathbf{y} = 0$. Pythagore : $\mathbf{x} \perp \mathbf{y} \Rightarrow \|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$.
- Projection sur la droite dirigée par $\mathbf{u}$: $\frac{\mathbf{u} \cdot \mathbf{x}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u}$, point de la droite le plus proche de $\mathbf{x}$. Centrage : $\mathbf{x} - \bar{x} \mathbf{1}$, orthogonal à $\mathbf{1}$. Coordonnées d'un nuage sur un axe unitaire : $X\mathbf{u}$.
- Moindres carrés : $X = (\mathbf{1} \ \mathbf{x})$, équations normales $X^T X \hat{\beta} = X^T \mathbf{y}$; résidu orthogonal aux colonnes de X .

A.5 C5 — Distances et similarités

- Distance : positive, séparante ($d = 0 \iff \mathbf{x} = \mathbf{y}$), symétrique, inégalité triangulaire.
- Sur $\mathbb{R}^p$: $d_1 = \sum_i |x_i - y_i|$, d_2 euclidienne, $d_\infty = \max_i |x_i - y_i|$.
- Écart-type $s_j = \frac{1}{\sqrt{n}} \|\mathbf{x}^j - \bar{x}_j \mathbf{1}\|$; réduction $x_{ij}^* = x_{ij}/s_j$: distances indépendantes des unités.
- Hamming : nombre de positions différentes. Jaccard : $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, distance $1 - J$.
- Similarité issue d'une distance : $s = 1/(1 + d)$. Corrélation : cosinus des vecteurs centrés.
- Matrice des distances D : symétrique, diagonale nulle.

A.6 C6 — Clustering

- Centre de gravité $\mathbf{g} = \frac{1}{n} \sum_i \mathbf{x}_i$. Inertie $I = \frac{1}{n} \sum_i \|\mathbf{x}_i - \mathbf{g}\|^2 = \sum_j s_j^2$.
- Intra-classe $W = \frac{1}{n} \sum_k \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2$, inter-classe $B = \frac{1}{n} \sum_k n_k \|\mathbf{g}_k - \mathbf{g}\|^2$. Huygens : $I = W + B$.
- k-means : affectation au centre le plus proche, recentrage ; W décroît à chaque étape ; convergence vers un minimum local.
- Classification ascendante hiérarchique : fusions successives des deux classes les plus proches (saut minimal, saut maximal, moyenne), dendrogramme. UPGMA : CAH par la moyenne.

A.7 C7 — Information et entropie

- Surprise $-\log_2 p$. Entropie $H(\mathbf{p}) = -\sum_i p_i \log_2 p_i$, $0 \leq H \leq \log_2 K$; maximale pour la loi uniforme.
- Contenu en information d'une position d'ADN : $R = 2 - H$.

- Loi jointe : matrice J , marges $J\mathbf{1}$ et $J^T\mathbf{1}$. Indépendance : $J = \mathbf{p}\mathbf{q}^T$, matrice de rang 1. Colonnes normalisées : lois conditionnelles, matrice de transition.
- Information mutuelle $I(X; Y) = H(X) + H(Y) - H(X, Y) \geq 0$, nulle si et seulement si indépendance.
- Loi normale : à variance fixée, entropie maximale, $H = \frac{1}{2} \log_2(2\pi e \sigma^2)$.

A.8 C8 — Variance, covariance, corrélation

- Matrice des données X ($n \times p$), matrice centrée X_c .
- Variance $s_j^2 = \frac{1}{n} \|\mathbf{x}^j - \bar{x}_j \mathbf{1}\|^2$; covariance : produit scalaire des vecteurs centrés, divisé par n ; corrélation $r_{jk} = s_{jk}/(s_j s_k)$.
- Matrice de covariance $S = \frac{1}{n} X_c^T X_c$, symétrique ; matrice des corrélations $R = \frac{1}{n} Z^T Z$, Z réduite.
- Inertie totale : $\text{tr}(S) = \sum_j s_j^2 = I$.
- Axe unitaire $\mathbf{u}$: coordonnées $X_c \mathbf{u}$, variance projetée $\mathbf{u}^T S \mathbf{u}$.

A.9 C9 — Analyse en composantes principales

- Axe de variance maximale : $\max_{\|\mathbf{u}\|=1} \mathbf{u}^T S \mathbf{u} = \lambda_1$, atteint pour $S \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$.
- Déterminant $\det \begin{pmatrix} p & q \\ r & s \end{pmatrix} = ps - qr$. Valeurs propres 2×2 : $\lambda^2 - \text{tr}(A) \lambda + \det(A) = 0$; vecteur propre par Gauss sur $(A - \lambda I) \mathbf{v} = \mathbf{0}$.
- S symétrique : valeurs propres réelles, positives ou nulles ; axes principaux orthogonaux. $\lambda_1 + \dots + \lambda_p = \text{tr}(S)$; part d'inertie de l'axe k : $\lambda_k / \text{tr}(S)$.
- ACP normée : R au lieu de S , $\text{tr}(R) = p$. Composante principale $\mathbf{c}_k = Z \mathbf{u}_k$, de variance λ_k ; composantes non corrélées.
- Individus : $c_{ik} = \mathbf{z}_i \cdot \mathbf{u}_k$, $\cos_{ik}^2 = c_{ik}^2 / \|\mathbf{z}_i\|^2$, $\text{ctr}_{ik} = c_{ik}^2 / (n \lambda_k)$.
- Variables : $r(\mathbf{z}^j, \mathbf{c}_k) = \sqrt{\lambda_k} u_{jk}$, contribution u_{jk}^2 .

A.10 Autres outils mathématiques

- Logarithme de base 2 : $\log_2(ab) = \log_2 a + \log_2 b$, $\log_2 2^k = k$, $\log_2 x = \ln x / \ln 2$.
- Loi de probabilité sur K modalités : $p_i \geq 0$, $\sum_i p_i = 1$.
- Moyenne $\bar{x} = \frac{1}{n} \sum_i x_i$, variance $\frac{1}{n} \sum_i (x_i - \bar{x})^2$.
- Symbole $\sum$: linéarité, $\sum_i (\alpha a_i + \beta b_i) = \alpha \sum_i a_i + \beta \sum_i b_i$; échange de deux sommes finies.

B Compétences à maîtriser

Chaque compétence : une méthode, un point de vigilance, un renvoi au cours. Les commandes R portent sur un tableau numérique $\mathbf{x}$ (individus en lignes, variables en colonnes).

B.1 Résoudre un système par le pivot de Gauss

1. Écrire la matrice augmentée $[A \mid \mathbf{b}]$, une colonne par inconnue, dans un ordre fixé.
2. Descente : choisir un pivot non nul, annuler les coefficients sous lui par $L_i \leftarrow L_i - \frac{a_{i1}}{a_{11}} L_1$; recommencer sur la sous-matrice.
3. Lire la forme échelonnée : pivot dans la dernière colonne, aucune solution ; sinon, une variable libre par colonne sans pivot.
4. Remontée : exprimer les variables liées en fonction des variables libres ; écrire les solutions sous forme vectorielle.
5. Vérifier dans le système **initial**.

Vigilance : noter chaque opération ; ne jamais multiplier une ligne par 0. Renvoi : C2, C3 (équilibre), C9 (vecteur propre) ; TD1, TD8.

B.2 Calculer produit scalaire, norme, cosinus, projection

1. $\mathbf{x} \cdot \mathbf{y} = \sum_i x_i y_i$; $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$.
2. Cosinus : $\frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$; signe positif, angle aigu ; nul, vecteurs orthogonaux.
3. Projection sur la droite de $\mathbf{u}$: $\frac{\mathbf{u} \cdot \mathbf{x}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u}$; contrôle : $\mathbf{x} - \text{projeté}$ est orthogonal à $\mathbf{u}$.

Vigilance : le cosinus compare des formes de profils, pas des niveaux. Renvoi : C4, C5.

B.3 Centrer et réduire un tableau de données

Tableau X : n individus en lignes, p variables en colonnes.

1. Moyenne de chaque colonne : $\bar{x}_j = \frac{1}{n} \sum_i x_{ij}$.
2. Centrer : $x_{ij} - \bar{x}_j$. La colonne centrée $\mathbf{x}^j - \bar{x}_j \mathbf{1}$ est orthogonale à $\mathbf{1}$: sa moyenne est nulle.
3. Écart-type : $s_j = \sqrt{\frac{1}{n} \sum_i (x_{ij} - \bar{x}_j)^2}$.
4. Réduire : $z_{ij} = (x_{ij} - \bar{x}_j) / s_j$, nombre sans unité.

5. Contrôler : chaque colonne de Z a une moyenne 0 et une variance 1.

```
xc <- sweep(x, 2, colMeans(x))      # centrage
s  <- sqrt(colMeans(xc^2))          # écarts-types en 1/n, comme le cours
z  <- sweep(xc, 2, s, "/")          # réduction
colMeans(z); colMeans(z^2)          # contrôles : 0 et 1
scale(x)                            # version de R, écart-type en n - 1
```

Vigilance : `scale()` et `sd()` divisent par $n-1$; l'écart est négligeable pour n grand et sans effet sur les corrélations. Réduire efface les unités et donne le même poids à chaque variable. Renvoi : C4, C5, C8.

B.4 Calculer une corrélation

1. Centrer les deux variables $\mathbf{x}$ et $\mathbf{y}$.
2. Covariance : $s_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$, produit scalaire des vecteurs centrés divisé par n .
3. Corrélation : $r = \frac{s_{xy}}{s_x s_y}$, cosinus des vecteurs centrés ; $r \in [-1, 1]$ par Cauchy-Schwarz.
4. Lire : le signe donne le sens de la liaison ; $|r|$ proche de 1, points proches d'une droite ; r proche de 0, pas de liaison linéaire.
5. Plusieurs variables : matrice des corrélations $R = \frac{1}{n} Z^T Z$, symétrique, de diagonale 1.

```
cor(x[, 1], x[, 2])                 # une corrélation
R <- cor(x)                         # matrice des corrélations
crossprod(as.matrix(z)) / nrow(z)  # la même, à partir de z
round(R, 2)
```

Vigilance : r ne voit que les liaisons linéaires ($y = x^2$ sur des valeurs symétriques donne $r = 0$) ; une valeur extrême peut créer ou masquer une corrélation ; corrélation n'est pas causalité. Renvoi : C5, C8.

B.5 Calculer une distance

1. Choisir la représentation : vecteur numérique, séquence, ensemble.
2. Numérique : d_1 , d_2 ou d_∞ . Séquences alignées : Hamming. Ensembles : Jaccard $1 - \frac{|A \cap B|}{|A \cup B|}$.
3. Variables d'unités différentes : réduire d'abord, $x_{ij}^* = x_{ij}/s_j$.
4. Ranger les distances dans la matrice D : symétrique, diagonale nulle.

```
d <- dist(scale(x))           # euclidienne sur variables réduites
dist(x, method = "manhattan") # d1 ; "maximum" pour d-infini
```

Vigilance : changer d'unité peut changer le plus proche voisin. Renvoi : C5.

B.6 Utiliser et interpréter les k-means

1. Réduire les variables si leurs unités diffèrent.
2. Fixer K ; lancer plusieurs initialisations et garder la partition de plus petite inertie intra-classe W .
3. Choisir K : tracer W en fonction de K ; retenir le coude.
4. Qualité : part d'inertie expliquée B/I (Huygens, $I = W + B$).
5. Interpréter : effectifs, centres des classes, confrontation à une classification connue.

```
km <- kmeans(scale(x), centers = 3, nstart = 25)
km$cluster           # classe de chaque individu
km$centers           # centres (variables réduites)
km$betweenss / km$totss # B / I
W <- sapply(1:6, function(k) kmeans(scale(x), k, nstart = 25)$tot.withinss)
plot(1:6, W, type = "b") # coude
table(km$cluster, groupe) # comparaison à un facteur connu
```

Vigilance : W décroît toujours quand K augmente ; un seul départ peut bloquer sur un minimum local ; `tot.withinss` vaut nW . La CAH (`h <- hclust(dist(scale(x)), method = "average")`), puis `cutree(h, k = 3)` donne une hiérarchie au lieu d'une partition. Renvoi : C6.

B.7 Lancer et interpréter une ACP

Lancer.

1. Choisir : ACP normée (unités différentes, cas usuel) ou sur variables centrées seulement (même unité, variances comparables).
2. Mettre de côté les variables qualitatives (espèce, sexe) : elles serviront d'éléments supplémentaires.

```
pca <- prcomp(x, scale. = TRUE) # ACP normée
lambda <- pca$sdev^2           # valeurs propres
100 * lambda / sum(lambda)     # parts d'inertie (%) ; summary(pca)
barplot(lambda)                # éboulis
pca$rotation                   # axes u_k (coefficients u_jk)
pca$x                          # composantes principales c_ik
```

Interpréter, dans cet ordre :

1. **Valeurs propres** : parts d'inertie, éboulis ; retenir les axes avant le coude ou de valeur propre supérieure à 1 (ACP normée).
2. **Variables** : corrélations variable-axe $\sqrt{\lambda_k} u_{jk}$, cercle des corrélations. Flèche proche du cercle : variable bien représentée ; flèches groupées : variables corrélées ; contributions u_{jk}^2 pour nommer un axe.

```
cor(x, pca$x)[, 1:2]          # coordonnées sur le cercle
pca$rotation[, 1:2]^2        # contributions des variables
```

3. **Individus** : plan principal ; vérifier le $\cos^2$ avant d'interpréter une proximité ; repérer les contributions fortes et isolées.

```
cos2 <- pca$x^2 / rowSums(pca$x^2)
ctr  <- sweep(pca$x^2, 2, colSums(pca$x^2), "/")
plot(pca$x[, 1:2], col = groupe) # plan principal, par groupe
```

4. **Facteur commun** : toutes les variables du même côté de l'axe 1, effet taille ; lire alors les axes suivants.
5. **Éléments supplémentaires** : couleurs du plan pour un facteur connu ; nouveaux individus projetés sans modifier les axes, par `predict(pca, nouveaux)`.

Vigilance : l'ACP ne voit que les liens linéaires ; une grande part d'inertie n'est pas forcément la question biologique ; le choix normée ou non décide du poids des variables. Renvoi : C8, C9 ; TD8, TD9.

C Vérifier avec sympy

sympy (Python) résout un système linéaire de façon **formelle** : fractions exactes, pas d'arrondi, et les variables libres restent des paramètres symboliques. Installation : `pip install sympy`.

C.1 Résoudre un système : linsolve

`linsolve((A, b), x, y, z)` résout $Ax = b$. b est le second membre, un vecteur colonne d'autant d'entrées que d'équations ; nul, il donne un système **homogène**.

Exemple — l'équilibre de la succession (C3) : $(P - I)\pi = 0$ avec $5(P - I) = \begin{pmatrix} -2 & 0 & 1 \\ 2 & -2 & 0 \\ 0 & 2 & -1 \end{pmatrix}$.

```
from sympy import Matrix, symbols, linsolve

A = Matrix([[-2, 0, 1],
            [ 2, -2, 0],
            [ 0, 2, -1]]) # 5 (P - I)
pi1, pi2, pi3 = symbols("pi1 pi2 pi3")

linsolve((A, Matrix([0, 0, 0])), pi1, pi2, pi3)
# {(pi3/2, pi3/2, pi3)}
```

π_3 reste un symbole : $\pi_1 = \pi_2 = \pi_3/2$. La contrainte $\pi_1 + \pi_2 + \pi_3 = 1$ donne $\pi = (1/4; 1/4; 1/2)$.

C.2 Lire la forme échelonnée : rref

```
R, pivots = A.rref()
R          # [[1, 0, -1/2], [0, 1, -1/2], [0, 0, 0]]
pivots     # (0, 1) : colonnes à pivot ; la colonne 2 (pi3) est libre
```

`rref` renvoie la forme échelonnée **réduite** et les indices des colonnes pivots : autant de variables libres que de colonnes sans pivot.

C.3 Reproduire le pivot : elementary_row_op

```
M = A.copy()
M = M.elementary_row_op("n->n+km", row=1, k=1, row2=0) # L2 <- L2 + L1
M = M.elementary_row_op("n->n+km", row=2, k=1, row2=1) # L3 <- L3 + L2
# [-2  0  1]
# [ 0 -2  1]
# [ 0  0  0]
```

sympy ne détaille pas les étapes tout seul; `elementary_row_op` applique une opération élémentaire à la fois (ici deux transvections), comme à la main.

C.4 Un pivot détaillé automatiquement

La fonction ci-dessous applique les opérations élémentaires une à une et les affiche : descente (annulation sous les pivots), forme échelonnée, remontée (normalisation puis annulation au-dessus), forme échelonnée réduite. Un système incompatible est signalé (« pivot dans la dernière colonne »).

```
from sympy import Matrix, Rational

def gauss_detaille(A, b):
    """Élimination de Gauss détaillée du système A x = b."""
    M = Matrix(A).row_join(Matrix(b)).applyfunc(Rational)
    n, m = M.shape
    ninconnues = m - 1

    def transvection(i, ref, c):
        # L_i <- L_i - c L_ref
        if c > 0:
            coef = "" if c == 1 else f"({c}) "
            return f"L{i+1} <- L{i+1} - {coef}L{ref+1}"
        coef = "" if c == -1 else f"(-{c}) "
        return f"L{i+1} <- L{i+1} + {coef}L{ref+1}"

    def etape(desc):
        print("\n" + desc); print(M)

    print("Matrice augmentée [A | b] :"); print(M)

    pivots, r = [], 0
    for j in range(ninconnues):
        p = next((i for i in range(r, n) if M[i, j] != 0), None)
        if p is None:
```

```

        continue # colonne sans pivot : variable libre
    if p != r:
        M = M.elementary_row_op("n<->m", row1=r, row2=p)
        etape(f"L{r+1} <-> L{p+1}")
    pivots.append((r, j))
    for i in range(r + 1, n):
        if M[i, j] != 0:
            c = M[i, j] / M[r, j]
            M = M.elementary_row_op("n->n+km", row=i, k=-c, row2=r)
            etape(transvection(i, r, c))
    r += 1
    if r == n:
        break

    if any(all(M[i, j] == 0 for j in range(ninconnues)) and M[i, ninconnues] != 0
           for i in range(n)):
        print("\nAucune solution : pivot dans la dernière colonne.")
        return M

    print("\nForme échelonnée :"); print(M)

    for (rr, j) in reversed(pivots):
        p = M[rr, j]
        if p != 1:
            M = M.elementary_row_op("n->kn", row=rr, k=1/p)
            etape(f"L{rr+1} <- (1/p) L{rr+1}")
        for i in range(rr):
            if M[i, j] != 0:
                c = M[i, j]
                M = M.elementary_row_op("n->n+km", row=i, k=-c, row2=rr)
                etape(transvection(i, rr, c))

    print("\nForme échelonnée réduite :"); print(M)
    return M

```

Exemple — l'équilibre de la succession :

```
gauss_detaille(Matrix([[-2, 0, 1], [2, -2, 0], [0, 2, -1]]), Matrix([0, 0, 0]))
```

Matrice augmentée [A | b] :

```
Matrix([[-2, 0, 1, 0], [2, -2, 0, 0], [0, 2, -1, 0]])
```

```
L2 <- L2 + L1
```

```
Matrix([[-2, 0, 1, 0], [0, -2, 1, 0], [0, 2, -1, 0]])
```

```
L3 <- L3 + L2
Matrix([[ -2, 0, 1, 0], [0, -2, 1, 0], [0, 0, 0, 0]])
```

Forme échelonnée :

```
Matrix([[ -2, 0, 1, 0], [0, -2, 1, 0], [0, 0, 0, 0]])
```

```
L2 <- (-1/2) L2
```

```
Matrix([[ -2, 0, 1, 0], [0, 1, -1/2, 0], [0, 0, 0, 0]])
```

```
L1 <- (-1/2) L1
```

```
Matrix([[1, 0, -1/2, 0], [0, 1, -1/2, 0], [0, 0, 0, 0]])
```

Forme échelonnée réduite :

```
Matrix([[1, 0, -1/2, 0], [0, 1, -1/2, 0], [0, 0, 0, 0]])
```

La forme réduite se lit $\pi_1 = \pi_3/2$ et $\pi_2 = \pi_3/2$; la contrainte $\pi_1 + \pi_2 + \pi_3 = 1$ donne $\pi = (1/4; 1/4; 1/2)$.

C.5 Coefficients symboliques

Avec des lettres, la solution s'exprime en fonction des paramètres. Canal à deux états : probabilité de fermeture p , d'ouverture q .

```
p, q = symbols("p q")
P = Matrix([[1 - p, q],
            [p, 1 - q]])
linsolve((P - Matrix.eye(2), Matrix([0, 0])), symbols("f o"))
# {(o*q/p, o)}
```

$f = \frac{q}{p} o$, donc $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$: l'équilibre en fonction des probabilités de transition.

Vigilance : `Matrix([0, 0])` est le second membre, pas un réglage; sa longueur vaut le nombre d'équations. `linsolve` renvoie un ensemble (accolades) de tuples; `EmptySet` signifie aucune solution.

Renvoi : C2 (pivot de Gauss), C3 (équilibre); TD1.