Problème 1 : analyse en composantes principales

1) $(M^\top M)^\top = M^\top (M^\top)^t op = M^\top M$. Donc la matrice $M^\top M$ est égale à sa transposée donc elle est symétrique.

   $M^\top$ a q lignes et p colonnes, et M a p lignes et q colonnes donc $M^\top M$ a q lignes et q colonnes.

2) $\langle w_i | w_j \rangle = (Mv_i)^\top (Mv_j) = v_i^\top M^\top M v_j = v_i^\top \lambda_j v_j = \lambda_j (v_i|v_j) = \lambda_j \delta_{i,j}$ (car $(v_1, \dots, v_q)$ est orthonormée).

3) D'après le théorème du rang, $\dim(\ker M) = 0$.

4) Soit $j \in [\![1, q]\!]$. Puisque $(v_j|v_j) = 1$ il vient $v_j \neq \vec{0}_q$ donc d'après le résultat de la question précédente, $w_j = M_j \neq \vec{0}_p$.

   Appliquons le résultat de la question 2 : il vient $\lambda_j = \langle w_j|w_j\rangle > 0$.

5) a) Soit $j \in [\![1, q]\!]$. $MM^\top w_j = MM^\top M v_j = M\lambda_j v_j = \lambda_j w_j$.

   b) Soient $t_1, \dots, t_q \in \mathbb{R}$ tels que $t_1 w_1 + \dots + t_q w_q = \vec{0}_p$.
   Alors $t_1 w_1 + \dots + t_q w_q = t_1 M v_1 + \dots + t_q M v_q = M(t_1 v_1 + \dots + t_q v_q) = \vec{0}_p$ donc avec le résultat de la question 3 il vient $t_1 v_1 + \dots + v_q v_q = \vec{0}_q$. Or $(v_1, \dots, v_q)$ est une base (orthonormée) donc $t_1 = \dots = t_q = 0$. On a bien montré que $(w_1, \dots, w_q)$ est libre.
   Remarquons (avec 5a) que $\forall j \in [\![1, q]\!], w_j = \dfrac{1}{\lambda_j} MM^\top w_j \in Im(MM^\top)$ donc $(w_1, \dots, w_q)$ est une famille libre de $Im(MM^\top)$ donc $\dim(Im(MM^\top)) \geqslant q$.
   Par ailleurs (en tant que rang d'un produit) $\mathrm{rg}(MM^\top) \leqslant \min(\mathrm{rg}(M), \mathrm{rg}(M^\top)) = q$.
   Par double inégalité, $\mathrm{rg}(MM^\top) = q$.

6) Constatons que puisque M possède p lignes, $\mathrm{rg}(M) \leqslant p$ donc $\mathrm{rg}(M^\top M) \leqslant \mathrm{rg}(M) \leqslant p$ donc $q \leqslant p$.

   a) Soit $X \in \ker(MM^\top)$. Pour tout $i \in [\![1, q]\!]$ il vient $X^\top MM^\top w_i = \lambda_i \langle X|v_i\rangle$, mais par ailleurs $X^\top MM^\top w_i = (MM^\top X)^\top w_i = \vec{0}_p^\top w_i = 0$ donc $\langle X|v_i\rangle = 0$ donc $X \in F^\perp$.
   Donc $\ker(MM^\top) \subset F^\perp$.
   Or $\dim(\ker(MM^\top)) = p - \mathrm{rg}(MM^\top)$, mais $\mathrm{rg}(MM^\top) \leqslant \mathrm{rg}(M) \leqslant q$ (car M a q colonnes) donc $\dim(\ker(MM^\top)) \geqslant p - q$.
   Puisque $(w_1, \dots, w_q)$ est libre, il vient $\dim(F^\perp) = p - \dim(F) = p - q$. Donc $\ker(MM^\top) = F^\perp$.

   b) Pour tout $X \in \mathbb{R}^p$ notons $|X| = \sqrt{\langle X|X\rangle}$ sa norme euclidienne.
   Pour $i \leqslant q$ on sait que $w_i \neq \vec{0}_p$ (d'après le résultat de la question 3) donc on peut poser $w_i' = \dfrac{1}{|w_i|} w_i$ le normalisé de w_i : on a vu à la question 2 que $(w_1, \dots, w_q)$ est orthogonale donc $(w_1', \dots, w_q')$ est orthonormée. Par ailleurs $MM^\top w_i' = \dfrac{1}{|w_i|} MM^\top w_i = \dfrac{1}{|w_i|} \lambda_i w_i = \mu_i w_i'$.
   Construisons alors une base orthonormée $(w_{q+1}', \dots, w_p')$ de $\ker(MM^\top)$:

Ainsi $\forall i \in [\![q+1,p]\!], MM^\top w'_i = \vec{0}_p = \mu_i w'_i$.
Puisque $\ker(MM^\top) = F^\perp$, $(\underbrace{w'_1,\dots,w'_q}_{\text{b.o.n de } F}, \underbrace{w'_{q+1},\dots,w'_p}_{\text{b.o.n de } F^\perp})$ est une base orthonormée de $\mathbb{R}^p$.

7) Puisque $(v_1,\dots,v_q)$ est une base orthonormée de $\mathbb{R}^q$, les coordonnées de X dans cette base sont $((X|v_1),\dots,(X|v_q))$. Par calcul de la norme euclidienne d'un vecteur dans n'importe quelle base orthonormée, il vient $\|X\|^2 = \sum_{j=1}^{q}(X|v_j)^2$.

8) Soient $i,j \in [\![1,q]\!]$, le coefficient ligne i colonne j de $M^\top M$ est $\sum_{k=1}^{p} x_{k,i}x_{k,j}$.

Donc $\mathrm{tr}(M^\top M) = \sum_{i=1}^{q}\sum_{k=1}^{p} x_{k,i}^2 = \sum_{k=1}^{p}\|I_k\|^2 = \tilde{N}$.

Soit ϕ l'endomorphisme de $\mathbb{R}^q$ canoniquement associé à $M^\top M$. Puisque $\forall j \in [\![1,q]\!], M^\top M v_j = \lambda_j v_j$, la matrice de ϕ dans la base $(v_1,\dots,v_q)$ est $\mathrm{diag}(\lambda_1,\dots,\lambda_q)$.

Puisque deux matrices semblables ont la même trace, $\mathrm{tr}(M^\top M) = \sum_{j=1}^{q}\lambda_j$ donc $\tilde{N} = \sum_{j=1}^{q}\lambda_j$.

9) a) Pour tout $X \in \mathbb{R}^q$, puisque u est unitaire il forme une b.o.n de D donc $\pi_D(X) = (X|u)u$.

Donc pour tout $i \in [\![1,p]\!]$, $\|\pi_D(I_i)\|^2 = (I_i|u)^2$ donc $N(D) = \sum_{i=1}^{p}(I_i|u)^2$.

Or $(I_i|u)$ est le i–ième coefficient du vecteur colonne Mu.

Donc $u^\top M^\top M u = (Mu)^\top(Mu) = \|Mu\|^2 = \sum_{i=1}^{p}(I_i|u)^2 = N(D)$.

b) Soit $j \in [\![1,q]\!], N(\mathrm{vect}(v_j))) = v_j^\top M^\top M v_j = v_j^\top \lambda_j v_j = \lambda_j v_j^\top v_j$ (car λ_j est un scalaire) donc $N(\mathrm{vect}(v_j))) = \lambda_j\|v_j\|^2 = \lambda_j$.

c) Soit $u \in \mathbb{R}^q$ un vecteur unitaire dirigeant D. Décomposons u dans la base orthonormée $(v_1,\dots,v_q)$: soient $t_1,\dots,t_q \in \mathbb{R}$ tels que $u = \sum_{j=1}^{q} t_j v_j$. Puisque u est unitaire il vient $\sum_{j=1}^{q} t_j^2 = 1$.

$$N(D) = u^\top M^\top M u = \left(\sum_{j=1}^{q} t_j v_j\right)^\top M^\top M\left(\sum_{j=1}^{q} t_j v_j\right) = \left(\sum_{j=1}^{q} t_j v_j^\top\right)\left(\sum_{k=1}^{q} t_k\lambda_k v_k\right).$$

Développons ce produit en remarquant que $\forall j,k \in [\![1,q]\!], v_j^\top v_k = (v_j|v_k) = \delta_{j,k}$ (le symbole de Kronecker), il vient $N(D) = \sum_{j=1}^{q} t_j^2\lambda_j$.

Puisque $\lambda_1 \geqslant \dots \geqslant \lambda_q \geqslant 0$, il vient $N(D) \leqslant \lambda_1\sum_{j=1}^{q} t_j^2 = \lambda_1 = N(\mathrm{vect}(v_1))$.

10) a) La projection π_S étant orthogonale, pour tout $i \in [\![1,p]\!]$, $I_i - \pi_S(I_i)$ et $\pi_S(I_i)$ sont orthogonaux.
Donc d'après Pythagore $\|I_i - \pi_S(I_i)\|^2 + \|\pi_S(I_i)\|^2 = \|I_i\|^2$.

En sommant sur i il vient $\sum_{i=1}^{p} \|I_i - \pi_S(I_i)\|^2 + N(S) = \tilde{N}$ ou encore $\sum_{i=1}^{p} \|I_i - \pi_S(I_i)\|^2 = \tilde{N} - N(S)$. Ainsi minimiser $\sum_{i=1}^{p} \|I_i - \pi_S(I_i)\|^2$ revient à maximiser $N(S)$.

b) La formule de projection sur un s.e.v dont on connaît une base orthonormée nous donne :

$$\forall i \in [\![1, p]\!], \pi_S(I_i)(I_i|a)a + (I_i|b)b$$

Or a et b sont orthogonaux donc d'après Pythagore :

$$\|\pi_S(I_i)\|^2 = (I_i|a)^2 + (I_i|b)^2 = \|\pi_{vect(a)}(I_i)\|^2 + \|\pi_{vect(b)}(I_i)\|^2$$

Il ne reste plus qu'à sommer sur i.

c) Avec le résultat précédent et celui de la question 9b on obtient $N(\text{vect}(v_1, v_2)) = \lambda_1 + \lambda_2$.

Soit maintenant S un plan quelconque, il s'agit de démontrer que $N(S) \leqslant \lambda_1 + \lambda_2$.

Soit $H = \text{vect}(v_2, \ldots, v_q)$ l'hyperplan orthogonal à v_1 : puisque $\dim(H) = q - 1$ et $\dim(S) = 2$ il vient $\dim(H \cap S) \geqslant 1$ (sinon d'après Grassmann on aurait $\dim(H + S) > q$, absurde).

On peut donc trouver un vecteur $a \in H \cap S$, non nul. Quitte à normaliser a on peut même supposer a unitaire. Soit b qui complète a en une base orthonormée de S.

D'après la question 9c $N(\text{vect}(b)) \leqslant \lambda_1$.

Maintenant si on reprend l'argument de la question 9c avec $\text{vect}(a)$, puisque la composante de a suivant v_1 est nulle, on trouve $N(\text{vect}(a)) \leqslant \lambda_2$.

Ainsi $N(S) = N(\text{vect}(a)) + N(\text{vect}(b)) \leqslant \lambda_1 + \lambda_2$ ce qui achève notre preuve.

Problème 2 : Les polynômes de Tchebychev de seconde espèce

Dans ce problème, n est un entier naturel et E désigne l'espace vectoriel réel $\mathbb{R}_n[X]$.

Pour $(P, Q) \in E^2$, on pose : $(P|Q) = \displaystyle\int_{-1}^{+1} P(t)Q(t)\sqrt{1 - t^2}\,\mathrm{d}t$.

1) Démontrer que $(.|.)$ munit E d'une structure d'espace vectoriel euclidien.

2) Pour $p \in \mathbb{N}$, on pose : $I_p = \displaystyle\int_{-1}^{+1} t^p \sqrt{1 - t^2}\,\mathrm{d}t$.

   a) Calculer I_p pour p impair.

   b) Calculer I_0 et I_2. On admet que : $I_4 = \dfrac{\pi}{16}$.

   c) Pour $n = 2$, en appliquant la méthode d'orthonormalisation de Schmidt à la base canonique $(1, X, X^2)$, déterminer une base orthonormée de $\mathbb{R}_2[X]$.

3) On considère l'application $T : \left| \begin{array}{ccc} E & \longrightarrow & E \\ P & \longmapsto & T(P) = (1 - X^2)P'' - 3XP' \end{array} \right.$.

a) Démontrer que T est bien définie et est un endomorphisme de E.

b) Pour $x \in [-1,+1]$, on pose : $F(x) = (1-x^2)^{\frac{3}{2}} P'(x)$.
Vérifier que : $\forall x \in [-1,+1]$, $F'(x) = \sqrt{1-x^2}\, T(P)(x)$.

c) En déduire, en intégrant par parties, que : $\forall (P,Q) \in E^2$, $(T(P)|Q) = (P|T(Q))$.

4) a) Déterminer la matrice M de l'endomorphisme T dans la base canonique de E.

b) Pour $\lambda \in \mathbb{R}$, calculer le déterminant de la matrice $M - \lambda I_{n+1}$.
Préciser les valeurs de λ pour lesquelles cette matrice n'est pas inversible.

c) Soit $k \in [\![0,n]\!]$, on pose $\lambda_k = -k(k+2)$.
Déterminer avec précision le rang de de la matrice $M-\lambda_k I_{n+1}$ et en déduire la dimension du noyau de $T - \lambda_k \mathrm{id}_E$.

d) Pour $k \in [\![0,n]\!]$, on admet qu'il existe un unique polynôme U_k de E tel que $T(U_k) = \lambda_k U_k$ et $U_k(1) = k+1$.
En examinant son coefficient dominant, déterminer le degré du polynôme U_k.

e) En utilisant la question (3,c), démontrer que $(U_0, U_1, \ldots, U_n)$ est une base orthogonale de E.

5) Soit $k \in [\![0,n]\!]$. On considère l'équation différentielle (E_k) : $y'' + (k+1)^2 y = 0$.

a) Résoudre (E_k).

b) On pose : $\forall \theta \in \mathbb{R}$, $f(\theta) = U_k(\cos\theta)\sin\theta$.
Démontrer que f est solution sur $\mathbb{R}$ de l'équation (E_k).

c) En déduire qu'il existe $a \in \mathbb{R}$ tel que : $\forall \theta \in \mathbb{R}$, $f(\theta) = a\sin(k+1)\theta$.

d) Démontrer que : $\forall \theta \in]0,\pi[$, $U_k(\cos\theta) = \dfrac{\sin(k+1)\theta}{\sin\theta}$.

e) Déterminer U_0, U_1 et U_2. Vérifier le résultat de la question (2,c).

Les polynômes U_k $(k \in \mathbb{N})$ sont les polynômes de Tchebychev de seconde espèce.

Problème 3 : estimation statistique d'espèces non recensées

1) $S_0 + \cdots + S_n$ est le nombre total des espèces donc $S_0 + \cdots + S_n = N$, plus précisément c'est la variable aléatoire constante à N.

2) a) Notons $\mathcal{Q}_k$ l'ensemble des parties de $[\![1,n]\!]$ de cardinal k, ainsi card$(\mathcal{Q}_k = \binom{n}{k})$.
L'événement $[X_j = 1]$ est l'union 2 à 2 disjointe des événements :

$$[\forall u \in A, i_u \in E_j \text{ et } \forall v \in [\![1,n]\!] \setminus A, i_v \notin E_j]$$

$$\begin{aligned}
P([X_j=1] &= \sum_{A\in\mathcal{Q}_k} P\left(\left(\bigcap_{u\in A}[i_u\in E_j]\right)\cap\left(\bigcap_{v\in[\![1,n]\!]\smallsetminus A}\overline{[i_v\in E_j]}\right)\right)\\
&= \sum_{A\in\mathcal{Q}_k}\left(\prod_{u\in A}P\left([i_u\in E_j]\right)\right)\left(\prod_{v\in[\![1,n]\!]\smallsetminus A}P(\overline{[i_v\in E_j]})\right) \text{ par indépendance}\\
&= \sum_{A\in\mathcal{Q}_k} p_j^k(1-p_j)^{n-k}\\
&= \binom{n}{k}p_j^k(1-p_j)^{n-k}
\end{aligned}$$

b) S_k est le nombre d'espèces qui ont été observées exactement k fois, et $X_j=1$ si et seulement si l'espère E_j a été observée exactement j fois, donc $\sum\limits_{j=1}^{N}X_j$ est bien le nombre d'espèces ayant été observées exactement k fois.

c) Pour tout $j\in[\![1,n]\!], E(X_j)=\binom{n}{k}p_j^k(1-p_j)^{n-k}$ (espérance d'une variable aléatoire de Bernoulli) et il n'y a plus qu'à utiliser la linéarité de l'espérance.

d) C'est une conséquence immédiate du résultat précédent :

$$\begin{aligned}
E(S_0) &= \sum_{j=1}^{N}(1-p_j)^n\\
E(S_1) &= n\sum_{j=1}^{N}p_j(1-p_j)^{n-1}\\
E(S_2) &= \frac{n(n-1)}{2}\sum_{j=1}^{N}p_j^2(1-p_j)^{n-2}
\end{aligned}$$

Sans connaissance sur les p_j il n'y a pas de moyen de calculer ces sommes.

3)

$$\begin{aligned}
\sum_{k=1}^{N}(\alpha b_k-\beta a_k)^2 &= \alpha^2\sum_{k=1}^{N}b_k^2-2\alpha\beta\sum_{k=1}^{N}a_kb_k+\beta^2\sum_{k=1}^{N}a_k^2\\
&= 2\alpha^2\beta^2-2\alpha\beta\sum_{k=1}^{N}a_kb_k
\end{aligned}$$

Or $\sum\limits_{k=1}^{N}(\alpha b_k-\beta a_k)^2\geqslant 0$ (en tant que somme de carrés qui sont positifs)

donc $2\alpha\beta\sum\limits_{k=1}^{N}a_kb_k\leqslant 2\alpha^2\beta^2$. Par ailleurs il est évident que $\alpha\geqslant 0$ et $\beta\geqslant 0$.

1^e cas : $\alpha>0$ et $\beta>0$, alors $\sum\limits_{k=1}^{N}a_kb_k\leqslant\alpha\beta$.

2^e cas : $\alpha = 0$ et $\beta = 0$, par exemple $\alpha = 0$, alors $\sum_{k=1}^{N} a_k^2 = \alpha^2 = 0$, c'est une somme nulle de termes tous positifs donc chaque terme est nul, l'inégalité $\sum_{k=1}^{N} a_k b_k \leqslant \alpha\beta$ est encore vraie car $0 \leqslant 0$. Le cas $\beta = 0$ se traite de façon similaire.

4) Pour tout $k \in [\![1, N]\!]$ posons $a_k = (1 - p_k)^{n/2}$ et $b_k = p_k(1 - p_k)^{(n-2)/2}$. Alors $a_k b_k = p_k(1 - p_k)^{n-1}$ et l'inégalité de Cauchy–Schwarz donne :

$$\sum_{k=1}^{N} p_k(1 - p_k)^{n-1} \leqslant \sqrt{\left[\sum_{k=1}^{N}(1 - p_k)^n\right] \cdot \left[\sum_{k=1}^{N} p_k^2(1 - p_k)^{n-2}\right]}$$

Il n'y a plus qu'à mettre au carré cette inégalité ce qui est possible car chaque membre est positif.

5) L'inégalité de la question 4 combinée avec les résultats de la question 4d) donne :

$$\left(\frac{E(S_1)}{n}\right)^2 \leqslant E(S_0).\frac{2E(S_2)}{n(n-1)}$$

La formule de la question 2d) prouve que $E(S_2) > 0$, on en déduit donc :

$$E(S_0) \geqslant \frac{n-1}{2n}\frac{E(S_1)^2}{E(S_2)}$$

6) La modélisation selon laquelle les événements $[i_1 \in E_{\ell_1}], \dots, [i_n \in E_{\ell_n}]$ sont mutuellement indépendants n'est plus raisonnable si on effectue le recensement de toute la planète : en fait cette modélisation correspond à un sondage avec remise (à chaque fois qu'on examine un individu on le remet en jeu pour des tirages successifs pour ne pas changer la proportion de chaque espèce) alors que si on examine une portion importante de la population il faut considérer qu'on fait des tirages sans remise (à chaque fois qu'on tire un individu d'une espèce, cela fait baisser la proportion de cette espèce dans la population restante).