Trois problèmes indépendants. Le problème 1 est une introduction à certaines méthodes de représentation de données. Le problème 2 est un sujet très "classique" sur une certaine famille de polynômes orhtogonaux (à savoir traiter). Le problème 3 présente une méthode statistique.

Problème 1 : analyse en composantes principales

La représentation intelligente de données statistiques fait appel à des structures euclidiennes. Le devoir suivant est une introduction à la technique de l'analyse en composantes principales, qui est une technique très répandue de la représentation de données statistiques multidimensionnelles.
Nous allons étudier une situation où p individus concourent à des épreuves sportives dans $q \geqslant 2$ disciplines : chaque individu participe à toutes les disciplines, et pour tout $(i,j) \in [\![1,p]\!] \times [\![1,q]\!]$ on note $x_{i,j}$ le score obtenu par l'individu i dans la discipline j (voir en fin de sujet une étude concrète).
Soit $M = (x_{i,j})_{\substack{1 \leqslant i \leqslant p \\ 1 \leqslant j \leqslant q}} \in \mathcal{M}_{p,q}(\mathbb{R})$ la matrice de ces scores.

Nous supposerons que les scores de chaque discipline sont centrés : $\forall j \in [\![1,q]\!], \sum_{i=1}^{q} x_{i,j} = 0$.

De plus nous supposerons que $\mathrm{rg}(M) = q$.
Nous noterons $(.|.)$ le produit scalaire canonique de $\mathbb{R}^q$ et $\langle .|. \rangle$ celui de $\mathbb{R}^p$. La notation $\|.\|$ désignera la norme euclidienne dans $\mathbb{R}^q$. Pour tout $n \in \mathbb{N}^*$ on note $\vec{0}_n$ le vecteur nul de $\mathbb{R}^n$. Pour simplifier les notations du calcul matriciel, tout vecteur de $\mathbb{R}^n$ sera identifié à une matrice colonne.

Nous admettrons le résultat suivant (théorème spectral) :
> Soit $n \in \mathbb{N}^*$, soit $S \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique, alors il existe une base orthonormée $(v_1, \dots, v_n)$ de $\mathbb{R}^n$ et des réels $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ tels que $\forall i \in [\![1,n]\!], Sv_i = \lambda_i v_i$

1) Démontrer que $M^\top M$ est une matice symétrique. Préciser le nombre de ses lignes et le nombre de ses colonnes.

Soient alors une base orthonormée $(v_1, \dots, v_q)$ de $\mathbb{R}^q$ et des réels $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_q$ tels que $\forall j \in [\![1,q]\!], M^\top M v_j = \lambda_j v_j$ (leur existence provenant du théorème spectral).
Pour tout $j \in [\![1,q]\!]$ soit $w_j = Mv_j$.

2) Soient $i,j \in [\![1,q]\!]$, montrer que $\langle w_i | w_j \rangle = \begin{cases} \lambda_i & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$.

3) Soit $X \in \mathbb{R}^q \setminus \left\{\vec{0}_q\right\}$, montrer que $MX \neq \vec{0}_p$ (indication : utiliser $rg(M) = q$).

4) En déduire que $\forall j \in [\![1,q]\!], \lambda_j > 0$.

5) a) Montrer que $\forall j \in [\![1,q]\!], MM^\top w_j = \lambda_j w_j$.
   b) Montrer que $(w_1, \dots, w_q)$ est libre. En déduire que $\mathrm{rg}(MM^\top) = q$.

6) Pour tout $i \in [\![1,p]\!]$ on pose $\mu_i = \begin{cases} \lambda_i & \text{si } i \leqslant q \\ 0 & \text{si } i > q \end{cases}$.

   Soit $F = \mathrm{Vect}(w_1, \dots, w_q)$.

   a) Soit $X \in \ker(MM^\top)$, montrer que $X \in F^\perp$ (indication : on pourra par exemple considérer $X^\top MM^\top w_i$).
   Montrer que $\dim(\ker(MM^\top)) \geqslant p - q$.
   En déduire que $\ker(MM^\top) = F^\perp$.

b) En déduire qu'il est possible de construire une base orthonormée $(w'_1, \dots, w'_p)$ de $\mathbb{R}^p$ et telle que :

$$\forall i \in [\![1,p]\!], MM^\top w'_i = \mu_i w'_i$$

(On retrouve donc le résultat du théorème spectral pour la matrice $MM^\top$).

7) Pour tout $X \in \mathbb{R}^q$ justifier que $\|X\|^2 = \sum_{j=1}^{q} (X|v_j)^2$.

8) Pour tout $i \in [\![1,p]\!]$ soit $I_i = (x_{i,j})_{1\leqslant j\leqslant q} \in \mathbb{R}^q$ le vecteur des scores de l'individu i.

On définit alors $\tilde{N} = \sum_{i=1}^{p} \|I_i\|^2$ ($\tilde{N}$ s'appelle l'inertie totale).

Montrer que $\tilde{N} = \mathrm{tr}(M^\top M)$.

En déduire que $\tilde{N} = \sum_{j=1}^{q} \lambda_j$ (on pourra par exemple écrire dans la base $(v_1, \dots, v_q)$ la matrice de l'endomorphisme canoniquement associé à $M^\top M$).

Pour tout sous–espace vectoriel F de $\mathbb{R}^q$ soient $\pi_F \in L(\mathbb{R}^q)$ le projecteur orthogonal sur F et $N(F) = \sum_{i=1}^{p} \|\pi_F(I_i)\|^2$.

9) Inertie projetée sur un axe

a) Soit $u \in \mathbb{R}^q$ un vecteur unitaire, soit $D = \mathrm{Vect}(u)$. Montrer que $N(D) = u^\top M^\top M u$.

b) Soit $j \in [\![1,q]\!]$. Montrer que $N(\mathrm{Vect}(v_j)) = \lambda_j$.

c) Démontrer que pour toute droite vectorielle D de $\mathbb{R}^q$, $N(D) \leqslant \lambda_1 = N(\mathrm{Vect}(v_1))$.

Pour obtenir la meilleure représentation plane du nuage de points $(I_i)_{1\leqslant i\leqslant p}$ on cherche le plan vectoriel S qui minimise $\sum_{i=1}^{p} \|I_i - \pi_S(I_i)\|^2$ (principe de meilleure approximation aux moindres carrés). On visualise alors le nuage de points $(I_i)_{1\leqslant i\leqslant p}$ en représentant sur S la famille $(\pi_S(I_i))_{1\leqslant i\leqslant p}$.

10) a) Montrer que chercher ce meilleur plan revient à chercher le plan S qui maximise $N(S)$. (principe de maximisation de l'inertie projetée).

b) Soient S un plan de $\mathbb{R}^q$, et (a,b) une base orthonormée de S.
Montrer que $N(S) = N(\mathrm{Vect}(a)) + N(\mathrm{Vect}(b))$.

c) Calculer $N(\mathrm{Vect}(v_1, v_2))$, et montrer que $\mathrm{Vect}(v_1, v_2)$ est le plan vectoriel permettant d'obtenir la meilleure représentation du nuage de points $(I_i)_{1\leqslant i\leqslant p}$.

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

En mission d'exploration galactique dans votre soucoupe volante, vous découvrez une planète abritant la vie. On notera N le nombre d'espèces différentes sur cette planètes, notées $E_1, \ldots, E_N$. Pour tout $k \in [\![1, N]\!]$ on note $p_k \in]0, 1[$ la proportion d'individus[1] de l'espèce E_k parmi les individus de cette planète. Hélas vous ne connaissez ni la valeur de N ni celle des p_k (à part le fait bien sûr que $p_1 + \cdots + p_N = 1$). Une rapide exploration vous permet d'étudier n individus (n étant très inférieur au nombre total d'individus habitant la planète) notés $i_1, \ldots, i_n$. On supposera que pour tous $k \in [\![1, n]\!]$ et $\ell \in [\![1, N]\!]$, $P([i_k \in E_\ell]) =$

[1] animaux, végétaux, autre…

p_ℓ, par ailleurs on supposera que pour tous $\ell_1, \dots, \ell_n \in [\![1, N]\!]$, les événements $[i_1 \in E_{\ell_1}], \dots, [i_n \in E_{\ell_n}]$ sont mutuellement indépendants (la connaissance des espèces des autres individus n'apporte aucun renseignement sur la connaissance de l'espèce d'un individu particulier). On supposera $n \geqslant 2$.
Pour tout $k \in [\![0, n]\!]$ notons S_k la variable aléatoire qui compte le nombre d'espèces dont vous avez recensé exactement k représentants (ainsi S_0 est le nombre d'espèces présentes sur la planète mais dont vous n'avez trouvé aucun représentant, c'est cette quantité que nous allons essayer d'estimer).

1) Quelle est la valeur de $S_0 + \cdots + S_n$? (Le résultat est très simple !)

2) Fixons $k \in [\![0, n]\!]$, pour tout $j \in [\![1, N]\!]$ soit la variable aléatoire X_j qui vaut 1 si l'espèce E_j a été observée exactement k fois, 0 sinon.

   a) Montrer que X_j suit une loi de Bernoulli de paramètre $\binom{n}{k} p_j^k (1 - p_j)^{n-k}$.

   b) Justifier que $\sum_{j=1}^{N} X_j = S_k$.

   c) En déduire que $E(S_k) = \binom{n}{k} \sum_{j=1}^{N} p_j^k (1 - p_j)^{n-k}$.

   d) Expliciter en particulier $E(S_0)$, $E(S_1)$ et $E(S_2)$, à chaque fois sous forme de sommes. Peut–on calculer ces sommes ?

3) L'inégalité de Cauchy–Schwarz

   Soient $a_1, \dots, a_N, b_1, \dots, b_N$ des réels positifs, posons $\alpha = \sqrt{\sum_{k=1}^{N} a_k^2}$ et $\beta = \sqrt{\sum_{k=1}^{N} b_k^2}$.

   Montrer que $\sum_{k=1}^{N} (\alpha b_k - \beta a_k)^2 = 2\alpha^2\beta^2 - 2\alpha\beta \sum_{k=1}^{N} a_k b_k$.

   En déduire que $\sum_{k=1}^{N} a_k b_k \leqslant \alpha\beta$.

4) Montrer que $\left[\sum_{k=1}^{N} p_k (1 - p_k)^{n-1}\right]^2 \leqslant \left[\sum_{k=1}^{N} (1 - p_k)^n\right] . \left[\sum_{k=1}^{N} p_k^2 (1 - p_k)^{n-2}\right]$.
   On pourra par exemple se ramener à l'inégalite de Cauchy–Schwarz, à condition de bien préciser les valeurs a_k et b_k qu'on utilise.

5) En déduire une inégalité entre $E(S_0)$, $E(S_1)$ et $E(S_2)$. En particulier, on explicitera une constante C ne dépendant que de n telle que :

$$E(S_0) \geqslant C \frac{E(S_1)^2}{E(S_2)}$$

Note historique : ce résultat date de 1984 et est dû à la statisticienne taïwannaise Anna Chao. Il a servi notamment à estimer le nombre d'espèces d'arbres non recensés dans une forêt de Guyane suite à l'étude d'une parcelle. En pratique on se contente de remplacer $E(S_1)$ et $E(S_2)$ par leurs valeurs respectivement observées.

6) Dans le cas où on a pu étudier tous les individus de la planète, nécessairement $S_0 = 0$. Ceci peut–il contredire le résultat précédent ? Expliquer ce paradoxe.