HEC 2003. Math1 option scientifique.

NUAGES DE POINTS ET APPROXIMATION D'UN NUAGE

Dans tout le problème n et p désignent des entiers naturels supérieurs ou égaux à 2 et on pose $E_p = \mathcal{M}_{p,1}(\mathbb{R})$.
L'espace E_p est muni de sa structure euclidienne canonique ; la norme euclidienne d'un vecteur x de E_p est notée $\|x\|$; le produit scalaire de deux vecteurs x et y de E_p est noté $\langle x, y\rangle$.
Si u est un vecteur non nul appartenant à E_p, D_u désigne la droite vectorielle engendrée par u et si x est un vecteur de E_p, $P_{D_u}(x)$ est le projeté orthogonal de x sur la droite D_u.
Si F est un sous-espace vectoriel de E_p, le supplémentaire orthogonal de F dans E_p est noté $F^\perp$.
Pour toute matrice A appartenant à $\mathcal{M}_{m,\ell}(\mathbb{R})$ on note Φ_A l'application linéaire de $\mathcal{M}_{\ell,1}(\mathbb{R})$ dans $\mathcal{M}_{m,1}(\mathbb{R})$ définie par : $\forall X \in \mathcal{M}_{\ell,1}(\mathbb{R})$, $\Phi_A(X) = AX$.
Pour tout r appartenant à $\mathbb{N}^*$ et toute famille $(u_i)_{1\leqslant i\leqslant r}$ de vecteurs de E_p, $\mathrm{Vect}(u_1, \ldots, u_r)$ est le sous-espace vectoriel de E_p engendré par les vecteurs $u_1, \ldots, u_r$.
Si g est une fonction définie sur un sous-espace vectoriel F de E_p et à valeurs dans $\mathbb{R}$, on désigne par $\max\limits_{\substack{x\in F\\ \|x\|=1}} g(x)$ ou $\max\{g(x); x \in F \quad \text{et} \quad \|x\| = 1\}$ le maximum, lorsqu'il existe, de la fonction g sur l'ensemble des vecteurs x de F dont la norme est égale à 1.

Partie I: Étude d'un exemple

Dans cette partie et uniquement dans celle-ci, on suppose que $p = 2$. On note (u_1, u_2) la base canonique de E_2.

1) On considère les vecteurs v_1, v_2 et v_3 appartenant à E_2 et dont les coordonnées dans la base (u_1, u_2) sont respectivement $(1, 2), (-3, -1), (2, -1)$.
On considère un réel m et on note, pour tout i appartenant à $\{1, 2, 3\}$, v_i' le projeté orthogonal de v_i sur la droite vectorielle engendrée par $u_1 + mu_2$.

a) Calculer en fonction de m la quantité : $\|v_1'\|^2 + \|v_2'\|^2 + \|v_3'\|^2$.

b) Déterminer la valeur m_0 de m pour laquelle cette quantité atteint son maximum ; ce maximum est noté λ_1.

2) Soit X la matrice $\begin{pmatrix} 1 & -3 & 2 \\ 2 & -1 & -1 \end{pmatrix}$.

a) Vérifier que λ_1 est une valeur propre de $\Phi_{X\,{}^tX}$; $u_1 + m_0u_2$ étant un vecteur propre associé à λ_1.

b) Déterminer l'autre valeur propre de $\Phi_{X\,{}^tX}$ et la comparer à λ_1.

Partie II: Les axes principaux d'inertie d'un nuage

Les notations introduites dans cette partie seront utilisées dans toute la suite du problème.
On définit la matrice $X = (x_{ij})_{\substack{1\leqslant i\leqslant p\\ 1\leqslant j\leqslant n}}$ appartenant à $\mathcal{M}_{p,n}(\mathbb{R})$ appelée nuage ; ses colonnes $c_1, \ldots, c_n$ sont appelées points du nuage ; X est donc un nuage de n points dans un espace de dimension p.
On définit la matrice $V = X\,{}^tX$.
On appelle F le sous-espace vectoriel de E_p engendré par les vecteurs colonnes $c_1, \ldots, c_n$ et on suppose que $\dim F = r$ et $p > r \geqslant 1$.
Pour tout vecteur v non nul de E_p, on pose $I(v) = \sum\limits_{j=1}^{n} \|P_{D_v}(c_j)\|^2$; cette quantité s'appelle l'inertie du nuage X sur la droite D_v.
Pour tout couple de vecteurs (v, w) appartenant à E_p^2 , on pose: $J(v, w) = \sum\limits_{j=1}^{n} \langle v, c_j\rangle\langle w, c_j\rangle$.

1)a) Montrer que la matrice V est diagonalisable et que ses valeurs propres sont des réels positifs ou nuls.

On note $\lambda_1, \ldots, \lambda_p$ les valeurs propres de V et on suppose que $\lambda_1 \geqslant \ldots \geqslant \lambda_p$
Justifier l'existence d'une base orthonormale $(e_1, \ldots, e_p)$ de E_p telle que:

$$\forall i \in [\![1, p]\!],\ Ve_i = \lambda_i e_i$$

b) • Montrer que le noyau de Φ_V est égal à celui de $\Phi_{{}^tX}$.
• En déduire que le rang de V est égal à r.
• Montrer que: $\lambda_{r+1} = \ldots = \lambda_p = 0$.
• Que peut-on dire de $\lambda_1, \ldots, \lambda_r$?
• Montrer que $(e_1, \ldots, e_r)$ est une base de F.

2)a) Montrer, pour tout vecteur v de norme 1 appartenant à E_p, l'égalité: $I(v) = {}^tvVv$.

b) Déterminer, pour tout i appartenant à $[\![1, p]\!]$, $I(e_i)$ à l'aide des nombres $\lambda_1, \ldots, \lambda_p$

c) On définit les sous-espaces vectoriels $F_1, \ldots, F_r$ de E_p par :

$$F_1 = F, \quad F_2 = F_1 \cap (D_{e_1}^{\perp}), \ldots, F_r = F_{r-1} \cap (D_{e_{r-1}}^{\perp})$$

• Montrer que : $\forall i \in [\![1, r]\!],\ F_i = \mathrm{Vect}(e_i, \ldots, e_r)$.
• Montrer que : $I(e_1) = \max\{I(v); v \in E_p \text{ et } \|v\| = 1\} = \max\{I(v); v \in F_1 \quad \text{et} \quad \|v\| = 1\}$.
• Montrer que : $\forall i \in [\![1, r]\!], I(e_i) = \max\{I(v); v \in F_i \quad \text{et} \quad \|v\| = 1\}$.

3) Soit w un vecteur unitaire de E_p tel que $I(w) = \max\{I(v); v \in E_p \quad \text{et} \quad \|v\| = 1\}$. Montrer que w appartient à F.

4) On suppose dans cette question que $\varepsilon_1, \ldots, \varepsilon_r$ sont r vecteurs de norme 1 appartenant à E_p et que $G_1, \ldots, G_r$ sont r sous-espaces vectoriels de E_p tels que

$$(\mathcal{S}) \begin{cases} G_1 = F \\ \varepsilon_1 \in G_1 \quad \text{et} \quad I(\varepsilon_1) = \max\{I(v); v \in G_1 \quad \text{et} \quad \|v\| = 1\} \\ \varepsilon_2 \in G_2 = G_1 \cap (D_{\varepsilon_1}^{\perp}), \quad \text{et} \quad I(\varepsilon_2) = \max\{I(v); v \in G_2 \quad \text{et} \quad \|v\| = 1\} \\ \vdots \\ \varepsilon_{r-1} \in G_{r-1} = G_{r-2} \cap (D_{\varepsilon_{r-2}}^{\perp}), \quad \text{et} \quad I(\varepsilon_{r-1}) = \max\{I(v); v \in G_{r-1} \quad \text{et} \quad \|v\| = 1\} \\ \varepsilon_r \in G_r = G_{r-1} \cap (D_{\varepsilon_{r-1}}^{\perp}), \quad \text{et} \quad I(\varepsilon_r) = \max\{I(v); v \in G_r \quad \text{et} \quad \|v\| = 1\} \end{cases}$$

Les droites vectorielles $D_{\varepsilon_1}, \ldots, D_{\varepsilon_r}$ sont appelées axes principaux d'inertie du nuage.

a) Vérifier que $(\varepsilon_1, \ldots, \varepsilon_r)$ est une base orthonormale de F et que $(\varepsilon_1, \ldots, \varepsilon_r, e_{r+1}, \ldots, e_p)$ est une base orthonormale de E_p.

b) Montrer que pour tout couple de vecteurs (v, w) appartenant à E_p :

$$J(v, w) = {}^tvVw = \langle v, \Phi_V(w)\rangle$$

c) On se donne deux vecteurs v_1 et v_2, unitaires, orthogonaux et appartenant à F.
Pour tout réel t, on pose $\varphi(t) = I(\cos t\ v_1 + \sin t\ v_2)$.
• Exprimer $\varphi(t)$ à l'aide de $I(v_1), I(v_2), J(v_1, v_2)$ et t.
• Montrer que φ est majorée sur $\mathbb{R}$ et qu'elle admet un maximum.
• On suppose que le maximum de φ est atteint en 0. Montrer que $J(v_1, v_2) = 0$.

d) • Montrer que pour tout (i, j) appartenant à $[\![1, r]\!]^2$, $J(\varepsilon_i, \varepsilon_j) = 0$ dès que $i \neq j$.
• Déterminer la forme de la matrice de Φ_V dans la base $(\varepsilon_1, \ldots, \varepsilon_r, e_{r+1}, \ldots, e_p)$.
• En déduire que pour tout $i \in [\![1, r]\!]$, ε_i est un vecteur propre de V associé à λ_i.

5) *Dans le langage des statisticiens les colonnes c_j de X représentent des individus d'une population statistique où p variables statistiques x_i, $(1 \leqslant i \leqslant p)$ ont respectivement pris les valeurs $x_{i1}, x_{i2}, \ldots, x_{in}$ $(1 \leqslant i \leqslant p)$, valeurs fixées de telle sorte que leur moyennes sont nulles, c'est à dire :* $\displaystyle\sum_{j=1}^{n} x_{ij} = 0,\ 1 \leqslant i \leqslant p.$

Calculer la covariance $\mathrm{Cov}(x_k, x_\ell)$ des variables x_k et x_ℓ lorsque k et ℓ appartiennent à $[\![1, p]\!]$ puis comparer la matrice V et la matrice $\big(\mathrm{Cov}(x_k, x_\ell)\big)_{\substack{1 \leqslant k \leqslant p \\ 1 \leqslant \ell \leqslant p}}$

Partie III: Une décomposition de la matrice X

Pour tout $i \in [\![1,p]\!]$ on note Π_i la matrice dans la base canonique de E_p, de la projection orthogonale de E_p sur D_{e_i} ; les vecteurs $e_1, \ldots, e_p$ ont été définis au II.1.a.

1) Montrer que : $\displaystyle\sum_{i=1}^{p} \Pi_i = I_p$, (où I_p est la matrice appartenant à $\mathcal{M}_p(\mathbb{R})$ dont tous les éléments sont nuls excepté les éléments diagonaux qui valent 1).

2) Déterminer $\Pi_i \Pi_j$ pour tout $(i,j) \in [\![1,p]\!]^2$ tel que $i \neq j$.

3) Calculer pour tout $i \in [\![r+1,p]\!]$, $\Pi_i X$ et en déduire que : $X = \displaystyle\sum_{i=1}^{r} \Pi_i X$.

4) Pour tout $s \in [\![1,r]\!]$, on pose $X_s = \displaystyle\sum_{i=1}^{s} \Pi_i X$.

a) Montrer que : $\operatorname{Im} \Phi_{X_s} \subset \operatorname{Vect}(e_1, \ldots, e_s)$.

b) Calculer $X_s\, {}^t\!X e_j$ pour tout $j \in [\![1,p]\!]$ et déterminer le rang de X_s.

Partie IV: Une norme euclidienne de matrices carrées

Pour tout entier naturel q non nul et toute matrice, carrée $A = (a_{ij})_{\substack{1 \leqslant i \leqslant q \\ 1 \leqslant j \leqslant q}}$ appartenant à $\mathcal{M}_q(\mathbb{R})$, on pose $\operatorname{tr}(A) = \displaystyle\sum_{i=1}^{q} a_{ii}$.

On sait que tr *définit une application linéaire de $\mathcal{M}_q(\mathbb{R})$ dans $\mathbb{R}$ et que si A et B appartiennent respectivement à $\mathcal{M}_{n,p}(\mathbb{R})$ et $\mathcal{M}_{p,n}(\mathbb{R})$ alors* $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. *On sait également que si deux matrices A et B sont semblables alors* $\operatorname{tr}(A) = \operatorname{tr}(B)$.

Pour tout M et N appartenant à $\mathcal{M}_{p,n}(\mathbb{R})$ on pose : $\Theta(M,N) = \operatorname{tr}(M\, {}^t\!N)$.

1) Montrer que $(M,N) \mapsto \Theta(M,N)$ est un produit scalaire sur $\mathcal{M}_{p,n}(\mathbb{R})$.
Pour toute matrice M appartenant à $\mathcal{M}_{p,n}(\mathbb{R})$, on note $|||M||| = \sqrt{\operatorname{tr}(M\, {}^t\!M)}$, *appelé ici norme euclidienne de M.*

2) Calculer pour tout $(i,j) \in [\![1,p]\!]^2$, $\Theta(\Pi_i X, \Pi_j X)$. On distinguera les cas $i = j$ et $i \neq j$, et on exprimera les résultats en fonction des nombres $\lambda_1, \ldots, \lambda_p$.

3) Calculer $|||X - X_s|||^2$ en fonction de $\lambda_1, \ldots, \lambda_r$, pour tout s appartenant à $[\![1,r]\!]$.

Partie V: La meilleure approximation du nuage

On rappelle que si H_1 et H_2 sont deux sous-espaces vectoriels de E_p, alors :

$$\dim(H_1 + H_2) = \dim H_1 + \dim H_2 - \dim(H_1 \cap H_2)$$

On considère un entier naturel s appartenant à $[\![1, r-1]\!]$ et une matrice N appartenant à $\mathcal{M}_{p,n}(\mathbb{R})$ telle que $\operatorname{rg}(N) \leqslant s$.

1) Justifier rapidement l'existence d'une base orthonormale $(a_1, \ldots, a_p)$ de E_p formée de vecteurs propres de $(X-N)\, {}^t\!(X-N)$. On note $\gamma_1, \ldots, \gamma_p$ les valeurs propres de $(X-N)\, {}^t\!(X-N)$ associées respectivement aux vecteurs $a_1, \ldots, a_p$ et on suppose que $\gamma_1 \geqslant \ldots \geqslant \gamma_p$.

2) Soit i un entier appartenant à $[\![1, r-s]\!]$ et G un sous-espace de E_p de dimension supérieure ou égale à i.

a) Montrer que: $\dim\big(G \cap \operatorname{Vect}(a_i, \ldots, a_p)\big) \geqslant 1$.

b) En déduire qu'il existe un vecteur unitaire u appartenant à G tel que $\| {}^t\!(X-N)u\|^2 \leqslant \gamma_i$.

c) On considère l'espace vectoriel $H = \big(\operatorname{Ker} \Phi_{{}^t\!N}\big) \cap \operatorname{Vect}(e_1, \ldots, e_{s+i})$.
- Montrer que : $\dim H \geqslant i$.
- En déduire : $\lambda_{s+i} \leqslant \gamma_i$.

3)a) Montrer que : $|||X - N|||^2 = \displaystyle\sum_{i=1}^{p} \gamma_i$.

b) En déduire que : $|||X - N|||^2 \geqslant \sum_{i=s+1}^{r} \lambda_i$.

c) En déduire que X_s réalise la meilleure approximation de X par des matrices de rang inférieur ou égal à s au sens de la norme euclidienne définie plus haut sur $\mathcal{M}_{p,n}(\mathbb{R})$.

4) Soit G un sous-espace vectoriel de E_p. On note P_G la projection orthogonale de E_p sur G, Π_G sa matrice dans la base canonique de E_p et $K(G) = \sum_{j=1}^{n} \|P_G(c_j)\|^2$.

La quantité $K(G)$ s'appelle l'inertie du nuage X sur le sous-espace G, et dans le cas où $G = E_p$, $K(G)$ est l'inertie totale *du nuage X.*

a) Montrer que : $K(G) = |||\Pi_G X|||^2$.

b) Montrer que : $K(G) = |||X|||^2 - |||X - \Pi_G X|||^2$.

c) On suppose toujours que s est un entier appartenant à $[\![1, r-1]\!]$ et $\dim G \leqslant s$.

• Montrer que : $K(G) \leqslant \sum_{i=1}^{s} \lambda_i$.

• Montrer que $K\big(\mathrm{Vect}(e_1, \ldots, e_s)\big)$ est le maximum des nombres $K(G)$, lorsque G parcourt l'ensemble des sous-espaces vectoriels de E_p dont la dimension est inférieure ou égale à s.

d) On suppose dans cette question que s appartient à $[\![1, p]\!]$, on ne suppose donc plus que $s \leqslant r - 1$.
Montrer que $K\big(\mathrm{Vect}(e_1, \ldots, e_s)\big)$ est le maximum des nombres $K(G)$, lorsque G parcourt l'ensemble des sous-espaces vectoriels de E_p dont la dimension est inférieure ou égale à s.

Partie VI: Non multa, sed multum

Dans cette partie, on propose une interprétation pratique des résultats théoriques précédents à propos d'une enquête de consommations.

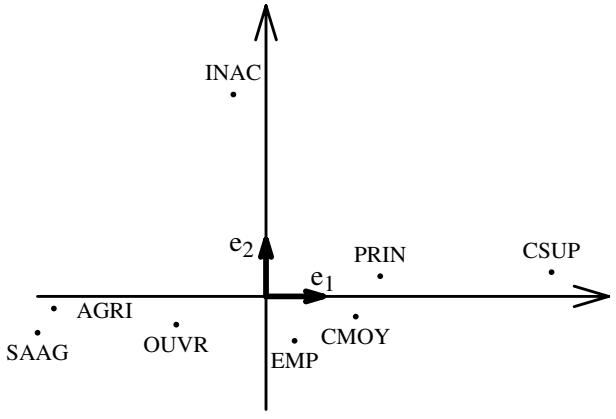
On a étudié les « consommations » annuelles de 8 denrées alimentaires (ce sont les 8 variables statistiques x_i, $(1 \leqslant i \leqslant 8)$ que l'on suppose centrées), par différentes catégories socio-professionnelles, à savoir : celles des exploitants agricoles (AGRI) représentées par la colonne c_1, des salariés agricoles (SAAG(= c_2)) , des professions indépendantes (PRIN(=c_3)), les cadres supérieurs (CSUP(= c_4)), des cadres moyens (CMOY(= c_5)), des employés (EMP(= c_6)), des ouvriers (OUV(= c_7)), des inactifs (INAC(= c_8)). Dans notre exemple un individu est donc une catégorie socio-professionnelle.

On a consigné les résultats de l'enquête dans une matrice $X = (x_{ij})_{\substack{1 \leqslant i \leqslant 8 \\ 1 \leqslant j \leqslant 8}}$. Par exemple x_{12} représente la consommation moyenne de la denrée 1 par la catégorie SAAG.

Les valeurs propres de la matrice $V = X\,{}^{\mathrm{t}}X$ sont approximativement $70, 20, 5, 3, 2, 0, 0$ et 0 associées respectivement à $e_1, \ldots, e_8$.

1) Quelle part de l'inertie totale est contenue dans l'inertie du nuage de points sur le sous-espace de base (e_1, e_2).

On a représenté dans le dessin ci-contre les projetés orthogonaux dans le plan de base (e_1, e_2) des 8 individus $(c_j)_{1 \leqslant j \leqslant 8}$, c'est-à-dire des 8 colonnes représentant les consommations moyennes de chaque catégorie socio-professionnelle.



2) Que représente le nuage de points du dessin pour le nuage X de l'enquête ?