

Mathématiques : Examen d'Analyse des données

Janvier 2008 - durée 3 heures

Calculatrice autorisée. Documents interdits.

Exercice 1

Soit une matrice réelle $X_{n \times p}$ et deux matrices réelles $Q_{p \times p}$ et $D_{n \times n}$ définissant toutes deux des métriques sur $\mathbb{R}^p$ et $\mathbb{R}^n$ respectivement.

Partie I

1. Expliquer en quoi le schéma ci-dessous correspond à un schéma de dualité.

$$\begin{array}{ccc} \mathbb{R}^p & \xrightarrow{Q} & \mathbb{R}^{p*} \\ {}^tX \uparrow & & \downarrow X \\ \mathbb{R}^{n*} & \xrightarrow{D} & \mathbb{R}^n \end{array}$$

2. Montrer que les matrices ${}^tX D X Q$ et $X Q {}^tX D$ sont M -symétriques avec M une métrique à définir.
3. Soit r le rang de X . Quel est le rang de ${}^tX D X Q$ et $X Q {}^tX D$.
4. On note $\Lambda_{r \times r}$ la matrice diagonale des r valeurs propre non nulles de ${}^tX D X Q$ et $U = [U_1, \dots, U_r]$ la matrice des vecteurs propres associés.
 - (a) Rappeler la propriété des valeurs propres et vecteurs propres d'une matrice réelle symétrique définie positive.
 - (b) Justifier que Q peut s'écrire sous la forme $Q^{1/2}Q^{1/2}$ avec $Q^{1/2}$ une matrice symétrique définie positive.
 - (c) En déduire que $Q^{1/2}U$ est la matrice des vecteurs propres d'une matrice symétrique définie positive.
 - (d) En déduire que les r valeurs propres non nulles de ${}^tX D X Q$ sont réelles et strictement positives.
 - (e) En déduire qu'il est possible de construire les r vecteurs U_i Q -orthonormés.
5. Montrer que $\text{Im } U \subset \text{Im } {}^tX$ puis en déduire que $\text{Im } U = \text{Im } {}^tX$.

On en déduit :

Proposition 0.1 *La matrice ${}^tX D X Q$ est Q -symétrique, ses r valeurs propres non nulles sont réelles positives et ses vecteurs propres forment une base Q -orthonormée de $\text{Im } {}^tX$.*

6. En déduire une propriété analogue.

Partie II

Les valeurs propres sont rangées dans l'ordre décroissant dans la matrice diagonale Λ : $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r$.

On note $U = [U_1, \dots, U_r]$ et $V = [V_1, \dots, V_r]$ les vecteurs propres de ${}^tX D X Q$ et $X Q {}^tX D$ associés aux valeurs propres $\{\lambda_1, \dots, \lambda_r\}$ respectivement Q -orthonormés et D -orthonormés.

1. (a) En déduire les deux relations définissant la matrice $U = [U_1, \dots, U_r]$.
- (b) En déduire les deux relations définissant la matrice $V = [V_1, \dots, V_r]$.
- (c) Montrer que $X Q U$ est vecteur propre de $X Q^t X D$. En déduire la formule de transition $V = X Q U \Lambda$.
- (d) En déduire la formule de transition pour U en fonction de V .
2. (a) Justifier que $U^t U Q$ représente la matrice de la projection Q -orthogonale sur $Im U$.
- (b) En utilisant la première formule de transition, démontrer que $X = V \Lambda^t U$.

On en déduit :

Proposition 0.2 *La décomposition en valeurs singulières (DVS) du triplet (X, Q, D) s'écrit $X = V \Lambda^t U = \sum_{i=1}^r \sqrt{\lambda_i} U_i^t V_i$.*

Partie III

1. Déterminer les expressions de (X, Q, D) pour décrire une acp et une acp normée.
2. Montrer que la DVS de (X, Q, D) est équivalente à la DVS de $(Z = D^{1/2} X Q^{1/2}, I_p, I_n)$.

Remarque 0.3 *Les différentes méthodes factorielles étudiées, ACP, AFC, ACM, AFD, peuvent se traduire par une DVS d'un triplet bien choisi. On dispose alors du théorème suivant pour trouver la solution optimale :*

Proposition 0.4 *Théorème d'approximation d'Eckart-Young.*

Soit une matrice réelle $X_{n \times p}$ de rang r , $X = V \Lambda^t U = \sum_{i=1}^r \sqrt{\lambda_i} U_i^t V_i$ la DVS de $(X Q D)$ et k un entier inférieur ou égal à r .

On note $U_k = [U_1, \dots, U_k]$ et $V_k = [V_1, \dots, V_k]$ les matrices extraites de U_k et V_k et $\Lambda = \text{diag}(\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k)$ la matrice diagonale des k premières valeurs singulières.

On cherche un élément Z_k de l'ensemble des matrices réelles E_k ($n \times p$) de rang k le plus proche de X au sens de la norme $\|X\|_{Q,D}^2 = \text{tr}(X Q^t X D)$.

Alors $\min_{E_k} \|X - X_k\|_{Q,D}^2 = \|X - Z_k\|_{Q,D}^2 = \sum_{i=k+1}^r \lambda_i$.

et le minimum est atteint pour $Z_k = V_k \Lambda_k^t U_k$.

Exercice 2

Une entreprise veut choisir un nom pour un nouveau produit, qui doit paraître de qualité supérieure, prestigieuse, luxueuse, convenant à un public masculin, connaisseur, raffiné, distingué, de niveau socio-économique élevé... L'entreprise hésite entre 12 marques (cf 1ère ligne du tableau ci-dessous). Pour faire un choix entre ces marques, un échantillon de clients potentiels a été interrogé. Onze attributs leur étaient proposés (cf. 1ère colonne du tableau). Pour chaque attribut, la personne interrogée devait désigner une ou plusieurs marques auxquelles l'attribut se rapportait le mieux. Les résultats obtenus sont les suivants : On traite ce tableau par une méthode d'analyse des données multidimensionnelles. Les principaux résultats sont rassemblés dans les tableaux et graphiques ci-dessous :

1. (a) Quelle est la méthode d'analyse utilisée ici ? Pourquoi utilise-t-on cette méthode ?
- (b) Quelle information nous apporte le test du Khi deux présenté en 4.
2. (a) Dans le tableau des fréquences lignes, on lit 20,35 à l'intersection de la ligne "sobre-élégant" et de la colonne "alezan". Comment a été obtenue cette valeur ? Que signifie-t-elle ?

- (b) Comparer qualitativement le profil de la marque "Corsaire" au profil moyen.
 - 3. (a) Au vu du tableau des valeurs propres, combien de variables factorielles faudrait-il étudier dans le cadre d'une étude complète des données observées ?
 - (b) Quel résultat manque-t-il pour aider au choix des axes ?
 - (c) Dans cette étude, on se limite aux trois premiers axes. Quel pourcentage d'inertie est expliqué par ces trois axes ?
 - 4. (a) Rappeler la définition de la notion de qualité de représentation d'un point.
 - (b) Les points sont-ils bien représentés dans le plan F1-F2 ?
 - 5. Etude de la première variable factorielle
 - (a) On considère le nuage des attributs. Quels sont les individus dont la contribution est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante. Comment peut-on interpréter cet axe en termes d'opposition entre attributs.
 - (b) Même question pour le nuage des marques envisagées.
 - 6. (a) Mener une étude analogue pour la deuxième variable.
 - 7. (a) Faire une synthèse des deux études précédentes en décrivant les résultats obtenus dans le premier plan factoriel.
 - 8. (a) A votre avis, quelle est la marque qui a finalement été retenue par l'entreprise ?
- Analyse de documents