**Université de Nantes, Département de mathématiques** **2006**
**Master MIM première année**

# Première session d'examen d'Analyse des données

MAI 2005

Documents interdits – Durée 3 heures

## I Question théorique (7 points)

On se place dans l'espace $\mathbb{R}^p$ muni du produit scalaire standard

$$<X,Y> = \sum_{j=1}^{p} x_j y_j \ , \ \forall X = \sum_{j=1}^{p} x_j e_j \ \text{ et } \ Y = \sum_{j=1}^{p} y_j e_j$$

où $e_1,e_2,\ldots,e_p$ désigne la base canonique de $\mathbb{R}^p$. On note désormais par la même lettre X, Y,… un point ou un vecteur de $\mathbb{R}^p$ et sa matrice colonne dans la base canonique de $\mathbb{R}^p$.

On considère le nuage

$$\mathcal{N}(I) = \{(X_i,1) \mid i \in I = \{1,2,\ldots,n\}\}$$

de n points de $\mathbb{R}^p$ affectés de poids uniformes $p_i=1$.

On note X la matrice ayant pour lignes $X_1, X_2,\ldots, X_n$.

**1.** **a)** Qu'appelle-t-on dispersion, $dis_0(\mathcal{N}(I))$, du nuage $\mathcal{N}(I)$ par rapport à l'origine 0 de $\mathbb{R}^p$?

**b)** Etant donné un vecteur unitaire $u$ de $\mathbb{R}^p$ qu'appelle-t-on dispersion, $dis_{0,u}(\mathcal{N}(I))$, de $\mathcal{N}(I)$ suivant $u$ d'origine 0?

**c)** Quelle relation lie $dis_0(\mathcal{N}(I))$ aux dispersions $dis_{0,ej}(\mathcal{N}(I))$ suivant les divers vecteurs $e_j$ de la base canonique?

**d)** Justifier la relation

$$dis_{0,u}(\mathcal{N}(I)) = {}^t u \, {}^t X \, X \, u$$

On note désormais U cette matrice de dispersion : $U= {}^tX \, X$ et on suppose que U est non nulle.

**e)** Donner des propriétés de la matrice U. En particulier, que sait-on de ses valeurs propres et espaces propres associés et quelle est la dispersion de $\mathcal{N}(I)$ expliquée par un vecteur propre unitaire?

**f)** *Exemple numérique* : $p$=3, $n$=2, $X_1 = (2,0,-1)$, $X_2 = (0,1,0)$.
Déterminer la matrice de dispersion U de ce nuage $(X_1, X_2)$, les valeurs propres et vecteurs propres associés de celle-ci. Préciser la direction de plus grande dispersion et un sous espace qui explique au moins 90% de la dispersion.

**2.** On revient au cas général mais on suppose, pour simplifier, que toutes les valeurs propres non nulles de U sont simples. On note $\lambda_1 > \lambda_2 > \ldots > \lambda_p$ les valeurs propres de U et $u_1, u_2, \ldots, u_p$ des vecteurs propres unitaires associés formant une base orthonormée.

On considère maintenant le nuage des points $\mathcal{N}(J)$ de $\mathbb{R}^n$ associés aux colonnes de la matrice X. On suppose, comme pour $\mathbb{R}^p$, que $\mathbb{R}^n$ est muni du produit scalaire euclidien standard.

**a)** Que vaut la matrice de dispersion V du nuage $\mathcal{N}(J)$ par rapport à 0?

**b)** On note $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$ les valeurs propres ordonnées par valeurs décroissantes de V et $v_1, v_2, \ldots, v_n$ des vecteurs propres unitaires associés.

Etablir l'égalité des premières valeurs propres $\lambda_1 = \mu_1$.

**c)** En admettant l'égalité des valeurs propres non nulles suivantes, établir les relations

$$\begin{cases} v_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} X u_\alpha \\ u_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}} {}^t X v_\alpha \end{cases}$$

pour tout $\alpha$ tel que $\lambda_\alpha = \mu_\alpha \neq 0$

**d)** Vérifier ces relations sur l'exemple numérique de la question 1f).

## II - Etude d'un tableau à l'aide d'une AFC (6 points)

Soit le tableau de contingence $T=\begin{array}{c} A \\ B \\ C \\ D \\ E \end{array}\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \\ 1 & 0 & 2 \\ 0 & 2 & 2 \end{bmatrix}$ croisant les 5 modalités A, B …,E d'une

$\begin{matrix} X & Y & Z \end{matrix}$

première variable et les 3 modalités X,Y,Z d'une seconde variable.

### 1. Profils ligne et colonne

Calculer le tableau des fréquences relatives F, des fréquences marginales $f_{i.}$ et $f_{.j}$ et les profils lignes et colonnes L et C.

### 2. Ajustement du nuage des profils lignes (7 pts)

On rappelle que la métrique utilisée dans le nuage des profils lignes est $D_J^{-1}$ avec $D_J$ la matrice diagonale ($f_{.j}$) et la matrice des poids est $D_I$ la matrice diagonale ($f_{i.}$).

On rappelle que la dispersion du nuage des profils lignes L suivant le vecteur *u* unitaire d'origine 0 est

$${}^t u\, D_J^{-1}\, {}^t L D_I L\, D_J^{-1} u = {}^t u\, D_J^{-1}\, {}^t X\, D_I^{-1}\, X\, D_J^{-1} u.$$

**a)** Construire la matrice $X = \left[\frac{f_{ij}}{\sqrt{f_{i.}} \times \sqrt{f_{.j}}}\right] = \left[\frac{n_{ij}}{\sqrt{n_{i.}} \times \sqrt{n_{.j}}}\right]$

**b)** On en déduit ${}^tXX = \begin{bmatrix} 11/18 & 1/6 & \sqrt{2}/9 \\ 1/6 & 1/2 & \sqrt{2}/6 \\ \sqrt{2}/9 & \sqrt{2}/6 & 13/18 \end{bmatrix}$ (résultat admis).

Que représente cette matrice ?

**c)** Calculer les valeurs propres de ${}^tXX$, $\lambda_0 \geq \lambda_1 \geq \lambda_2$ (Pour vérification: $\lambda_1 = \frac{1}{2}$ et $\lambda_2 = \frac{1}{3}$).

**d)** Que représente un vecteur propre associé à $\lambda_0$ (il n'est pas demandé de le calculer)?

**e)** Déterminer les vecteurs propres unitaires pour la norme classique $u_1^*$ et $u_2^*$ associé à $\lambda_1$ et $\lambda_2$. Montrer que les vecteurs $D_J^{1/2} u_1^*$ et $D_J^{1/2} u_2^*$ sont unitaires pour la métrique $D_J^{-1}$.

**f)** Quelle est l'inertie du nuage et le % d'inertie expliqué par les axes 1 et 2.

### 3. Représentation du nuage

On note $F_1$, $F_2$ les facteurs principaux associés aux profils lignes.

**a)** Justifier la relation $F_1 = LD_J^{-1/2} u_1^*$.

**b)** Calculer les facteurs principaux $F_1$ et $F_2$.

**c)** Calculer les facteurs principaux $G_1$ et $G_2$ pour les profils colonnes à l'aide des formules de transition $G_1 = \sqrt{\lambda_1} D_J^{-1/2} u_1^*$ et $G_2 = \sqrt{\lambda_2} D_J^{-1/2} u_2^*$

**d)** Représenter dans un même plan les profils des deux variables.

# III Analyse de documents (7 points)

L'activité de différents vendeurs a été étudiée à partir de quatre variables quantitatives :
- contact : nombre de contacts nouveaux clients,
- rencontre : proportion de contacts avec rendez-vous,
- appels : appels téléphoniques reçus,
- visites : nombre de nouveaux comptes visités.

Les vendeurs ont été classés selon leur réussite à un concours de vente :
- G : gagnant,
- C : prix de consolation,
- S : sans succès.

Le tableau contient 15 individus de chaque classe, un extrait est donné ci-dessous:

| vendeur | contact | rencontre | appels | visites | classe |
|---|---|---|---|---|---|
| KZV | 130 | 62 | 148 | 42 | G |
| BOR | 122 | 70 | 186 | 44 | G |
| NUA | 89 | 68 | 171 | 32 | G |

Les résultats de l'analyse discriminante sont présentés en annexe.

**1.** **a)** Qu'appelle-t-on fonction linéaire discriminante?
**b)** Rappeler le critère utilisé pour déterminer les fonctions linéaires discriminantes.
**c)** Combien de fonctions linéaires discriminantes peut-on déterminer dans cet exemple.

**2.** Interpréter les informations apportées en j.

Dans la suite, on suppose que les trois classes suivent des lois multinormales.

**3.** **a)** Quelle est la dimension de ces lois?
**b)** Donner une estimation de la moyenne du groupe G.
**c)** Donner une estimation de la matrice des covariances sous l'hypothèse où elle est identique dans les trois classes.

**4.** Le test de Kullback est présenté au f. Il porte sur la différence entre les matrices des covariances.
**a)** Quelle est l'hypothèse nulle ?
**b)** Interpréter le résultat obtenu ici.

**5.** Les tests de Bartlett sont présentés au k. Ils portent sur l'égalité des moyennes.
**a)** Quelle est l'hypothèse nulle pour le test portant sur F1-F2 et sur F2 seul respectivement.
**b)** Interpréter le résultat obtenu ici.

**6.** La qualité de classement obtenue a été calculée sur l'échantillon d'apprentissage.
**a)** Comment interpréter le tableau l. Comment est déterminé le classement *a posteriori*?
**b)** Déterminer les individus mal classés dans ce tableau.

**7.** Les tableaux m et n propose le bilan du classement en AFD linéaire et quadratique.
**a)** Expliquer la différence entre ces deux méthodes.
**b)** Quel est le % de bon classement pour un individu issu de la population C en AFD linéaire.
**c)** Quel est le % de bon classement globale en AFD linéaire.

**8.** Proposer deux méthodes pour améliorer l'évaluation de la qualité du classement en AFD.

*fin du I*

**3. a)** Rappeler comment s'expriment les composantes $a_\alpha$ d'un vecteur quelconque $Y = \sum_{\alpha=1}^{p} a_\alpha u_\alpha$ de $\mathbb{R}^p$ en fonction de Y et des $u_\alpha$.

**b)** En déduire la relation

$$\left( \sum_{\alpha=1}^{p} u_\alpha {}^t u_\alpha \right) Y = Y \text{ pour tout } Y \in \mathbb{R}^p$$

puis la relation

$$\left( \sum_{\alpha=1}^{p} u_\alpha {}^t u_\alpha \right) = I.$$

**c)** Vérifier cette relation sur l'exemple numérique de la question 1f).

**d)** Etablir la relation

$$X = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha} v_\alpha {}^t u_\alpha$$

**e)** Vérifier cette relation sur l'exemple numérique de la question 1f).

**f)** Commenter.