Université de Nantes, Département de mathématiques **2004 - 2005**
Master professionnel 1ère année Ingénierie Mathématique

Correction de la première session d'examen d'Analyse des données

I Question théorique (7 points)

1. a) $\text{dis}_0(\mathcal{N}(\text{I})) = \sum_{i=1}^{n} p_i OX_i^2 = \sum_{i=1}^{n} OX_i^2 = {}^tXX$

b) d) $\text{dis}_{0,u}(\mathcal{N}(\text{I})) = \sum_{i=1}^{n} p_i < X_i, u >^2 = {}^tu{}^tXXu$

c) $\text{dis}_0(\mathcal{N}(\text{I})) = \sum_{j=1}^{p} \text{dis}_{0,\text{ej}}(\mathcal{N}(\text{I}))$

e) U est une matrice symétrique, (semi) définie symétrique. Ses valeurs propres sont positives et les sous espaces propres sont orthogonaux deux à deux. La dispersion expliquée par un vecteur propre unitaire est égale à la valeur propre correspondante.

f) $U=\begin{pmatrix} 2 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 2 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}=\begin{pmatrix} 4 & 0 & -2 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}$

$|U-\lambda I| = (1-\lambda)[(4-\lambda)(1-\lambda)-4]=(1-\lambda)[\lambda^2-5\lambda]$ dont les racines sont $\lambda_1=5$ $\lambda_2=1$ et $\lambda_3=0$

Soit $u_1\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ qui vérifie $\begin{cases} -x-2z=0 \\ y=0 \\ -2x-4z=0 \end{cases}$ donc de la forme $\begin{pmatrix} -2z \\ 0 \\ z \end{pmatrix}$. u_1 est unitaire pour $z=-\frac{1}{\sqrt{5}}$

Soit $u_2\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ qui vérifie $\begin{cases} 3x-2z=0 \\ y=y \\ -2x=0 \end{cases}$ donc de la forme $\begin{pmatrix} 0 \\ y \\ 0 \end{pmatrix}$. u_2 est unitaire pour y=1.

Soit $u_3\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ qui vérifie $\begin{cases} 4x-2z=0 \\ y=0 \\ -2x+z=0 \end{cases}$ donc de la forme $\begin{pmatrix} x \\ 0 \\ 2x \end{pmatrix}$. u_1 est unitaire pour $x=\frac{1}{\sqrt{3}}$

u_1 est la direction de plus grande dispersion ($\frac{5}{6}$ <0.9). Un sous espace permettant d'expliquer plus de 90% (ici 100%) est celui engendré par (u_1, u_2).

2. a) $V = X\,{}^tX$

b) Pour $\lambda_\alpha \neq 0$, u_α vérifie ${}^tXXu_\alpha = \lambda_\alpha u_\alpha$ donc $X{}^tXXu_\alpha = \lambda_\alpha Xu_\alpha$. Comme $Xu_\alpha \neq \vec{O}$, Xu_α est un vecteur propre de $X{}^tX$ associé à la valeur propre λ_α.

Pour $\mu_\alpha \neq 0$, v_α vérifie $X{}^tXv_\alpha = \mu_\alpha v_\alpha$ donc ${}^tXX{}^tXv_\alpha = \mu_\alpha {}^tXv_\alpha$. Comme ${}^tXv_\alpha \neq \vec{O}$, ${}^tXv_\alpha$ est un vecteur propre de ${}^tXX$ associé à la valeur propre μ_α.

Les deux matrices ont donc leurs valeurs propres non nulles communes.

c) u_α vérifie ${}^tXXu_\alpha = \lambda_\alpha u_\alpha$, on en déduit donc $X{}^tXXu_\alpha = \lambda_\alpha Xu_\alpha$ donc Xu_α est vecteur propre de $X\,{}^tX$ associé à λ_α de norme λ_α (${}^t(Xu_\alpha)Xu_\alpha = \lambda_\alpha$) donc $v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} Xu_\alpha$

d) $V=\begin{pmatrix} 2 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 2 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix}=\begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$ donc $\mu_1=5$ et $\mu_2=1$ et $v_1\begin{pmatrix} -1 \\ 0 \end{pmatrix}$ $v_2\begin{pmatrix} 0 \\ 1 \end{pmatrix}$

On a bien : $v_1 = \frac{1}{\sqrt{\lambda_1}} X u_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2/\sqrt{5} \\ 0 \\ -1/\sqrt{5} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\quad v_2 = \frac{1}{\sqrt{\lambda_2}} X u_2 = \begin{pmatrix} 2 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$u_1 = \frac{1}{\sqrt{\lambda_1}} {}^t X v_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2/\sqrt{5} \\ 0 \\ -1/\sqrt{5} \end{pmatrix}$ $\quad u_2 = \frac{1}{\sqrt{\lambda_2}} {}^t X v_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

Exercice II:

1. $F \begin{pmatrix} 1/12 & 0 & 0 \\ 1/12 & 1/12 & 0 \\ 0 & 0 & 1/6 \\ 1/12 & 0 & 1/6 \\ 0 & 1/6 & 1/6 \end{pmatrix}$ $\quad f_{i.} \begin{pmatrix} 1/12 \\ 1/6 \\ 1/6 \\ 1/4 \\ 1/3 \end{pmatrix}$ $\quad f_{.j} \begin{pmatrix} 1/4 & 1/4 & 1/2 \end{pmatrix}$ $\quad L \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \\ 1/3 & 0 & 2/3 \\ 0 & 1/2 & 1/2 \end{pmatrix}$ $\quad C \begin{pmatrix} 1/3 & 0 & 0 \\ 1/3 & 1/3 & 0 \\ 0 & 0 & 1/3 \\ 1/3 & 0 & 1/3 \\ 0 & 2/3 & 1/3 \end{pmatrix}$

2. a) $X \begin{pmatrix} 1/\sqrt{3} & 0 & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & 0 \\ 0 & 0 & 1/2\sqrt{3} \\ 1/3 & 0 & 1/3\sqrt{2} \\ 0 & 1/2\sqrt{3} & 1/2\sqrt{6} \end{pmatrix}$ $\quad$ **b)** ${}^tXX = \begin{bmatrix} 11/18 & 1/6 & \sqrt{2}/9 \\ 1/6 & 1/2 & \sqrt{2}/6 \\ \sqrt{2}/9 & \sqrt{2}/6 & 13/18 \end{bmatrix}$ représente la matrice ${}^t L D_I^{-1} L D_J^{-1}$ dont les vecteurs propres unitaires sont les axes principaux.

c) On résoud $|{}^tXX - \lambda I| = 0$

En multipliant la matrice par 18, on recherche alors la valeur propre $\lambda' = 18\lambda$, donc $\lambda'=18$ est solution, vérifiant :

$(11-\lambda')[(9-\lambda')(13-\lambda')-18] - 3[-3\lambda'+27] + 2\sqrt{2}\,[-9\sqrt{2}+2\sqrt{2}\lambda']$

$= 1089 + 11\lambda'^2 - 242\lambda' - 99\lambda' - \lambda'^3 + 22\lambda'^2 + 9\lambda' - 81 - 36 + 8\lambda' = -\lambda'^3 + 33\lambda'^2 - 324\lambda' + 972 = -[\lambda'-18][\lambda'^2 - 15\lambda' + 54]$

Soit $\lambda'=9$ et $\lambda'=6$ sont également solution.

On trouve donc $\lambda_0 = 1 \quad \lambda_1 = \frac{1}{2} \quad \lambda_2 = \frac{1}{3}$

d) Un vecteur propre associé à λ_0 représente l'axe (OG_I), axe trivial.

e) ${}^t u_1^* D_J^{1/2} D_J^{-1} D_J^{1/2} u_1^* = {}^t u_1^* u_1^* = 1$ donc $u_1 = D_J^{1/2} u_1^*$ (de même pour u_2)

f) L'inertie est égale à $\lambda_1 + \lambda_2 = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$

La proportion d'inertie projetée sur u_1 est donc de $\frac{1/2}{5/6} = \frac{3}{5}$ donc 60% et donc 40% sur u_2.

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

b) $F_1 = \begin{pmatrix} 4/\sqrt{6} \\ 2/\sqrt{6} \\ -2/\sqrt{6} \\ 0 \\ -1/\sqrt{6} \end{pmatrix}$ $F_2 = \begin{pmatrix} 1/\sqrt{3} \\ -1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \\ -1/\sqrt{3} \end{pmatrix}$

c) $G_1 = \sqrt{\lambda_1} D_J^{-1/2} u_1^* = \begin{pmatrix} 2/\sqrt{3} \\ 0 \\ -1/\sqrt{3} \end{pmatrix}$ $G_2 = \sqrt{\lambda_2} D_J^{-1/2} u_2^* = \begin{pmatrix} 1/3 \\ -1 \\ 1/3 \end{pmatrix}$

III Analyse de documents (7 points)

1. **a)** Une fonction discriminant est une combinaison linéaire des variables. Ce sont les vecteurs propres de $W^{-1}B$.

b) Il faut maximiser le rapport entre la variance inter et la variance intra de Xu, soit $\frac{{}^t uBu}{{}^t uTu}$

c) Le nombre est r=min(q-1,p), q le nombre d classe et p de variables soit ici q-1=2.

2. Le tableau j nous donne les fonctions discriminantes.

3. **a)** La dimension est p=4.

b) L'estimation est (104,60,155,37) d'après a.

c) On obtient une estimation de Σ en divisant W (c.) par 42:

253,3	-76,6	-92,9	29,1
-76,6	80,5	61,8	-8,0
-92,9	61,8	739,5	14,3
29,1	-8,0	14,3	28,8

4. **a)** L'hypothèse nulle est l'absence de différence entre les matrices des covariances intra.

b) La probabilité d'obtenir la valeur observée sous H_0 est 0.464. On accepte donc H_0.

5. **a)** L'hypothèse nulle pour le test portant sur F1-F2 est l'absence de différences globales entre groupes, et sur F2 l'absence de différence pour le dernier axe discriminant.

b) Pour F1-F2, on rejette l'hypothèse H_0 avec un risque de 1ère espèce très faible (<0.001), par contre le dernier axe ne semble pas montrer de différences significative (P=0.8).

6. **a)** Le tableau indique le groupe d'appartenance (a priori) et d'affectation par l'AFD (a posteriori). Il donne ensuite la valeur du score dans les 3 groupes et les coordonnées F1 F2.

L'affectation a posteriori est celle donnée par le score le plus grand.

b) Les individus mal classés sont 17 et 20.

7. **a)** L'analyse discriminante linéaire suppose l'égalité des matrices des covariances dans les différents groupes alors que celle quadratique utilise des matrices différentes dans chacun des groupes.

a) Il est de 13/15 soit 87%.

b) Il est de 42/45 soit 93%.

8. L'évaluation du taux d'erreur sur l'échantillon d'apprentissage est trop optimiste. Deux méthodes permettent une meilleure estimation : l'utilisation d'un échantillon test ou la validation croisée (bootstrap) (voir le cours).

fin du I

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

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.