Université du Québec à Montréal — Département de Mathématiques

STT3100-MAT8081 . Analyse Multivariée
Devoir 2 . 15 % de la note finale
Automne 2014

Partie I: partie théorique

Exercice 1

Supposons que nous sommes dans le cas d'une analyse discriminante avec k groupe. Soient $\mathbf{E}$ et $\mathbf{H}$ les matrices de sommes de carrés et produits intra- et inter-groupes respectivement. Soit A la matrice $d \times s, [s = rang(\mathbf{H})]$, définie par $A = [a_1, \ldots, a_s]$, où les $a_i \in R^d, i = 1 \ldots, s$, sont les vecteurs propres associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s$ de $\mathbf{E}^{-1}\mathbf{H}$. Les vecteurs a_i sont les s fonctions discriminantes qui distinguent davantage les k groupes.

i. Montrer que

$$A^T\mathbf{H}A = \Lambda \qquad \text{et que} \qquad A^T\mathbf{E}A = I_s,$$

où I_s est la matrice identité $s \times s$, et Λ est la matrice diagonale $s \times s$ avec $\Lambda_{ii} = \lambda_i, i = 1, \ldots, s$.

ii. Application: Un chercheur en psychiatrie s'intéresse à savoir si l'effet des familles (traitements ou groupes) a un impact sur la Schizophrénie (SZ). Afin de répondre à la question, le chercheur a mesuré les 4 variables (symptômes) suivantes : Délire, hallucination, manie et dépression sur 521 sujets provenant de 36 familles atteintes de la maladie. Les matrices $\mathbf{E}/\nu_E$ et $\mathbf{H}/\nu_H$, calculées à partir des données collectées, sont données comme suit :

Matrice ***E***	**Délire**	**hall**	**manie**	**dépres**
délire	0.1824	0.0770	0.2368	0.0535
hallucinations		0.2000	0.1989	0.0881
manie			1.1260	-0.0092
dépression				0.7461

Matrice ***H***	**Délire**	**hall**	**manie**	**dépres**
délire	0.2707	0.3724	-0.2078	-0.2814
hallucinations		0.6862	-0.2857	-0.3957
manie			1.1578	0.4399
dépression				0.4846

a. Utiliser les corrélations et les covariances inter- et intra-familles entre les variables pour interpréter (dans le context de l'exercice) l'association entre les variables deux à deux (Délire-manie), (manie-Dépression) et (Délire-Dépression). Indication: utiliser le logiciel R pour calculer les deux matrices de corrélations associées aux $\mathbf{E}/\nu_E$ et $\mathbf{H}/\nu_H$.

b. À l'aide de la fonction eigen() du logiciel R, calculer la première fonction discriminante et la statistique de Roy. Au seuil 0.05, tester si la maladie est héréditaire. Indication: les quantiles de la loi beta multidimensionnelles sont donnés dans la table A.10 du livre, page 575. Pour plus de détails à propos de la statistique de Roy consultez la section 6.1.4 du livre.

Exercice 2

Soit $Y_1, \ldots, Y_n$ un échantillon aléatoire de R^d. Sans perte de généralité, nous supposons que les Y_i sont centrés, c.-à-d. $\bar{Y} = 0_d$. Dans les notes de cours, nous avons montré que l'ACP cherche une combinaison linéaire $Z = a^T Y$ avec une variance totale maximale. Dans cet exercice, nous allons montrer que l'ACP peut être vue comme une approche qui cherche à projeter les données sur des directions qui minimisent l'erreur quadratique moyenne (distance) entre les vecteurs Y_i et leurs projections sur les composantes principales. Autrement dit, les composnates principales sortantes de l'ACP minimisent la somme de carrés résiduelles suivante:

$$SCR = \sum_{i=1}^{n} ||Y_i - (a^T Y_i)a||^2.$$

i. Montrer que

$$Arg \min_{\substack{a \in R^d \\ ||a||=1}} SCR = Arg \max_{\substack{a \in R^d \\ ||a||=1}} \sum_{i=1}^{n} (a^T Y_i)^2.$$

ii. Déduire que le vecteur qui minimise SCR est exactement la première composante principale sortante de l'ACP vue dans les notes de cours.

Exercice 3

Soit R la matrice de corrélation $d \times d$ suivante:

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1d} \\ r_{12} & 1 & r_{ij} & \vdots \\ \vdots & r_{ij} & \ddots & r_{(d-1)d} \\ r_{1d} & \dots & r_{(d-1)d} & 1 \end{pmatrix}.$$

Soit $R_{(\alpha)}$ une deuxième matrice de corrélation définie par

$$R_{(\alpha)ij} = \alpha\, r_{ij}, i \neq j, \qquad R_{(\alpha)ii} = 1,$$

avec

$$-1 \leq \alpha\, r_{ij} \leq 1, \qquad \text{pour tout} \quad i \neq j \in \{1, \dots, d\}.$$

i. Montrer que tout vecteur propre a_α de R_α est aussi vecteur propre de R.

ii. Soient λ et δ les valeurs propres de R et de R_α, respectivement, qui sont associées au vecteur propre a_α. Déduire que

$$\frac{\lambda - 1}{\delta - 1} = \frac{1}{\alpha}.$$

Conclusion: Cet exercice nous montre que les composantes principales sortantes d'une ACP avec matrice de corrélation peuvent servir comme composantes principales pour une autre matrice de corrélation (de type R_α). Ainsi, les composantes principales issues des matrices de corrélations ne dépendent pas des corrélations r_{ij} ni des corrélations $r_{(\alpha)ij}$, mais elles dépendent du ratio des corrélations $r_{ij}/r_{(\alpha)ij} = 1/\alpha$.

Partie II: partie pratique

Exercice 4

Le jeu de données *FOOTBALL* de la page 24 des notes de cours du chapitre 6 contient 6 variables anthropométriques mesurées sur 90 hommes formant trois groupes. Le jeu de données se trouve dans Moodle → Laboratoires → Données.

i. Dans cette question nous voulons faire une analyse discriminante (analyse de classification) à l'aide de la fonction discriminante linéaire normale:

   a. Calculer les trois erreurs de classification à l'aide des estimés apparentes, les estimés plug-in et les estimés de la validation croisée. Comparer les estimés. Que peut-on conclure?

   b. En utilisant l'option MAHALANOBIS, lequel des trois groupes est significativement différent des autres? Justifier vos réponses.

   c. Tester l'homogénéité des matrices de variances-covariances des trois groupes au seuil $\alpha = 1\%$.

ii. Refaire l'analyse discriminante sur le même jeu de données à l'aide de la fonction discriminante quadratique, et calculer les erreurs de classification à l'aide des estimés apparentes et les estimés de la validation croisée. Comparer les estimés. Que peut-on conclure?

iii. Parmi les 6 variables, lesquelles distinguent davantage les trois groupes? [*Indication*: vous pouvez faire une analyse MANOVA à l'aide de proc GLM, calculer le vecteur a_1 (la fonction discriminante linéaire), standardiser a_1, etc...].

iv. Dans cette question nous voulons faire une analyse en composantes principales sur le jeu de données FOOTBALL en gardant seulement les deux groupes qui sont non significativement différents (qui sont homogènes):

   a. Faire une ACP sur ce nouveau jeu de données en utilisant la matrice de variances-covariances et la matrice de corrélation.

   b. À la lumière des critères vus en classes, choisir le nombre de composantes approprié à retenir dans les deux cas.

c. Laquelle des deux analyses (ACP avec **S** ou ACP avec **R**) est appropriée pour ce jeu de données.

d. Les quatre variables WDIM, CIRCUM, FBEYE et JAW sont des mesures verticales du cran tandis que les variables EYEHD et EARHD sont des mesures horizontales du cran. Cette analyse reflète-elle la structure (pattern) des 6 variables? Expliquer.

v. Dans cette question nous voulons faire une analyse factorielle sur le jeu de données de la question **iv)**:

a. Faire une analyse factorielle sur ce jeu de données et déterminer le nombre approprié de facteurs à retenir.

b. Vérifier s'il y a un regroupement naturel entre les 6 variables. Utiliser la rotation des facteurs pour mieux visualiser l'association entre facteurs et variables.

c. Comparer le *pattern* des variables sortant de cette analyse à celui obtenu par l'ACP.