Université du Québec à Montréal — Département de Mathématiques

STT3100-MAT8081 . Analyse Multivariée
Devoir 1 . 15 % de la note finale
Automne 2014

Partie I: partie théorique

Exercice 1

Méthode d'orthogonalisation de Gram-Schmidt: Soient $x_1, x_2, \ldots, x_n$, n vecteurs linéairement indépendants de R^d ($n \leq d$). On construit les vecteurs $u_1, u_2, \ldots, u_n$ à partir des vecteurs x_i de la façon suivante:

i. $u_1 = x_1/||x_1||$

ii. $u_i = \frac{x_i - \sum_{j=1}^{i-1} u_j^T x_i u_j}{||x_i - \sum_{j=1}^{i-1} u_j^T x_i u_j||}$, $i = 1, \ldots, n$.

a. Montrer que les u_i sont orthogonaux entre eux et ont une norme égale à 1.

b. Montrer que $V(x_1, x_2, \ldots, x_n) = V(u_1, u_2, \ldots, u_n)$, c'est-à-dire que les x_i et les u_i engendrent le même espace vectoriel. Pour établir b), prendre un vecteur Xa de $V(x_1, x_2, \ldots, x_n)$ et montrer que ce vecteur peut également s'écrire comme Ub où U est la matrice $d \times n$ des u_i et montrer que $b = Ca$ où C est la matrice $n \times n$ de changement de base telle que X=UC.

c. Montrer que la matrice de changement de base C est triangulaire supérieure et que ses éléments satisfont C_{ii}=distance entre x_i et sa projection sur $V(u_1, u_2, \ldots, u_{i-1})$ (ou entre x_1 et l'origine lorsque $i = 1$) et que C_{ki}=coordonnée de u_k dans la projection de x_i sur $V(u_1, u_2, \ldots, u_{i-1})$, $k = 1, 2, \ldots, i-1$.

d. Soit $S = X^T X$, une matrice $n \times n$. Montrer que $S = C^T C$; on dit que C est la décomposition de Cholesky de S. Donner un algorithme itératif pour construire les éléments de C à partir de ceux de S (C_{ii} s'écrit en fonction de S_{ii} et de $C_{ki}, k =$

$1, ..., i-1$ et $C_{ij}, j > i$, s'écrit en fonction de S_{ij} et de $\{C_{ki}, C_{kj} : k = 1, \ldots, i\}$. Justifier vos résultats. (Suggestion utilisez le fait que pour $j \geq i$, $S_{ij} = \sum_{k=1}^{i} C_{ki}C_{kj}$).

Remarque: Étant donné une matrice définie positive quelconque S, cet algorithme permet de calculer la décomposition de Cholesky de S, une matrice triangulaire supérieure C telle que $S = C^T C$.

Application: Prendre $d = 5$ et $n = 3$ et poser

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & -3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

a. Construire les vecteurs u_i et calculer C.

b. Calculer S et recalculer C à l'aide de l'algorithme du problème 1 d).

Exercice 2

Soit $y = (y_1, y_2)^\top \in \Re^2$ un vecteur aléatoire distribué selon une loi normale de moyenne $\mu = (\mu_1, \mu_2)^\top$ et de variance Σ, i. e. $y \sim N_2(\mu, \Sigma)$. Soit $U = [u_1; u_2]$ une matrice 2×2 orthogonale et $D = diag(\lambda_1, \lambda_2), \lambda_1 > \lambda_2 > 0$, une matrice diagonale 2×2 telle que:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = UDU^\top.$$

Soit $w_1 = u_1^\top y$, $w_2 = u_2^\top y$, $\alpha_1 = u_1^\top \mu$ et $\alpha_2 = u_2^\top \mu$.

i. Montrer que w_1 et w_2 sont les coordonnées de y sur u_1 et u_2 respectivement.

ii. Montrer que α_1 et α_2 sont les coordonnées de μ sur u_1 et u_2 respectivement.

iii. Déduire la loi de w.

a. Les iso-contours d'une loi normale bidimensionnelle sont des ellipses centrés en μ. On les obtient comme ceci:

$$C = \{y \in \Re^2 : f_y(y) = c\},$$

où $f_y(y)$ est la fonction de densité de y et c est une constante positive.

b. Écrire l'equation de l'ellipse en fonction de c, w, λ_1, λ_2, α_1 et α_2; montrer que l'ensemble C décrit une ellipse dans $\Re^2$.

c. Déduire les axes principaux de l'ellipse.

Application: Prendre $\mu = (2, 3)^\top$, $\sigma_1^2 = \sigma_2^2 = 3$, $\sigma_{12} = 2$ et $c = 4$.

1. Écrire un programme R permettant de calculer les points, (y_1, y_2), de l'ellipse C. [Indication: écrire l'équation de l'ellipse sous la forme $\cos(\theta)^2 + \sin(\theta)^2 = 1$, générer un vecteur des θ, à l'aide de ceci: θ = seq(0, 2*pi, length.out=200), et calculer les 200 points (y_{i1}, y_{i2})].

2. Faire un graphique de l'ellipse pour les 200 points (y_{i1}, y_{i2}) et rajouter les axes principaux.

Exercice 3

Soient A et B deux matrices $d \times d$ avec A semi-définie positive et B définie positive et soit b un vecteur de R^d.

a. Montrer que

$$\max_a \frac{(a^T b)^2}{a^T B a} = b^T B^{-1} b,$$

avec le maximum est atteint pour tout vecteur $a = cB^{-1}b$, avec $c \in R$. [Indication: utiliser l'inégalité de Cauchy–Schwartz pour répondre à cette question, voir définiton à la fin de la partie I du devoir].

b. Montrer que

$$\max_a \frac{a^T A a}{a^T B a} = \lambda_1,$$

où λ_1 est la plus grande valeur propre de $B^{-1}A$. Conclure que le maximum est atteint pour le vecteur propre de $B^{-1}A$ qui est associé à λ_1.

Exercice 4

Application de l'exercice 3: Suppposons maintenant que nous sommes dans une situation ou on rejette l'hypothèse nulle dans un test multivarié à deux échantillons

$$\begin{cases} \text{échantillon I:} & y_{11}, \ldots, y_{n1} \quad i.i.d \sim N_d(\mu_1, \Sigma), \\ \text{échantillon II:} & y_{12}, \ldots, y_{n2} \quad i.i.d \sim N_d(\mu_2, \Sigma). \end{cases}$$

Nous somme donc dans la situation où l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ est rejetée. On suppose aussi que Σ est inconnue, comme dans la section 2.2 (b) du chapitre 3 des notes de cours.

a. Montrer que si $H_0 : \mu_1 = \mu_2$ est rejetée, alors il existe au moins un vecteur $a \in R^d$ (c.-à-d. une direction dans R^d) tel que l'hypothèse univariée $H_0 : a^T\mu_1 = a^T\mu_2$ est rejetée.

b. Supposons maintenant les variables aléatoires Z_1 et Z_2 définies par $Z_{i1} = a^T Y_{i1}$ et $Z_{i2} = a^T Y_{i2}$ pour $i = 1, \ldots, n$. Donnez la forme explicite de la statistique de Student, $t_z(a)$, du test $H_0 : \mu_{z1} = \mu_{z2}$ en fonction de a, $\bar{y}_1$, $\bar{y}_2$ et S_p où μ_{z1} et μ_{z2} sont les moyennes théoriques des deux variables Z_1 et Z_2 respectivements. La matrice S_p est la matrice de variances-covariances combinée donnée dans la page 17 des notes de cours (chapitre 3), et $\bar{y}_1$ et $\bar{y}_2$ se sont les vecteurs moyens des deux échantillons 1 et 2 respectivement.

c. Montrer que

$$\max_a \left\{[t_z(a)]^2\right\} = T^2,$$

où T^2 est la statistique de Hotelling qui teste $H_0 : \mu_1 = \mu_2$. Montrer que le maximum est atteint pour $a = S_p^{-1}(\bar{y}_1 - \bar{y}_2)$. Le vecteur a est appelé la fonction discriminante, (c.-à-d. c'est la direction selon laquelle les deux populations se distinguent davantage). La statistique T^2 est donnée dans la page 17 du chapitre 3 des notes de cours.

d. Que peut-on conclure alors lorsque nous rejetons $H_0 : \mu_1 = \mu_2$ par la statistique T^2? Les entrées du vecteur $a = S_p^{-1}(\bar{y}_1 - \bar{y}_2)$, sont-elles interpétables? Justifiez vos réponses.

Rappel 1

INÉGALITÉ DE CAUCHY–SCHWARTZ (EXTENSION): Soient a et b deux vecteurs quelconques de R^d, et soit B une matrice $d \times d$ définie positive. Alors nous avons les résultats suivants:

$$\begin{cases} \text{inégalité de Cauchy–Schwartz} & (a^T b)^2 \leq (a^T a)(b^T b), \\ \text{une extension de cette inégalité donne} & (a^T b)^2 \leq (a^T B a)(b^T B^{-1} b). \end{cases}$$

Partie II: partie pratique

Exercice 5

Le fichier *ExeV.dat* contient les mesures de d=3 variables mesurées sur $n = 51$ individus.

a. Tester la normailté univariée pour chacune des trois variables? Citer au moins deux statistiques vues dans le cours pour cet effet?

b. La normalité multivariée est-elle acceptable pour ces données ?

c. Utiliser R et SAS pour trouver la statistique de Hotelling et le p-value qui testent l'hypothèse $H_0 : \mu = (14.6, 26, 21)^T$?

Exercice 6

Exercice 5.18 du livre (page 150): La Table 5.6 du livre est disponible par moodle dans la section *Laboratoires –> Données*.

a. À l'aide du logiciel R, calculer la statistique de Hotelling T^2 ainsi que la statitistique F associée et tester l'hypothèse $H_0 : \mu_1 = \mu_2$.

b. Calculer la fonction discriminante a et interpréter ces entrées dans le contexte de l'exercice.

c. En utilisant un modèle de régression, comme dans les notes de cours, recalculer la statistique F à l'aide de SAS.

d. Si on rejette l'hypothèse nulle, tester l'importance de chaque variable à la présence des autres cinq variables [donnez la statistique du test appropriée et le p-value associé].

e. Maintenant, à l'aide de la correction de Bonferroni, refaire les tests univariés de l'importance de chaque variable dans la distinction des deux groupes.

f. En comparant les résultats de d) et e) que peut-on conclure de la correction de Bonferroni?