Examen d'analyse de données. Corrigé

1. **La métrique définie dans le test du Khi deux par l'écart *D* entre la loi théorique proposée (ensemble des *m* probabilités $\pi_1, \pi_2, \ldots \pi_m$, pour les *m* valeurs possibles $y_1, y_2, \ldots y_m$) et la loi observée (fréquences relatives des *m* valeurs $y_1, y_2, \ldots y_m$) est-elle une distance sur l'ensemble des lois finies à *m* éléments ?**
On dispose d'un ensemble de valeurs numériques discrètes, parmi 3 valeurs possibles. On souhaite tester, au risque de 5%, l'hypothèse que les probabilités de ces trois valeurs sont de 0.4 pour la première, 0.4 pour la deuxième, et 0.2 pour la troisième. Sur un jeu de 800 observations, on compte pour ces trois valeurs respectivement 300, 350, et 150 éléments. Quel est alors le résultat du test du Khi deux ?
La métrique s'écrit, en notant N_k la fréquence observée de y_k, et N le nombre total d'observations : $D = (N_1 - N\pi_1)^2 / N\pi_1 + (N_2 - N\pi_2)^2 / N\pi_2 + \ldots + (N_m - N\pi_m)^2 / N\pi_m = N [(f_1 - \pi_1)^2 / \pi_1 + (f_2 - \pi_2)^2 / \pi_2 + \ldots + (f_m - \pi_m)^2 / \pi_m]$, en posant $f_k = N_k / N$ la fréquence relative de y_k.
Pour 2 lois discrètes, de probabilités π_k et θ_k, on a donc :
$D(\pi, \theta) = N [(\theta_1 - \pi_1)^2 / \pi_1 + (\theta_2 - \pi_2)^2 / \pi_2 + \ldots + (\theta_m - \pi_m)^2 / \pi_m]$, et il ne s'agit pas d'une distance, car elle n'est pas symétrique.

Pour l'application du test du Khi deux, le risque $\alpha = 5\%$ correspond à un seuil de $-2\ (3) = 6$. En effet, la loi du Khi deux à 2 degrés de liberté, qui intervient ici, est une loi exponentielle X de paramètre $\lambda = 0.5$, d'où $\alpha = P(D > \eta) \cong 1 - F_X(\eta) = \exp(-\eta/2)$, soit $\eta = -2 \log(0.05)$.
L'écart D est ici : $(300 - 320)^2 / 320 + (350 - 320)^2 / 320 + (150 - 160)^2 / 160 = 75/16 = 4.69$
$D < 6$: on accepte donc l'hypothèse.

2. **On considère une matrice de données $A = \begin{bmatrix} 4 & -1 \\ 6 & 0 \\ 5 & 1 \end{bmatrix}$. Ecrivez la matrice centrée *B*, puis la matrice de variance-covariance associée. Déterminez les composantes principales et les deux axes principaux. Quelles sont les projections orthogonales des 3 lignes de *B* sur le premier axe principal ?**
La matrice centrée est $B = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$. La matrice de covariance est $V = 1/3 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Ses valeur propres sont 1 et 3. Le premier axe principal correspond à la valeur 3. Il a pour équation : $2x + y = 3x$, soit $y = x$. Le deuxième axe principal a pour équation : $2x + y = x$, soit $y = -x$. Il s'agit de l'axe orthogonal au précédent. La matrice de passage peut s'écrire $Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$.
Les composantes des lignes de *B* sur une base orthonormée portée par ces deux axes sont : $(-\sqrt{2}, 0)$; $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$; $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$

3. On souhaite faire la classification hiérarchique d'un nuage de *n* points en dimension quatre, en disposant d'une fonction d'écart entre points, *d*.

a) Indiquez une procédure détaillée, utilisant la fonction *d*, permettant de passer de *n* classes à une seule classe.

Soit $x_1, x_2, \ldots, x_n$ le nuage de points.
$k = n$ (k désigne le nombre de classes)
tableau : absent (n) = 0
création du tableau des x
pour i = 1 : n
moy (i) = x (i)
eff (i) = 1
fin

delta = infini
pour i = 1 : n
pour j = 1 : i
dist (i, j) = d (xi, xj)
si dist (i, j) < delta,
delta = dist (i, j)
c1 = i
c2 = j
fin
fin

C = In (matrice binaire des classes)
tant que k > 1
*moy (c1) = (moy (c1) * eff (c1) + moy (c2) * eff (c2))/(eff (c1) + eff (c2))*
eff (c1) = eff (c1) + eff (c2)
C(c1, :) = C (c1, :) + C(c2, :) (Regrouper classes c1 et c2 dans c1)
C(c2, :) = 0
absent (c2) = 1
pour i = 1 : n
si i ≠ c1 & absent (i) = 0
dist (c1, i) = d (moy(c1), moy (i))
fin
fin
trier le tableau des distances (des k – 1 distances nouvelle classe vs classes existantes)
Concaténer les tableaux triés distances existantes et nouvelles distances
delta = distance min
c1 = imin
c2 = jmin
fin

b) Proposez une méthode pour arrêter la classification, pour un ensemble de classes significatif.

On peut choisir un critère d'arrêt sur un seuil d'inertie, ou sur des variations d'inertie.
Si *inert* = inertie intra classe/inertie totale

Inertie intraclasse = somme des inerties des classes, plus précisément somme sur toutes les classes des sommes des carrés des distances entre éléments de chaque classe,

Inertie totale = somme des carrés des distances entre toutes les paires de points.

Soit on met un seuil (entre 0 et 1, par exemple à 0.5) pour la variable *inert*.

Soit on met un seuil sur les variations, ou (*inert* suivant – *inert*) / *inert*

c) Si *v* est le coût d'un appel de la fonction *d*, quelle est la complexité de l'étape a) dans le cas où l'algorithme construit une seule classe (non réduite à un point) de taille croissante, de 2 à *n* points.

On peut observer le comportement de la boucle sur la construction des classes. Pour les opérations algébriques, les additions sont en O (n^2), les multiplications et les divisions en o (n).
Pour les appels de la fonction *d*, il y en a n (n–1)/2 au départ, et k – 1 à l'étape k de la boucle, soit n^2 en tout.
Pour les comparaisons, on peut observer le pire cas, quand on compare les (k – 1) nouvelles distances aux (k – 1) (k – 2)/2 existantes : on trie les nouvelles, et on compare aux valeurs existantes, dans le pire cas : $k - 1)^2$ (k – 2)/2 comparaisons, ce qui donne en tout O (n^4).
C'est ce nombre de comparaisons qui pourra coûter le plus ici.
Si une multiplication coûte 64 cycles, on aura dans le pire cas : $192\ n + 3\ n^2 + n^2\ v + n^4/8$

d) Indiquez différentes solutions possibles pour la fonction d (qui peut ou non être une distance), et pour un choix particulier que vous proposerez précisez la complexité numérique de l'étape a)

Si on calcul la distance euclidienne entre les moyennes, il y a, en dimension 4 ,3 additions, 4 multiplications, et une racine carrée. La racine carrée est inutile, et on peut se contenter de calculer le carré de la distance euclidienne.

e) Que peut-on dire de la complexité de la classification hiérarchique dans le cas général ?

Le nombre de passages dans la boucle sera limité, mais il faut calculer les inerties successives. Si on construit une seule classe croissante, un calcule à chaque étape les carrés des distances du nouveau point à ceux existants, soit n – k, pour k variant de n à K.

Dans le pire cas on double le nombre de calculs de distance.