La Revue de Modulad

ANALYSE DISCRIMINANTE SUR VARIABLES CONTINUES ET QUALITATIVES
NOTICE SCIENTIFIQUE DU LOGICIEL ADM

Jean-Jacques Daudin et Mohamed Soukal

INAPG
Département de Mathématique et Informatique
16 rue Claude Bernard 75005
tel: 45 35 16 42

Résumé :
Le modèle Manova-loglinéaire ("location model") permet d'utiliser conjointement des variables continues et qualitatives dans une analyse discriminante. Cet article décrit ce modèle ainsi qu'une méthode de sélection des variables basée sur le critère d'Akaiké. Le logiciel ADM permet d'estimer les paramètres de ce modèle, les critères de sélection, les probabilités à postériori et les proportions de bien et de mal classés.

Mots-clés : Analyse Discriminante, Critère d'Akaiké, Mélange de variables, Modèle Manova-loglinéaire

1. Introduction.

Le mélange de variables continues et qualitatives se rencontre souvent en Analyse Discriminante. Par exemple, les notes de maladie pour les animaux ou pour des plantes sont utilisées conjointement avec des variables continues mesurant le rendement de différentes races ou espèces. Dans le domaine médical, les variables qualitatives comme le sexe, le statut social ou des résultats de tests sont utilisées comme variables discriminantes avec des variables continues pour obtenir un diagnostic.

La principale nouveauté induite par cette situation par rapport à celles abordées classiquement, est que les variables qualitatives induisent des sous-populations dans lesquelles les variables continues peuvent avoir des comportements discriminants différents. Par exemple, les hommes et les femmes peuvent avoir des symptômes différents pour une même maladie; une variété A de céréale produit davantage que la variété B au printemps et moins en été. Ce changement de modèle dans des sous-populations est appelé renversement ("reversal") par Moore(1973) et Krzanowski(1977). En termes d'analyse de la variance, il y a interaction entre le facteur population et la variable qualitative considérée sur la ou les variables discriminantes continues.

Une pratique courante pour traiter cette situation consiste à recoder les variables qualitatives en données binaires puis à utiliser l'Analyse Discriminante Linéaire sur l'ensemble des variables continues et binaires. Pour prendre en compte d'éventuels renversements, Knoke(1982) et Vlachonikolis et Marriot(1982) ont étudié la possibilité d'inclure parmi les variables discriminantes les produits des variables continues par les variables binaires. Cette procédure empirique a l'avantage d'utiliser directement les logiciels de discrimination linéaire, mais le nombre très important de variables discriminantes obtenues par les transformations ci-dessus implique la nécessité de la sélection, et le critère du F de Fisher est inadapté dans ce cas.

Une autre procédure empirique consiste à transformer les variables continues en variables qualitatives par découpage en classes, puis à utiliser les méthodes de discrimination sur variables qualitatives. Cette méthode comporte un risque de perte d'information et peut nécessiter l'estimation d'un trop grand nombre de paramètres.

Il existe plusieurs modèles qui prennent en compte la nature des variables: la discriminante logistique, les modèles non paramétriques et le modèle "Manova-loglinéaire". C'est ce dernier modèle qui est celui utilisé dans le logiciel ADM; nous le présentons au § 2. Dans le §3 nous traitons le problème de la sélection des variables de ce modèle. Le § 4 décrit le logiciel ADM et en donne les résultats sur un exemple simple.
Pour plus de détails sur toutes ces méthodes, on pourra consulter l'ouvrage de R. Tomassone et al. (1988).

2. Le modèle Manova-loglinéaire.

Olkin et Tate(1961) ont défini un modèle (appelé "location model") qui a été repris et développé par Krzanowski (1975 et 1980).

Soit $G_1, G_2, \ldots G_g$, g groupes, $\mathbf{z} = (z_1, z_2, \ldots z_q)$, q variables qualitatives et $\mathbf{y} = (y_1, y_2, \ldots y_p)$, p variables continues. La distribution conditionnelle de y dans le groupe G_i et pour $\mathbf{z} = \mathbf{a}$ fixé est gaussienne avec un vecteur moyenne $\mathbf{m}_{a,i}$ et une matrice de variance Σ définies par les relations (1) et (2) :

$$\mathbf{m}_{a,i} = E(\mathbf{y} / G_i, \mathbf{z} = \mathbf{a}) = \mathbf{w}(\mathbf{a}, i)\,\mathbf{b} \qquad (1)$$

$$\Sigma = V(\mathbf{y} / G_i, \mathbf{z} = \mathbf{a}) \qquad (2)$$

où $\mathbf{w}(\mathbf{a},i)$ est la matrice d'incidence associé à un modèle d'analyse de variance multivariable où les facteurs sont les variables qualitatives et le facteur groupe, et où $\mathbf{b}$ est un vecteur de paramètres.

Par exemple, si q=2 et p=1, on obtient un modèle d'analyse de variance à 3 facteurs:

$$E(\mathbf{y} / G_i, z_1 = l, z_2 = m) = m + a_i + b_l + c_m + (ab)_{il} + \ldots \text{etc.}$$

$$V(\mathbf{y} / G_i, z_1 = l, z_2 = m) = s^2$$

Il faut noter que, comme c'est le cas en analyse de variance multivariable, on considère que le modèle est le même pour toutes les variables continues (seuls changent les paramètres). Les interactions entre le facteur population et les facteurs associés à $\mathbf{z}$ permettent de modéliser les renversements. Les termes importants du modèle pour la discrimination sont ceux où apparaît l'indice i .

Pour compléter le modèle sur $(\mathbf{y}, \mathbf{z})$ dans chaque population G_i on considère un modèle loglinéaire sur la probabilité $\mathrm{Prob}(\mathbf{z} = \mathbf{a} / G_i)$:

$$\ln(\mathrm{Prob}(\mathbf{z} = \mathbf{a} / G_i)) = \mathbf{v}(\mathbf{a}, i)\,\mathbf{h}$$

où $\mathbf{v}(\mathbf{a},i)$ est le vecteur d'incidence associé à un modèle d'analyse de la variance et $\mathbf{h}$ est le vecteur des paramètres

Par exemple, si q = 2 on a :

$$\ln (\mathrm{Prob} (z_1 = l , z_2 = m \ / G_i)) \quad = \quad m + a_i + b_l + c_m + (ab)_{il} + \ldots\ldots \mathrm{etc.}$$

Finalement, la loi de (**y**, **z**) dans G_i est obtenue en faisant le produit des 2 lois précédentes :

$$g(\mathbf{y},\mathbf{z}) \quad = \quad (2\pi)^{-p/2} \det (\Sigma)^{-1/2} \exp \ [\ -1/2 \, (\mathbf{y} - \mathbf{w} (\mathbf{z}, i) \, \mathbf{b})' \, \Sigma^{-1} (\mathbf{y} - \mathbf{w} (\mathbf{z}, i) \, \mathbf{b}) \ + \ \mathbf{v} (\mathbf{z}, i) \, \mathbf{h} \,]$$

Le logarithme de la vraisemblance pour une unité pour laquelle on a obtenu (**y**, **z**) est donc dans le groupe G_i :

$$L_i \ = \ -1/2 \, (\mathbf{y} - \mathbf{w} (\mathbf{z}, i) \, \mathbf{b})' \, \Sigma^{-1} (\mathbf{y} - \mathbf{w} (\mathbf{z}, i) \, \mathbf{b}) \ + \ \mathbf{v} (\mathbf{z}, i) \, \mathbf{h} \ \ - \ p/2 \ln (2\pi) - 1/2 \ln (\det (\Sigma)^{-1/2} \ )$$

La règle optimale d'affectation d'une unité consiste donc à affecter l'unité au groupe G_i pour lequel la quantité :

$$(\mathbf{y} \ - \ 1/2 \, \mathbf{w} (\mathbf{z}, i) \, \mathbf{b})' \ \Sigma^{-1} \ \mathbf{w} (\mathbf{z}, i) \, \mathbf{b} \ + \ \mathbf{v} (\mathbf{z}, i) \, \mathbf{h}$$

est maximum.

Les probabilités a posteriori sont données par :

$$\mathrm{Prob} (G_i \ / \ \mathbf{y}, \mathbf{z} \ ) \quad = \quad p (G_i) \ \exp (L_i) \, / \, [\sum_k (p (G_k) \exp (L_k)) \ ]$$

où $p(G_i)$ est la probabilité à priori du groupe G_i. On vérifie que la probabilité a posteriori est de forme logistique. La différence essentielle entre le modèle Manova-loglinéaire et le modèle logistique réside dans les hypothèses de base, plus fortes dans le premier, et dans la procédure d'estimation des paramètres.

3. Estimation des paramètres et selection des variables.

Il faut estimer les paramètres **b** et **h** des 2 modèles. La procédure d'estimation du maximum de vraisemblance est bien connue aussi bien pour le modèle d'analyse de variance multivariable (Morrison 1974), que pour le modèle loglinéaire (Bishop et al. 1976). De nombreux logiciels existent pour les deux cas (BMDP,GENSTAT,SAS).

Il reste à choisir les variables discriminantes, et plus généralement, les termes à inclure dans chacun des 2 modèles. Il s'agit de trouver un modèle s'adaptant bien aux données d'une part, mais ne comportant pas trop de paramètres d'autre part. Si le modèle ne s'ajuste pas bien aux données, et surtout si l'on omet des paramètres importants pour discriminer les groupes, on obtient une mauvaise discrimination. Par contre, si on estime trop de paramètres, la fiabilité de ces estimations est mauvaise et la discrimination n'est pas meilleure.

Pour donner un exemple de cette situation, considérons le cas de modèles complets, c'est à dire comportant toutes les interactions de tous les ordres. On retrouve dans ce cas les estimations données par les analyses discriminantes linéaires effectuées dans chaque cellule définie par $\mathbf{z} = \mathbf{a}$. Or il est très fréquent qu'il n'y ait que quelques unités de l'échantillons pour lesquelles $\mathbf{z} = \mathbf{a}$, voire pas d'unité du tout. Il est clair que dans cette situation ces analyses discriminantes sont peu fiables pour la plupart. On est donc amené à regrouper des cellules correspondant à différentes valeurs de $\mathbf{z}$, pour obtenir de meilleures estimations. Ceci se fait dans le modèle manova-loglinéaire en supposant la nullité de certaines interactions. Le problème de la sélection consiste à rechercher les termes qu'il convient de garder dans le modèle. Il consiste également à sélectionner les variables discriminantes continues.

Daudin (1986) a proposé une méthode de sélection basé sur le critère d'Akaiké. Le nombre de modèles du type Manova-loglinéaire que l'on peut ajuster sur un ensemble de données est très important, aussi il est généralement impossible d' obtenir le modèle optimal. On peut, par contre, utiliser des méthodes sous optimales, telles qu'une procédure d'élimination progressive des termes du modèle à partir d'un modèle de base. Il s'agit d'éliminer de l'analyse

- les variables continues et qualitatives qui n'ont pas de pouvoir discriminant
- les termes du modèle que l'on peut supposer nuls dans le modèle loglinéaire ou dans le modèle d'analyse multivariable.

Certaines variables continues peuvent n'apporter de discrimination qu'en présence d'une variable qualitative particulière comme on peut le remarquer sur l'exemple suivant où l'on considère un modèle d'analyse de variance univariable :

$$m_{ij} = m + a_i + b_j + (ab)_{ij}$$

où i est l'indice relatif au facteur population, et j celui du facteur associé à une variable qualitative. Si

$a_i = 0 \quad (i = 1, g)$ et $(ab)_{ij} \neq 0$, y ne peut apporter aucune discrimination directe, mais par contre elle possède un pouvoir discriminant dans chaque cellule engendrée par la variable qualitative. La procédure suivante permet de sélectionner en même temps les variables discriminantes continues et les termes discriminants du modèle d'analyse de variance.

3.1 Selection des variables continues.

Les paramètres **b** du modèle d'analyse de la variance peuvent être divisés en 2 parties : $[\mathbf{b}_1, \mathbf{b}_2]$ où $\mathbf{b}_1$ contient les paramètres liés au facteur groupe et $\mathbf{b}_2$ les autres termes. Ces deux sous ensembles de paramètres jouent des rôles différents dans l'analyse discriminante: alors que $\mathbf{b}_1$ représente les différences entre les moyennes des variables entre les groupes, $\mathbf{b}_2$ sert seulement à déterminer les moyennes dans chaque cellule définie par les variables discriminantes qualitatives. L'importance de ce dernier terme réside seulement dans la réduction de Σ qu'il apporte. On note AIC ($\mathbf{b}_1, \mathbf{b}_2$, S) , la valeur du critère d'Akaiké pour le modèle d'analyse de variance sur les variables continues appartenant à un sous-ensemble S des p variables:

$$\mathrm{AIC}(\mathbf{b}_1, \mathbf{b}_2, S) = -(1/2)\,[\,n \ln(\det(\Sigma)) + ns\ln(2\pi) + s] - sk_1$$

où n est la taille de l'échantillon , s = card(S) , k_1 est le nombre de paramètres linéairement indépendants du modèle d'analyse de variance sur une variable.

Les valeurs de AIC ne sont pas comparables pour deux ensembles de variables S et S' à cause du terme ln (det (Σ)). C'est pourquoi on utilise une modification du critère d'Akaiké , DAIC, défini par :

$$\mathrm{DAIC}(\mathbf{b}_1, \mathbf{b}_2, S) = \mathrm{AIC}(\mathbf{b}_1, \mathbf{b}_2, S) - \mathrm{AIC}(\mathbf{0}, \mathbf{b}_2, S)$$

On peut considérer DAIC comme l'accroissement du critère d'Akaiké provoqué par la prise en compte du facteur population dans le modèle d'analyse de la variance multivariable sur l'ensemble S.
La paire optimale (S_{opt} , $\mathbf{b}_{1opt}$) est alors définie par la relation:

$$\mathrm{DAIC}(\mathbf{b}_{1opt}, \mathbf{b}_2, S_{opt}) = \sup_{S,\,\mathbf{b}_1} [\,\mathrm{DAIC}(\mathbf{b}_1, \mathbf{b}_2, S)\,]$$

3.2 Choix de b2 et h

Après avoir déterminé (S_{opt}, $\mathbf{b}_{1opt}$) il reste à déterminer $\mathbf{b}_2$ et **h**. Le critère d'Akaiké s'applique sans difficulté pour les termes du modèle d'analyse de la variance multivariable, car le nombre de variables ne change pas dans ce cas. Il suffit de déterminer $\mathbf{b}_2$ qui maximise AIC ($\mathbf{b}_{1opt}$, $\mathbf{b}_2$, S_{opt})

Il en est de même pour le modèle loglinéaire pour lequel Sakamoto(1982) a décrit l'utilisation de ce critère.

4 Description du logiciel ADM.

4.1 Caractéristiques générale de ADM.

Le programme ADM est composé de trois parties principales : analyse de variance, modèle loglinéaire, et calculs finaux (calcul des probabilités a posteriori, table de classement). Cette troisième partie est optionnelle.

4.1.1 Analyse de variance.

4.1.1.1 Paramètres en entrée:

- nombre de variables continues par individu
- nombre d'individus
- nombre de variables qualitatives(ou "facteurs") par individu.
- numéro du facteur population parmi les facteurs.
- noms des variables continues.
- noms des facteurs.
- modèle d'analyse de variance.
- pour chaque facteur, le nombre de niveaux de ce facteur (i.e. le nombre entier, n, tel que les valeurs possibles de ce facteur appartiennent à l'intervalle [1,n].
- le format de lecture des données (format Fortran)

4.1.1.2 Résultats

a) Pour le modèle défini en paramètre (appelé modèle fourni):

- matrice des paramètres par variable.
- matrice des résidus.
- nombre de degrés de liberté du modèle
- nombre de degrés de liberté de l'erreur

b) Pour le modèle construit par l'élimination du facteur population dans le modèle fourni (par exemple si le modèle fourni est 1+2+3+1.2+2.3, et si le facteur population est le facteur 1, le modèle construit par l'élimination du facteur population est 2+3+2.3), on réitère les calculs et édite les sorties du type précédent (§a).

c) Critère d'Akaike pour le modèle fourni, et différence entre cette dernière valeur et celle du même critère, calculé pour le modèle défini en b). (cf DAIC défini au § 3.1)

4.1.2 Modèle loglinéaire.

4.1.2.1 Paramètres en entrée.

- modèle (il peut-être différent de celui de l'analyse de la variance).
- nombre maximum d'itérations dans l'algorithme de recherche des estimateurs du maximum de vraisemblance.
- Précision minimale demandée (écart-relatif maximal entre deux itérations successives)
- constante additionnelle; c'est une valeur (typiquement 0.1 ou 0.5) ajoutée à chaque case de la table de contingence, si une au moins des cases des tables marginales de la table de contingence associées au modèle, est nulle.

4.1.2.2 Résultats.

- Nombre d'itérations utilisées pour atteindre la précision demandée.
- Statistique du rapport de vraisemblance.
- Valeur du critère d'Akaike (cf §3.2)

4.1.3 Calculs finaux.

Ces calculs peuvent être omis suivant la valeur (0 ou 1) du paramètre qui les déclenche ou non.

4.1.3.1 Paramètres en entrée.

- Seuil pour la probabilité de classement, si on procède à un classement avec doute. (si la probabilité a posteriori du groupe le plus probable est inférieure à ce seuil, on ne classe pas l'unité). Ce seuil doit être compris entre 0 et 1. (les valeurs typiques sont .6, .8 et .9)
- Seuil pour le rejet; c'est la valeur de la probabilité a posteriori en deçà de laquelle on rejette le groupe pour l'unité considérée. Ce seuil doit être compris entre 0 et 1.(valeurs typiques: 0.05, 0.1)
- probabilités a priori; l'option standard est l'équiprobabilité. Dans le cas contraire,il faut donner autant de probabilité qu'il y a de groupes, dont la somme vaut 1.
- classement avec coût; si cette option est utilisée, l'utilisateur doit donner la matrice des coûts de mauvais classement.

4.1.3.2 Résultats.

- pour chaque individu : numéro de l'individu,numéro du groupe d'origine, numéro du groupe d'affectation, probabilité a posteriori de chaque groupe auquelle est éventuellement ajoutée une " * " si le groupe est rejeté .(cf 4.1.3.1 seuil de rejet)
- tables de classement par resubstitution; (classement systématique, et éventuellement classement avec doute et classement par la règle de coût minimal)

4.1.3 Structure du programme source.

Le source est très commenté, style littéraire délirant. Pour chaque procédure, le source est découpé comme suit:

- Déclaration systématique des paramètres et des variables.
- En commentaire, explication détaillée de
  + paragraphe importe : variables importées.
  + paragraphe exporte : variables exportées.
  + paragraphe local : variables locales.

Des variables peuvent être à la fois importées et exportées.

- En commentaire algorithme détaillé de la procédure en pseudo Pascal.
- Programme Fortran commenté.

ADM est écrit en Fortran 77, norme Modulad, en caractères minuscules sauf les variables.

Un programme annexe, SADM, accompagne ADM. SADM permet la saisie, avec contrôle de validité, en conversationnel, du fichier paramètre utilisé par ADM. On peut évidemment s'en passer et créer ce fichier paramètre avec un éditeur de texte quelconque.

4.2 Exemple de résultats donnés par ADM.

Les données sont les suivantes:

Facteur population	Variable qualitative	Variable continue 1	Variable continue 2
1	1	5	2
1	1	6	3
1	1	7	3
1	2	8	2
1	2	9	3
1	2	10	3
2	1	10	3
2	1	9	4
2	1	8	2
2	2	7	3
2	2	6	3
2	2	5	2

Il s'agit donc d'un problème de discrimination de 2 populations avec une variable qualitative et deux variables continues.

Les résultats donnés par ADM se présentent de la façon suivante :

1) Les résulats de l'analyse de variance multivariable avec le modèle donné par l'utilisateur. On obtient l'estimation des paramètres du modèle pour chaque variable (la contrainte sur les paramètres consiste à annuler le dernier effet, interaction etc....). La méthode d'inversion est le sweeping (voir Dempster 1969). On obtient également la matrice de variance résiduelle ("matrice résidus").

2) Les résultats de l'analyse de variance multivariable avec le même modèle que précedemment sauf tous les termes concernant le facteur population. Ceci permet d'obtenir le critère DAIC (cf §3). Le pouvoir discriminant est d'autant plus élevé que DAIC est élevé. On recherchera donc à obtenir le modèle maximisant DAIC.

3) Les résultats du modèle loglinéaire . L'estimation des paramètres du modèle est faite par l'algorithme "Iterative Proportionnal Fitting" (Bischop et al. 1976). Dans le cas où une ou plusieurs marginales sont nulles, il est nécessaire d'ajouter une constante à toutes les cases de la table de contingence multidimensionnelle. La valeur du critère d'Akaiké est donnée pour chaque modèle (Sakamoto,1982).

4) Le calcul de la vraisemblance de chaque observation dans chaque population, des probabilités à postériori, de l'affectation des observations et de la table de classement. Il est possible d'utiliser les probabilités à postériori pour (cf R. Tomassone et al 1988) :

faire un classement avec doute.
donner les populations rejetées pour chaque observation
faire un classement minimisant le coût moyen de mauvais classement. Dans ce cas l'utilisateur doit donner la matrice des coûts de classement.

```
*                      ANALYSE DISCRIMINANTE MIXTE                                   *
* J.J. DAUDIN ET M. SOUKAL : INSTITUT NATIONAL AGRONOMIQUE DE PARIS-GRIGNON *
******************************************************************************

**************************************************
  PARAMETRES ANALYSE DE VARIANCE
      NOMBRE DE VARIABLES        2
      NOMBRE D'INDIVIDUS        12
      NOMBRE DE FACTEURS         2
      FACTEUR POPULATION :
                   NUMERO        1
                   NOM       POP
      NOMS DES VARIABLES
   V1         V2
      NOMS DES FACTEURS
   POP        FAC
      MODELE
      1+2+1.2
      NIVEAU DES FACTEURS
     2          2
      FORMAT DE LECTURE DES INDIVIDUS
      (2I2,2F2.0)

**************************************************


                  ANALYSE DE VARIANCE
                 *********************


*****ANALYSE AVEC LE MODELE COMPLET*****

       MATRICE DES PARAMETRES PAR VARIABLE
              V1             V2
1     1     0.60000E+01   0.26667E+01
2     2     0.30000E+01   0.71526E-06
3     3     0.30000E+01   0.33333E+00
4     4    -0.60000E+01  -0.33333E+00

       MATRICE RESIDUS
              V1             V2
 V1         0.10000E+01   0.50000E+00
 V2         0.50000E+     0.50000E+00

                DEGRES DE LIBERTE DU MODELE :       3
                DEGRES DE LIBERTE DE L ERREUR :     8

*****ANALYSE AVEC LE MODELE SANS LES TERMES INCLUANT LE FACTEUR POPULATION****

       MATRICE DES PARAMETRES PAR VARIABLE
              V1             V2
1     1     0.75000E+01   0.26667E+01
2     2     0.10729E-05   0.16667E+00

       MATRICE RESIDUS
              V1             V2
 V1         0.35000E+01   0.55000E+00
 V2         0.55000E+00   0.41667E+0


                      DEGRES DE LIBERTE DU MODELE :       1
                      DEGRES DE LIBERTE DE L'ERREUR :    10


  CRITERE D' AKAIKE POUR LE MODELE FOURNI          -67.473
  DIFFERENCE DU CRITERE D' AKAIKE POUR LE MODELE FOURNI
  ET CE MEME MODELE SANS LE FACTEUR POPULATION     10.373

                    ANALYSE LOG-LINEAIRE
                   *********************


    ************************************************
      PARAMETRES ETUDE LOG-LINEAIRE
           MODELE
           1+2+1.2
           NOMBRE MAXIMUM DE BOUCLES POUR
           LA CONVERGENCE                          40
           PRECISION DU CALCUL                  0.010
           CONSTANTE ADDITIONNELLE DE LA
           TABLE DE CONTINGENCE SI UNE CASE
           DES TABLES MARGINALES EST NULLE      0.100

    ************************************************

      **** MODELE HIERARCHIQUE COMPLET OBTENU
           A PARTIR DU MODELE LOG-LINEAIRE ****
               1+2+1.2

  CONVERGENCE EN    1 ITERATIONS
  KHI2 DU RAPPORT DE VRAISEMBLANCE     0.000
  VALEUR DU CRITERE D' AKAIKE         -8.000
```

```
                    CALCULS FINAUX
                   ***************

  ***************************************************
    PARAMETRES CALCULS FINAUX
        PROBABILITE MINIMUM DE CLASSEMENT
        AVEC DOUTE                                      0.600
        L'ETOILE QUI PRECEDE LA PROBABILITE
        A POSTERIORI SIGNIFIE QUE LA CLASSE
        EST REJETEE AU SEUIL                            0.100
        POUR TOUS LES NIVEAUX, PROBABILITE
        A PRIORI EQUIVALENTE                            0.500

  ***************************************************


                GROUPE            PROBABILITE A POSTERIORI POUR
INDIVIDU   ORIGINE  AFFECTATION   CHAQUE NIVEAU DU FACTEUR POPULATI
    1         1          1          1.000   * 0.000
    2         1          1          1.000   * 0.000
    3         1          1          0.969   * 0.031
    4         1          1          0.999   * 0.001
    5         1          1          0.999   * 0.001
    6         1          1          1.000   * 0.000
    7         2          2        * 0.000     1.000
    8         2          2        * 0.072     0.928
    9         2          2        * 0.001     0.999
   10         2          2        * 0.007     0.993
   11         2          2        * 0.000     1.000
   12         2          2        * 0.000     1.000


                               TABLE DE CLASSEMENT
              CLASS.      CLASS  EFFECTUE
              REEL

                         1    2
                      ---------
                  1:     6    0
                  2:     0    6


                              TABLE DE CLASSEMENT
              CLASS.  CLASSES   CLASS. AVEC DOUTE
              REEL    DOUTEUX

                              1    2
                           ---------
                  1:     0:   6    0
                  2:     0:   0    6
```

Si on utilise le même modèle sans la deuxième variable continue, on obtient DAIC=11.03, ce qui est meilleur que la valeur de ce critère pour l'analyse discriminante avec les deux variables continues (DAIC=10.37). Par contre, si on supprime l'interaction dans le modèle d'analyse de la variance, on obtient DAIC=-3.26, et une très mauvaise discrimination. Par curiosité, le lecteur est invité à utiliser une analyse discriminante linéaire classique sur ces données. Le résultat est une très mauvaise discrimination, car les données présentent un" renversement", qui n'est pas pris en compte dans l'analyse discriminante classique.

BIBLIOGRAPHIE

Akaiké , H . (1973) Information theory and an extension of the maximum likelihood principle. 2nd Internat.Symposium on information theory. Ed. Petrov B.N. Budapest.

BISHOP, Y.M.M.,FIENBERG,S.E.,HOLLAND,P.W.(1976) *Discrete multivariate analysis* .MIT Press.

DAUDIN,J.J.(1987) Selection of variables in Mixed-Variable Discriminant Analysis. Biometrics 38, 191-200.

DEMPSTER,A.P.(1969) *Elements of Continuous Multivariate Analysis* .Addison-Wesley Publishing Company.

KNOKE , J . D. (1982) Discriminant analysis with discrete and continuous variables . Biometrics 38 , 191-200.

KRZANOWSKI , W . J. (1975) Discrimination and classification using both binary and continuous variables . J. Amer. Statist. Assoc. 70 , 782-90

KRZANOWSKI , W . J . (1977) The performance of Fischer's linear discrimination function under non optimal conditions . Technometrics 19 , 191-200.

KRZANOWSKI ,W . J . (1980) Mixtures of continuous and categorical variables in discriminant analysis . Biometrics 36 , 486-499

KRZANOWSKI , W . J . (1983) Stepwise location model choice in mixed variable discrimination . Appl . Statist. 32 , 3 , 260-66

MOORE , D. H . (1973) Evaluation of five discrimination procedures for binary variables . J.Amer. Statist.Assoc. 68 , 399-404.

OLKIN, I. TATE,R.F. (1961) Multivariate correlation models with mixed discrete and continuous variables . Annals of Mathematical Statistics 32 , 448-465 .

RAO , C . R . (1973) *Linear statistical inference and its applications* .Wiley . New-York.

SAKAMOTO , Y . (1982) Efficient use of Akaike's information criterion for model selection in high dimensional contingency table analysis . Metron 15 , 1-2 , 257-276 .

TOMASSONE R., DANZART M., DAUDIN J.J., MASSON J.P.(1988) *Discrimination et Classement*. Masson Ed.Paris.

VLACHONIKOLIS , J.G. and MARRIOT, F.H.C. (1982) Discrimination with mixed binary and continuous data. Appl. Statist. 31 , 23-31.