L'ANALYSE DISCRIMINANTE DÉCISIONNELLE : PRINCIPES ET APPLICATION

R. PALM*

RÉSUMÉ

Les principes de l'analyse discriminante sont décrits et illustrés par l'exemple relatif au classement de trois espèces d'iris proposé par FISHER [1936].

SUMMARY

The principles of discriminant analysis are described and illustrated by the iris data set published by FISHER [1936].

1. INTRODUCTION

L'objectif de l'*analyse discriminante décisionnelle*[1] est de définir une règle permettant de classer un individu dans un groupe particulier, parmi les g groupes possibles. Cette affectation de l'individu à un groupe donné se fait sur base de p caractéristiques, c'est-à-dire de p variables, observées sur cet individu et la règle de classement est établie en fonction de ces mêmes p caractéristiques, observées sur des échantillons provenant des g groupes ou populations.

L'analyse discriminante décisionnelle vise donc à résoudre les problèmes de classement. Elle se différencie des méthodes de *classification*[2], dans la mesure où les classes sont définies au départ de l'analyse. Bien qu'il y ait des points communs, comme nous le verrons par la suite, elle se différencie aussi de l'*analyse factorielle discriminante*, appelée encore *analyse des variables canoniques*[3], dont l'objectif est de décrire et de comparer des échantillons ou des populations, et qui constitue un complément de l'analyse de la variance multivariée.

*Chargé de cours associé à la Faculté universitaire des Sciences agronomiques de Gembloux.

1. En anglais : *discriminant analysis, predictive discriminant analysis, classification procedure.*

2. En anglais : *cluster analysis, classification.*

3. En anglais : *factorial discriminant analysis, descriptive discriminant analysis, canonical variate analysis.*

Un exemple fréquemment étudié et initialement proposé par FISHER [1936] concerne des observations relatives à trois espèces d'iris (*Iris setosa*, *Iris versicolor* et *Iris virginica*). Les données complètes se trouvent, notamment, dans HAND *et al.* [1994], KENDALL *et al.* [1983] et dans la documentation SAS [SAS, 1989]. Quatre variables (longueur et largeur des pétales et des sépales) ont été observées sur 50 fleurs de chacune des espèces. A partir de ces données, on se propose de définir une règle permettant d'affecter à une espèce déterminée un iris dont on ne connaîtrait pas l'espèce. Nous reprendrons cet exemple classique pour illustrer les principes de l'analyse discriminante.

Un autre exemple d'utilisation de l'analyse discriminante pour résoudre un problème de taxonomie est donné par DEBOUCHE *et al.* [1979]. Il s'agit de différencier deux espèces de poissons du genre *Clarias* dans le bassin du Sénégal (*Clarias lazera* et *Clarias anguillaris*). La différenciation de ces deux espèces nécessitait le prélèvement de l'arc branchial, donc le sacrifice de l'individu. L'objectif de l'étude de ces auteurs était d'établir une fonction discriminante basée sur des mesures faciles à réaliser de manière à déterminer l'animal sans le sacrifier.

Des exemples d'application dans le domaine médical sont donnés, notamment, par ALBERT et HARRIS [1987] et par SAPORTA [1990]. Il s'agit, par exemple, de poser un diagnostic sur la base d'analyses biochimiques ou de réaliser un pronostic en fonction de diverses caractéristiques observées sur les patients.

L'analyse discriminante trouve aussi son application dans les problèmes d'interprétation d'images provenant de satellites. On parle dans ce cas de *classification supervisée*[4]. Le problème peut être schématisé de la façon suivante. En fonction de données de terrain, on connaît, par exemple, l'utilisation agricole d'une série de parcelles témoins. On connaît aussi, à partir des images, la réponse spectrale de ces différentes parcelles et on s'efforce de définir une règle permettant d'affecter des zones non observées sur le terrain à l'une ou l'autre de ces utilisations. Un exemple est donné dans la documentation SAS [1989]. L'expression classification supervisée, utilisée dans ce contexte, s'oppose à la notion de *classification non supervisée*[5], qui a pour objet la définition de groupes ou de classes.

Des références pour de nombreux autres exemples sont données par HUBERTY [1994].

Dans cette note, nous détaillerons une technique particulière de classement, appelée analyse discriminante linéaire, qui s'applique aux situations où les populations sont supposées normales et de même matrice de variances et covariances. Nous examinerons d'abord le cas le plus simple, où on dispose de deux populations univariées (paragraphe 2), puis, nous passerons au cas de deux populations bivariées (paragraphe 3) et, enfin, au cas général de g populations multivariées (paragraphe 4). Nous examinerons ensuite brièvement d'autres méthodes d'analyse discriminante (paragraphe 5) et nous étendrons les méthodes décrites au cas où les probabilités d'appartenance à un groupe donné sont inégales (paragraphe 6). Nous présenterons alors les méthodes de validation des résultats (para-

4. En anglais : *supervised classification.*
5. En anglais : *unsupervised training.*

graphe 7) et, enfin, nous clôturerons par quelques recommandations pratiques (paragraphe 8).

2. DEUX POPULATIONS CARACTÉRISÉES PAR UNE SEULE VARIABLE

2.1. Règle de classement

Considérons tout d'abord le cas de deux populations normales à une dimension ($g = 2$ et $p = 1$) de même variance et supposons qu'on ait prélevé, dans chacune de ces deux populations et de manière indépendante, un échantillon aléatoire et simple.

Pour être plus concret, prenons le cas des observations relatives à la longueur des pétales des 50 *Iris versicolor* et des 50 *Iris virginica* auxquels nous avons fait allusion dans l'introduction. Pour ces deux espèces, les moyennes et les écarts-types (en cm) sont donnés dans le tableau 1. Les données de départ sont présentées avec une décimale. Toutefois, pour assurer une meilleure concordance des valeurs numériques obtenues manuellement avec celles résultant de l'utilisation des logiciels, nous avons volontairement laissé, dans le tableau 1, des décimales non significatives.

Tableau 1. Moyennes et écarts-types estimés des longueurs des pétales des 50 individus d'*Iris versicolor* et d'*Iris virginica* (en cm).

Espèces	Moyennes	Ecarts-types
Iris versicolor	4,260	0,470
Iris virginica	5,552	0,552

L'hypothèse de normalité des deux distributions est acceptée sur la base du test de SHAPIRO-WILK, proposé par le logiciel SAS [SAS, 1989]. D'autre part, la valeur F_{obs} relative au test d'égalité des deux variances vaut 1,38. Elle est non significative et conduit donc à l'acceptation de l'hypothèse d'égalité de la dispersion des longueurs des pétales au sein des deux espèces.

Une estimation globale de la variance commune des populations est obtenue en additionnant les sommes des carrés des écarts et les degrés de liberté relatifs à chacun des deux échantillons. On obtient ainsi :

$$\hat{\sigma}^2 = \frac{(49)(0,470^2) + (49)(0,552^2)}{49 + 49} = 0,263\,,$$

soit un écart-type estimé de 0,513.

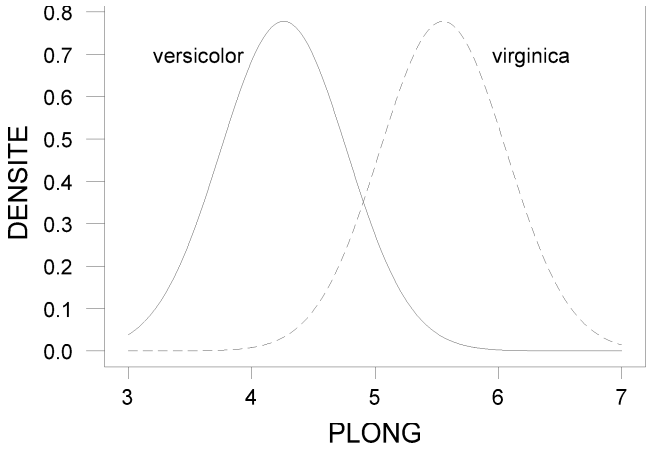

Figure 1. Représentation des populations des longueurs de pétales (PLONG) pour *Iris versicolor* et *Iris virginica*.

Les conditions d'application énoncées ci-dessus sont bien remplies, si du moins on considère les deux échantillons comme aléatoires, simples et indépendants.

La figure 1 donne une représentation graphique des deux populations normales de moyennes $\hat{m}_1 = 4,260$, pour *Iris versicolor*, et $\hat{m}_2 = 5,552$, pour *Iris virginica*, et de même écart-type $\hat{\sigma} = 0,513$.

En supposant, *a priori*, qu'un individu ait la même probabilité de provenir de l'une ou l'autre des deux populations et en désignant par x_i la longueur des pétales d'un individu i d'origine inconnue, il est naturel de définir la règle de classement suivante :

- si $x_i < x_0$, on classe l'individu dans la population *Iris versicolor*,
- si $x_i > x_0$, on classe, par contre, l'individu dans la population *Iris virginica*,

la valeur x_0 étant égale à :

$$x_0 = (\hat{m}_1 + \hat{m}_2)/2 = 4,91\,,$$

Si x_i est égal à x_0, on classe indifféremment l'individu dans l'une ou l'autre population.

Une solution tout à fait équivalente consiste à calculer les carrés des deux distances réduites suivantes :

$$d_{1i}^2 = \left(\frac{x_i - \hat{m}_1}{\hat{\sigma}}\right)^2 \quad \text{et} \quad d_{2i}^2 = \left(\frac{x_i - \hat{m}_2}{\hat{\sigma}}\right)^2 ,$$

et à classer l'individu i dans l'espèce *versicolor* si $d_{1i}^2 < d_{2i}^2$ et dans l'espèce *virginica* si $d_{1i}^2 > d_{2i}^2$.

Cela revient, en fait, à calculer la distance de l'individu i par rapport aux centres de gravité estimés des deux populations et à affecter cet individu à la population dont il est le plus proche du centre de gravité.

Une troisième solution est basée sur l'examen des fonctions de densité de probabilité des deux populations pour l'individu i :

$$f_1(x_i) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \hat{m}_1}{\hat{\sigma}}\right)^2\right]$$

$$\text{et} \quad f_2(x_i) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \hat{m}_2}{\hat{\sigma}}\right)^2\right]$$

et à classer l'individu dans la population pour laquelle la densité de probabilité est la plus grande : on classe l'individu i dans l'espèce *versicolor* si $f_1(x_i) > f_2(x_i)$ et on classe l'individu i dans l'espèce *virginica* si $f_1(x_i) < f_2(x_i)$. Cette solution est identique à la solution précédente, car la densité de probabilité est une fonction décroissante du carré de la distance réduite définie ci-dessus.

On notera que les fonctions de densité de probabilité dont il est question sont, en réalité, des fonctions estimées puisque les moyennes et l'écart-type commun sont, en pratique, inconnus.

A titre d'illustration, considérons un iris dont on sait qu'il appartient soit à l'espèce *versicolor*, soit à l'espèce *virginica* et dont la longueur des pétales est égale à 4,7. Il serait affecté à l'espèce *versicolor* car :

$$x_i = 4,7 < x_0 = 4,91\,,$$
$$d_{1i}^2 = 0,74 < d_{2i}^2 = 2,76\,,$$
$$f_1(x_i) = 0,54 > f_2(x_i) = 0,20\,.$$

Une quatrième solution peut encore être envisagée pour le classement d'un individu. Cette solution, équivalente aux trois précédentes, se base sur la probabilité d'appartenance *a posteriori* et est présentée au paragraphe suivant.

2.2. Probabilité d'appartenance *a posteriori*

Etant donnée l'observation x_i, on peut se demander quelle est la probabilité que l'individu i appartienne à l'une ou l'autre des deux populations. L'application du théorème de Bayes [Dagnelie, 1998] permet d'établir ces deux probabilités :

$$\mathrm{P}(\mathrm{A1}|x_i) = f_1(x_i) \big/ \left[f_1(x_i) + f_2(x_i)\right]$$

$$\text{et} \quad \mathrm{P}(\mathrm{A2}|x_i) = f_2(x_i) \big/ \left[f_1(x_i) + f_2(x_i)\right]\,,$$

A1 et A2 désignant les événements "l'individu i appartient à la population 1" et "l'individu i appartient à la population 2".

Pour deux populations normales de même variance, elles peuvent encore s'écrire :

$$\mathrm{P}(\mathrm{A1}|x_i) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x_i - \hat{m}_1}{\hat{\sigma}}\right)^2\right]}{\sum_{k=1}^{2}\exp\left[-\frac{1}{2}\left(\frac{x_i - \hat{m}_k}{\hat{\sigma}}\right)^2\right]} = \frac{\exp\left(-\frac{1}{2}d_{1i}^2\right)}{\sum_{k=1}^{2}\exp\left(-\frac{1}{2}d_{ki}^2\right)}$$

$$\mathrm{P}(\mathrm{A2}|x_i) = \frac{\exp\left[-\frac{1}{2}\left(\frac{x_i - \hat{m}_2}{\hat{\sigma}}\right)^2\right]}{\sum_{k=1}^{2}\exp\left[-\frac{1}{2}\left(\frac{x_i - \hat{m}_k}{\hat{\sigma}}\right)^2\right]} = \frac{\exp\left(-\frac{1}{2}d_{2i}^2\right)}{\sum_{k=1}^{2}\exp\left(-\frac{1}{2}d_{ki}^2\right)}\,.$$

Ces probabilités d'appartenance *a posteriori*, c'est-à-dire quand on dispose de l'observation x_i, permettent de classer un individu, en l'affectant à la population pour laquelle la probabilité d'appartenance *a posteriori* est la plus grande. Il s'agit, ici aussi, en réalité de probabilités estimées et non de probabilités théoriques, puisqu'elles sont calculées à partir de densités de probabilité estimées.

La probabilité d'appartenance *a posteriori* à une population donnée étant une fonction décroissante de la distance d_{ki}, on obtient évidemment le même résultat de classement que précédemment. Rappelons que toutes les règles d'affectation supposent qu'*a priori* un individu a la même probabilité de provenir de chacune des deux popultions. Nous examinerons, au paragraphe 6, le cas où ces probabilités sont inégales.

Pour l'exemple ci-dessus, on a :

$$\mathrm{P}(\textit{versicolor}|x_i = 4,7) = 0,54/(0,54 + 0,20) = 0,73$$

et
$$\mathrm{P}(\textit{virginica}|x_i = 4,7) = 0,20/(0,54 + 0,20) = 0,27\,.$$

2.3. Probabilité de classement erroné *a priori*

Dans les conditions que nous avons envisagées dans les paragraphes précédents (deux populations normales de même variance, m_1 inférieure à m_2 et probabilités *a priori* égales), la probabilité de classer, dans la population 2, un individu i pris au hasard alors qu'il appartient à la population 1 est égale à :

$$\begin{aligned}\mathrm{P}(\mathrm{C2}|\mathrm{A1}) = \mathrm{P}(x_i > x_0) &= 1 - \Phi\left(\frac{x_0 - m_1}{\sigma}\right)\\ &= 1 - \Phi\left(\frac{m_2 - m_1}{2\,\sigma}\right) = 1 - \Phi(\Delta_{12}/2)\,.\end{aligned}$$

Dans cette relation, Δ_{12} représente l'écart réduit, pris en valeur absolue, qui existe entre les moyennes des deux populations :

$$\Delta_{12} = \frac{|m_1 - m_2|}{\sigma} .$$

La probabilité définie ci-dessus est la probabilité conditionnelle de l'événement C2, sous la condition de la réalisation de l'événement A1. C2 représente l'événement "classer l'individu i dans la population 2" et A1 représente l'événement "l'individu i appartient à la population 1".

De même, la probabilité de classer l'individu dans la population 1, s'il appartient à la population 2, est égale à :

$$\mathrm{P}(\mathrm{C1}|\mathrm{A2}) = 1 - \Phi(\Delta_{12}/2) .$$

Les deux probabilités conditionnelles sont donc identiques et sont aussi égales à la probabilité de classement erroné d'un individu pris au hasard dans l'une des deux populations.

En effet, à partir de la définition des probabilités conditionnelles, on peut écrire :

$$\mathrm{P}(\mathrm{C2}|\mathrm{A1}) = \mathrm{P}(\mathrm{C2 \text{ et } A1})/\mathrm{P}(\mathrm{A1}) \quad \text{et} \quad \mathrm{P}(\mathrm{C1}|\mathrm{A2}) = \mathrm{P}(\mathrm{C1 \text{ et } A2})/\mathrm{P}(\mathrm{A2})$$

et la probabilité de classement erroné est égale à :

$$\begin{aligned} \mathrm{P}(\mathrm{C2 \ \text{ et } A1}) + \mathrm{P}(\mathrm{C1 \ \text{ et } A2}) &= \mathrm{P}(\mathrm{C2}|\mathrm{A1})\mathrm{P}(\mathrm{A1}) + \mathrm{P}(\mathrm{C1}|\mathrm{A2})\mathrm{P}(\mathrm{A2}) \\ &= \mathrm{P}(\mathrm{C2}|\mathrm{A1}) = \mathrm{P}(\mathrm{C1}|\mathrm{A2}) , \end{aligned}$$

si les probabilités d'appartenance aux deux populations, P(A1) et P(A2), sont égales.

Une estimation naturelle de cette probabilité est obtenue en remplaçant Δ_{12} par l'estimation suivante :

$$\hat{\Delta}_{12} = \frac{|\hat{m}_1 - \hat{m}_2|}{\hat{\sigma}} .$$

Pour les deux espèces d'iris considérées, on a :

$$\hat{\Delta}_{12} = 2,52 \quad \text{et} \quad 1 - \Phi(1,26) = 1 - 0,89 \approx 0,10 ,$$

et on peut donc s'attendre à un taux d'erreur de l'ordre de 10 %.

Nous verrons, au paragraphe 7.1, que l'estimation de Δ_{12} est biaisée et qu'une correction peut être apportée afin de réduire ce biais.

3. DEUX POPULATIONS CARACTÉRISÉES PAR DEUX VARIABLES

3.1. Règle de classement

On considère maintenant deux populations normales à deux dimensions ($g = 2$ et $p = 2$), de matrices de variances et covariances égales, dans lesquelles ont été prélevés deux échantillons aléatoires, simples et indépendants.

Soient σ_{x_1}, σ_{x_2} et ρ les écarts-types marginaux et la corrélation des deux variables x_1 et x_2 et soient m_{11} et m_{12} les moyennes des deux variables pour la population 1 et m_{21} et m_{22}, les moyennes des deux variables pour la population 2. Ces paramètres sont, en général, inconnus et doivent être estimés à partir des deux échantillons, comme nous le verrons dans l'exemple ci-dessous.

Pour un individu i à classer, caractérisé par le couple d'observations x_{1i} et x_{2i}, on peut calculer les densités de probabilité à deux dimensions de chacune des deux populations [DAGNELIE, 1998] :

$$f_1(x_{1i}, x_{2i}) = \frac{1}{2\pi\,\hat{\sigma}_{x_1}\,\hat{\sigma}_{x_2}\sqrt{(1-\hat{\rho}^2)}} \exp\left[-\frac{1}{2}d_{1i}^2\right]$$

et

$$f_2(x_{1i}, x_{2i}) = \frac{1}{2\pi\,\hat{\sigma}_{x_1}\,\hat{\sigma}_{x_2}\sqrt{(1-\hat{\rho}^2)}} \exp\left[-\frac{1}{2}d_{2i}^2\right]$$

avec :

$$d_{1i}^2 = \frac{1}{(1-\hat{\rho}^2)}\left[\left(\frac{x_{1i}-\hat{m}_{11}}{\hat{\sigma}_{x_1}}\right)^2 - 2\,\hat{\rho}\left(\frac{x_{1i}-\hat{m}_{11}}{\hat{\sigma}_{x_1}}\right)\left(\frac{x_{2i}-\hat{m}_{12}}{\hat{\sigma}_{x_2}}\right) + \left(\frac{x_{2i}-\hat{m}_{12}}{\hat{\sigma}_{x_2}}\right)^2\right]$$

et

$$d_{2i}^2 = \frac{1}{(1-\hat{\rho}^2)}\left[\left(\frac{x_{1i}-\hat{m}_{21}}{\hat{\sigma}_{x_1}}\right)^2 - 2\,\hat{\rho}\left(\frac{x_{1i}-\hat{m}_{21}}{\hat{\sigma}_{x_1}}\right)\left(\frac{x_{2i}-\hat{m}_{22}}{\hat{\sigma}_{x_2}}\right) + \left(\frac{x_{2i}-\hat{m}_{22}}{\hat{\sigma}_{x_2}}\right)^2\right]$$

et, comme dans le cas univarié, on affecte l'individu i à la population pour laquelle la densité de probabilité est la plus grande.

Pour simplifier l'écriture, on peut utiliser les notations matricielles et on a alors :

$$d_{1i}^2 = [\boldsymbol{x}_i - \hat{\boldsymbol{m}}_1]' \;\hat{\boldsymbol{\Sigma}}^{-1}\,[\boldsymbol{x}_i - \hat{\boldsymbol{m}}_1] \quad \text{et} \quad d_{2i}^2 = [\boldsymbol{x}_i - \hat{\boldsymbol{m}}_2]' \;\hat{\boldsymbol{\Sigma}}^{-1}\,[\boldsymbol{x}_i - \hat{\boldsymbol{m}}_2]$$

avec :

$$\boldsymbol{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix}, \quad \hat{\boldsymbol{m}}_1 = \begin{bmatrix} \hat{m}_{11} \\ \hat{m}_{12} \end{bmatrix}, \quad \hat{\boldsymbol{m}}_2 = \begin{bmatrix} \hat{m}_{21} \\ \hat{m}_{22} \end{bmatrix} \quad \text{et} \quad \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}_{x_1}^2 & \hat{\mu}_{11} \\ \hat{\mu}_{11} & \hat{\sigma}_{x_2}^2 \end{bmatrix}.$$

Dans ces relations, $\hat{\boldsymbol{\Sigma}}$ représente la matrice des variances et covariances estimées, les variances des deux variables étant sur la diagonale et la covariance, $\hat{\mu}_{11}$, étant l'élément hors diagonale.

Puisque les densités de probabilité sont des fonctions décroissantes des distances d_{ki}^2 définies ci-dessus, on peut aussi comparer directement les distances et classer l'individu i dans la population 1 si $d_{1i}^2 < d_{2i}^2$ et dans la population 2 si

Tableau 2. Paramètres des distributions relatives à la longueur des pétales et à la largeur des pétales pour *Iris versicolor* et *Iris virginica*.

Paramètres	*Iris versicolor*	*Iris virginica*
Longueurs moyennes	4,260	5,552
Largeurs moyennes	1,326	2,026
Matrices de variances et covariances	$\begin{bmatrix} 0,22082 & 0,07310 \\ 0,07310 & 0,03911 \end{bmatrix}$	$\begin{bmatrix} 0,30459 & 0,04882 \\ 0,04882 & 0,07543 \end{bmatrix}$

$d_{1i}^2 > d_{2i}^2$. Les distances d_{1i} et d_{2i} sont appelées distances au sens de MAHALANOBIS et leur interprétation est donnée au paragraphe suivant.

Comme dans le cas univarié également, on peut calculer les probabilités d'appartenance *a posteriori*. On retrouve d'ailleurs les mêmes formules que précédemment (paragraphe 2.2) :

$$\mathrm{P(A1}|\boldsymbol{x}_i) = \frac{\exp\left(-\frac{1}{2}d_{1i}^2\right)}{\sum_{k=1}^{2}\exp\left(-\frac{1}{2}d_{ki}^2\right)} \quad \text{et} \quad \mathrm{P(A2}|\boldsymbol{x}_i) = \frac{\exp\left(-\frac{1}{2}d_{2i}^2\right)}{\sum_{k=1}^{2}\exp\left(-\frac{1}{2}d_{ki}^2\right)},$$

et on classe alors l'individu dans la population pour laquelle la probabilité d'appartenance *a posteriori* est la plus grande.

Pour illustrer la règle de classement dans le cas de deux variables et deux populations, reprenons les deux espèces d'iris examinées au paragraphe précédent et ajoutons, comme seconde variable, la largeur des pétales. Nous supposons, sans les vérifier pour le moment, que les conditions d'application énoncées ci-dessus sont remplies. Nous discuterons, au paragraphe 8.1, l'importance et la manière de vérifier ces conditions.

Le tableau 2 reprend les paramètres des distributions relatives à la longueur et à la largeur des pétales pour les deux espèces. En désignant par $\hat{\mathbf{\Sigma}}_1$ et $\hat{\mathbf{\Sigma}}_2$ les deux matrices de variances et covariances estimées à partir de chacun des échantillons, l'estimation commune de la matrice des variances et covariances est donnée par la relation :

$$\hat{\mathbf{\Sigma}} = \left(49\hat{\mathbf{\Sigma}}_1 + 49\hat{\mathbf{\Sigma}}_2\right) \Big/ 98 = \begin{bmatrix} 0,26270 & 0,06096 \\ 0,06096 & 0,05727 \end{bmatrix}.$$

On en déduit que l'écart-type estimé des longueurs est égal à 0,513, que l'écart-type estimé des largeurs est égal à 0,239 et, enfin, que le coefficient de corrélation vaut 0,497.

Supposons qu'on souhaite classer un iris dont la longueur des pétales serait de 4,7 et la largeur des pétales de 1,6. Les distances, élevées au carré, de cet iris

aux centres de gravité des deux populations sont égales à :

$$d_{1i}^2 = 1,422 \quad \text{et} \quad d_{2i}^2 = 3,972\,,$$

et les densités de probabilité valent :

$$f_1(x_{1i},\, x_{2i}) = 0,734 \quad \text{et} \quad f_2(x_{1i},\, x_{2i}) = 0,205\,.$$

On en déduit les probabilités *a posteriori* d'appartenance aux deux populations :

$$\mathrm{P}(\mathrm{A1}|\boldsymbol{x}_i) = 0,734/(0,734 + 0,205) = 0,782$$

et

$$\mathrm{P}(\mathrm{A2}|\boldsymbol{x}_i) = 0,205/(0,734 + 0,205) = 0,218\,.$$

L'iris est, par conséquent, classé dans l'espèce *versicolor*.

3.2. Distance généralisée de MAHALANOBIS

Nous avons vu, au paragraphe précédent, que le classement d'un individu se fait en comparant les carrés des deux distances d_{1i} et d_{2i}.

La distance d_{1i} est une mesure de l'éloignement du point i, de coordonnées x_{1i} et x_{2i}, du centre de gravité du nuage des points de l'échantillon 1, dont les coordonnées sont $\hat{m}_{11}$ et $\hat{m}_{12}$. Toutefois, il ne s'agit pas d'une distance euclidienne classique, mais d'une expression de l'éloignement du point i du centre de gravité qui tient compte des variances respectives et de la corrélation des variables x_1 et x_2.

Pour interpréter cette distance, il suffit de se rappeler que tous les individus caractérisés par des couples d'observations conduisant à la même valeur d_{1i}^2 ont des densités de probabilité égales. Donc, pour une distribution normale à deux dimensions, tous les points qui sont situés à une distance constante d du centre de gravité sont situés sur une même ellipse de concentration. La probabilité pour qu'un individu soit situé dans cette ellipse est directement fonction de la distance considérée. En effet, on peut démontrer que cette probabilité est égale à :

$$\mathrm{P}(\chi^2 \leq d^2)\,,$$

la variable χ^2 étant une variable χ^2 de PEARSON à 2 degrés de liberté. Ainsi, par exemple, la probabilité qu'un individu pris au hasard dans la population soit situé à une distance inférieure à $\sqrt{1,39}$ est égale à 0,5 et la probabilité qu'il soit situé à une distance inférieure à $\sqrt{4,61}$ est égale à 0,9.

La figure 2 donne une représentation graphique des observations relatives aux deux espèces d'iris ainsi que les ellipses de concentration correspondant à une probabilité de 0,9.

Les deux ellipses présentent deux points d'intersection et la droite passant par ces deux points divise l'espace $(x_1,\, x_2)$ en deux parties, la zone 1 comprenant le centre de gravité estimé de la population d'*Iris versicolor* et la zone 2 comprenant le centre de gravité estimé de la population d'*Iris virginica*. Les iris seront donc

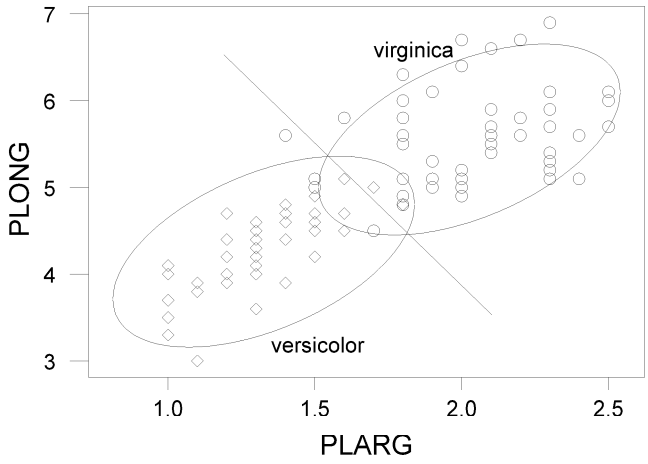

Figure 2. Ellipses de concentration et diagrammes de dispersion de la longueur (PLONG) et de la largeur (PLARG) des pétales pour *Iris versicolor* et *Iris virginica*.

classés dans l'espèce *versicolor* s'ils sont situés dans la zone 1 et dans l'espèce *virginica* s'ils sont situés dans la zone 2.

La position de la droite qui sépare les deux espèces est indépendante de la valeur de la probabilité retenue pour les ellipses de concentration, à condition que cette probabilité soit suffisamment grande pour garantir l'intersection des ellipses. Tous les points de cette droite sont tels que les fonctions de densité de probabilité des deux populations sont égales.

Par ailleurs, la probabilité de classement erroné *a priori* dépend de la séparation des deux populations. Cette séparation peut se mesurer par la distance au sens de MAHALANOBIS entre les centres de gravité des deux populations :

$$\Delta_{12}^2 = (\boldsymbol{m}_1 - \boldsymbol{m}_2)' \, \boldsymbol{\Sigma}^{-1} (\boldsymbol{m}_1 - \boldsymbol{m}_2) \, ,$$

dont une estimation (biaisée) est donnée par :

$$\hat{\Delta}_{12}^2 = (\hat{\boldsymbol{m}}_1 - \hat{\boldsymbol{m}}_2)' \, \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{m}}_1 - \hat{\boldsymbol{m}}_2) \, ,$$

et, comme pour le cas univarié, la probabilité *a priori* de classement erroné est égale à :

$$1 - \Phi(\hat{\Delta}_{12}/2) \, .$$

Pour l'exemple considéré, on a :

$$\hat{\Delta}_{12}^2 = 10,07 \, .$$

et la probabilité de classement erroné est égale à :

$$1 - \Phi(1,587) \approx 0,06 \, .$$

Le classement sur la base des deux variables est donc légèrement meilleur que le classement sur la base de la seule longueur des pétales, car, dans ce cas, la probabilité de classement erroné *a priori* était de 0,10 (paragraphe 2.3). Nous verrons, au paragraphe 7.1, comment améliorer l'estimation de Δ^2_{12} et donc aussi l'estimation de la probabilité de classement erroné.

3.3. Fonction linéaire discriminante

L'équation de la droite séparant l'espace à deux dimensions en deux sous-espaces peut être obtenue par l'intermédiaire de la *fonction linéaire discriminante*[6].

Cette fonction est du type :

$$y = a + b_1 x_1 + b_2 x_2 \,.$$

Il s'agit d'une combinaison linéaire, définie à une constante près, telle que si on réalise une analyse de la variance univariée sur les valeurs de y, le critère de classification étant les populations, la valeur F_{obs} sera maximum.

Cette fonction permet de remplacer les deux variables initiales par une seule variable dérivée, appelée *variable canonique discriminante*[7].

Nous présenterons, au paragraphe 4.2, une méthode générale de calcul des variables canoniques, quels que soient les nombres de variables et les nombres de populations considérés. Dans le cas particulier de deux populations, cette variable canonique peut s'obtenir de manière assez simple par régression multiple. On définit une variable binaire qui possède une valeur donnée pour tous les individus issus de la première population (par exemple, $y = 0$) et une autre valeur pour tous les individus issus de la deuxième population (par exemple, $y = 1$). On calcule ensuite l'équation de régression suivante :

$$y = a + b_1 x_1 + b_2 x_2 \,.$$

Les valeurs y_i, calculées pour chacun des individus par cette relation, représentent les scores des individus, c'est-à-dire les valeurs de la variable canonique.

Si $\bar{y}_1$ et $\bar{y}_2$ sont les moyennes observées de la variable canonique pour les deux groupes, la séparation des deux populations se fait en fonction de la valeur :

$$y_0 = (\bar{y}_1 + \bar{y}_2)/2 \,.$$

Pour classer un individu d'origine inconnue, il suffit de calculer son score, y_i, et de le classer dans la population 1 si $y_i < y_0$ et dans la population 2 si $y_i > y_0$, et l'équation :

$$a + b_1 x_1 + b_2 x_2 = y_0 \,,$$

6. En anglais : *linear discriminant function.*
7. En anglais : *canonical variable.*

est l'équation de la droite qui sépare l'espace à deux dimensions en deux sous-espaces.

Pour l'exemple considéré au paragraphe 3.1 et en donnant à y la valeur 0 pour les iris de l'espèce *versicolor* et la valeur 1 pour les iris de l'espèce *virginica*, l'équation de régression suivante est obtenue :

$$y = -1,5815 + 0,1976x_1 + 0,6634x_2 \,.$$

Si, dans cette équation, on remplace x_1 et x_2 successivement par les moyennes des longueurs et des largeurs des pétales de l'espèce *versicolor* et de l'espèce *virginica* on obtient :

$$\bar{y}_1 = -1,5815 + (0,1976)(4,260) + (0,6634)(1,326) = 0,14 \,,$$

$$\bar{y}_2 = -1,5815 + (0,1976)(5,552) + (0,6634)(2,026) = 0,86$$

et
$$y_0 = \frac{0,14 + 0,86}{2} = 0,5 \,.$$

Pour l'iris à classer dont il a été question au paragraphe 3.1, on a :

$$y = -1,5815 + (0,1976)(4,7) + (0,6634)(1,6) = 0,409 \,,$$

et l'iris est affecté à l'espèce *versicolor*.

Le fait que y_0 soit situé à mi-chemin entre les deux valeurs $y = 0$ et $y = 1$, qui ont été attribuées aux deux espèces, résulte de l'égalité des effectifs des deux échantillons et n'est pas une propriété générale.

En posant :
$$-1,5815 + 0,1976x_1 + 0,6634x_2 = 0,5 \,,$$
on obtient l'équation :
$$x_2 = 3,1376 - 0,2979x_1 \,,$$
qui divise l'espace en deux sous-espaces (figure 2).

D'autre part, le coefficient de détermination multiple, R^2, de l'équation de régression est directement lié à la distance $\hat{\Delta}^2_{12}$.

On a en effet, pour des effectifs égaux :

$$\hat{\Delta}^2_{12} = \frac{4(n-1)R^2}{n(1-R^2)}$$

et pour des effectifs inégaux [DAGNELIE, 1975] :

$$\hat{\Delta}^2_{12} = \frac{(n_1+n_2)(n_1+n_2-2)R^2}{n_1 n_2(1-R^2)} \,.$$

Pour l'exemple considéré, le coefficient de détermination est égal à 0,72 et le carré de la distance de MAHALANOBIS est égal à :

$$\hat{\Delta}^2_{12} = \frac{(4)(49)(0,72)}{(50)(1-0,72)} = 10,08 \,,$$

qui, aux erreurs d'arrondis près, est bien identique à la valeur donnée précédemment.

On peut noter que les principes qui ont été présentés dans le paragraphe 3, et en particulier l'établissement de la règle de classement à partir de la régression multiple, peuvent facilement être étendus au cas où on dispose de plus de deux variables. L'espace des variables n'est plus, dans ce cas, un plan, mais un espace à p dimensions et la division de cet espace en deux sous-espaces se fera à l'aide d'un hyperplan.

4. NOMBRES QUELCONQUES DE POPULATIONS ET DE VARIABLES

4.1. Règle de classement

Soit g populations normales à p dimensions, de matrices de variances et covariances égales, dans lesquelles ont été prélevés g échantillons aléatoires, simples et indépendants.

Soit $\hat{\boldsymbol{m}}_k$ les vecteurs des moyennes des g populations et soit $\hat{\boldsymbol{\Sigma}}$ la matrice commune de variances et covariances, ces paramètres étant estimés à partir des différents échantillons.

Pour un individu, caractérisé par un vecteur d'observations $\boldsymbol{x}_i$, la densité de probabilité correspondant à la population h s'écrit :

$$f_h(\boldsymbol{x}_i) = \frac{1}{\sqrt{(2\pi)^p |\hat{\boldsymbol{\Sigma}}|}} \exp\left(-\frac{1}{2} d_{hi}^2\right)$$

avec : $$d_{hi}^2 = (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_h)' \, \hat{\boldsymbol{\Sigma}}^{-1} \, (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_h) \ ,$$

et la probabilité *a posteriori* d'appartenance à la population h est égale à :

$$\mathrm{P}\,(\mathrm{A}h|\boldsymbol{x}_i) = \frac{\exp\left(-\frac{1}{2} d_{hi}^2\right)}{\sum_{k=1}^{g} \exp\left(-\frac{1}{2} d_{ki}^2\right)} \ .$$

Si les probabilités *a priori* sont égales, l'individu i sera classé dans la population pour laquelle $f_h(\boldsymbol{x}_i)$ est maximum ou pour laquelle d_{hi}^2 est minimum, ou encore pour laquelle $\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i)$ est maximum.

4.2. Fonctions linéaires discriminantes

On peut calculer $g(g-1)/2$ fonctions linéaires discriminantes qui, égalées chaque fois à une constante, déterminent $g(g-1)/2$ hyperplans délimitant g régions auxquelles peuvent être associées les différentes populations considérées.

Pour obtenir l'hyperplan séparant deux populations quelconques, h et l, on peut partir du rapport des fonctions de densité de probabilité relatives à ces deux populations, qui s'appelle le *rapport de vraisemblance*[8] :

$$L_{hl} = \frac{f_h(\boldsymbol{x})}{f_l(\boldsymbol{x})} = \frac{\left(1/\sqrt{(2\pi)^p|\hat{\boldsymbol{\Sigma}}|}\right) \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_h\right)' \hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_h\right)\right]}{\left(1/\sqrt{(2\pi)^p|\hat{\boldsymbol{\Sigma}}|}\right) \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_l\right)' \hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_l\right)\right]}.$$

Après simplification et en prenant le logarithme de l'expression, on obtient :

$$\log_e(L_{hl}) = \left(\hat{\boldsymbol{m}}_h\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} - \frac{1}{2}\hat{\boldsymbol{m}}_h'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{m}}_h\right) - \left(\hat{\boldsymbol{m}}_l\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} - \frac{1}{2}\hat{\boldsymbol{m}}_l'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{m}}_l\right).$$

Les vecteurs $\boldsymbol{x}$ qui annulent cette expression appartiennent à l'hyperplan qui sépare la population h de la population l.

La fonction :

$$\hat{\boldsymbol{m}}_h\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} - \frac{1}{2}\hat{\boldsymbol{m}}_h'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{m}}_h$$

est appelée fonction linéaire discriminante de la population h et le calcul pour un individu donné de la fonction pour chacun des groupes permet également de classer cet individu dans le groupe pour lequel la fonction linéaire discriminante est maximum.

On notera que les fonctions discriminantes dont il est question ici n'ont pas la même signification que celle donnée au paragraphe 3.3. Pour retrouver les fonctions équivalentes à celles du paragraphe 3.3, il suffit de soustraire les fonctions relatives à deux groupes donnés.

Pour illustrer les principes ci-dessus, considérons les trois espèces d'iris, en nous limitant cependant à deux variables, afin de permettre les représentations graphiques.

Le figure 3 reprend une partie des résultats obtenus par le logiciel Minitab [X, 1994]. Dans cette figure, le groupe 1 représente l'espèce *setosa*, le groupe 2 représente l'espèce *versicolor* et le groupe 3 représente l'espèce *virginica*. PLONG et PLARG représentent, respectivement, la longueur et la largeur des pétales.

La première partie de cette figure donne les moyennes, les écarts-types et les matrices de variances et covariances estimés par espèce et pour l'ensemble des trois espèces. Dans la seconde partie, on trouve les distances, élevées au carré, entre les groupes :

$$\hat{\Delta}^2_{hl} = \left(\hat{\boldsymbol{m}}_h - \hat{\boldsymbol{m}}_l\right)' \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{m}}_h - \hat{\boldsymbol{m}}_l).$$

On voit immédiatement que la distance la plus faible concerne les espèces *versicolor* et *virginica* et que la distance la plus élevée est celle qui sépare *virginica* de *setosa*. Les risques de confusion des espèces seront donc plus importants pour

8. En anglais : *likelihood ratio*.

```
Variable   Pooled    Means for Group
             Mean        1        2        3
PLONG      3.7580   1.4620   4.2600   5.5520
PLARG      1.1993   0.2460   1.3260   2.0260

Variable   Pooled    Stdev for Group
            Stdev        1        2        3
PLONG      0.4303   0.1737   0.4699   0.5519
PLARG      0.2047   0.1054   0.1978   0.2747

Pooled Covariance Matrix
            PLONG    PLARG
PLONG     0.18519
PLARG     0.04267  0.04188

Covariance Matrix for Group  1
             PLONG     PLARG
PLONG     0.030159
PLARG     0.006069  0.011106

Covariance Matrix for Group  2
             PLONG     PLARG
PLONG     0.220816
PLARG     0.073102  0.039106

Covariance Matrix for Group  3
             PLONG     PLARG
PLONG     0.304588
PLARG     0.048824  0.075433

Squared Distance Between Groups
                1        2        3
1           0.000   48.189  112.225
2          48.189    0.000   14.064
3         112.225   14.064    0.000

        Linear Discriminant Function for Group
                1        2        3
Constant   -5.900  -50.850  -91.927
PLONG       8.548   20.527   24.612
PLARG      -2.834   10.749   23.302
```

Figure 3. Analyse discriminante linéaire pour trois espèces caractérisées par deux variables : résultats fournis par Minitab.

Tableau 3. Différences entre les fonctions discriminantes relatives aux trois couples d'espèces d'iris, dans le cas de deux variables.

Variables	*setosa*-*versicolor*	*setosa*-*virginica*	*versicolor*-*virginica*
constante	44,95	86,03	41,08
PLONG	–11,980	–16,064	–4,085
PLARG	–13,583	–26,136	–12,553

versicolor et *virginica* que pour *virginica* et *setosa*. On remarque aussi que la distance au carré qui sépare *versicolor* de *virginica*, qui est de 14,06, est un peu différente de la valeur donnée au paragraphe 3.2, suite à la modification de la matrice des variances et covariances estimées liée à la prise en compte des données de l'espèce *setosa*.

La troisième partie de la figure donne les fonctions linéaires discriminantes définies ci-dessus.

Si on souhaite affecter un iris caractérisé par une longueur des pétales de 4,7 et une largeur des pétales de 1,6, on peut remplacer, dans ces fonctions, PLONG et PLARG par ces valeurs et on obtient :

$$y_1 = -5,900 + (8,548)(4,7) - (2,834)(1,6) = 29,74\,,$$

$$y_2 = -50,850 + (20,527)(4,7) + (10,749)(1,6) = 62,83$$

et $$y_3 = -91,927 + (24,612)(4,7) + (23,302)(1,6) = 61,03\,.$$

La valeur la plus grande est obtenue pour y_2 et l'iris d'origine inconnue est donc affecté à l'espèce *versicolor*.

Le tableau 3 donne les différences entre les fonctions pour tous les couples d'espèces. En égalant à zéro les trois fonctions, on obtient les équations des trois droites suivantes :

$$\text{PLONG} = 3,752 - 1,134\ \text{PLARG},$$

$$\text{PLONG} = 5,362 - 1,629\ \text{PLARG},$$

$$\text{PLONG} = 10,068 - 3,077\ \text{PLARG}.$$

Comme le schématise la figure 4, ces trois droites se coupent au point de coordonnées :

$$\text{PLONG} = 0,0661 \quad \text{et} \quad \text{PLARG} = 3,251.$$

La première droite, notée 1|2 dans le graphique, sépare *setosa* de *versicolor*; la deuxième droite, notée 1|3, sépare *setosa* de *virginica* et la troisième droite, notée 2|3, sépare *versicolor* de *virginica*.

Les trois demi-droites représentées en trait plein dans la figure 4 divisent l'espace en trois régions, une région étant affectée à chacune des espèces. Toutefois, la demi-droite en trait plein séparant *virginica* et *setosa* ne présente aucun

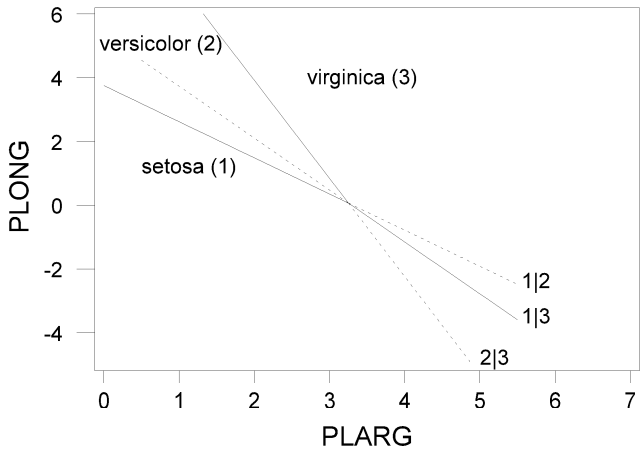

Figure 4. Représentation schématique des trois droites séparant les trois espèces d'iris.

intérêt pratique, dans la mesure où on n'observe jamais d'iris dont la largeur des pétales dépasse 3,25 cm alors que la longueur des pétales est inférieure à 0,066 cm.

La figure 5 reprend les diagrammes de dispersion pour les trois espèces ainsi que les deux droites séparant l'espace utile en trois régions. Cette figure indique clairement que la séparation la moins nette concerne *Iris versicolor* et *Iris virginica*. Par contre, l'espèce *Iris setosa* est bien séparée des deux autres.

4.3. Analyse des variables canoniques discriminantes

Nous avons signalé, au paragraphe 3.3, que la fonction linéaire discriminante permettait de remplacer les variables initiales par une seule variable dérivée, appelée variable canonique discriminante. Cette propriété se vérifie d'ailleurs quel que soit le nombre de variables, p, pour autant que le nombre de populations soit égal à deux.

En fait, la notion de variable canonique est une notion qui n'est pas limitée aux problèmes de classement que nous étudions dans la présente note. Il s'agit d'un concept plus général permettant, notamment, de décrire g échantillons caractérisés par p variables. Le principe consiste, comme dans le cas de l'analyse en composantes principales, à calculer des nouvelles variables à partir des variables initiales. Ces nouvelles variables sont des combinaisons linéaires des variables initiales réduites. Elles sont non corrélées et d'importance décroissante.

La première variable canonique discriminante est telle que le rapport de la variance entre les groupes à la variance dans les groupes soit maximum. La seconde variable canonique vise aussi à maximiser ce rapport avec, en plus, la

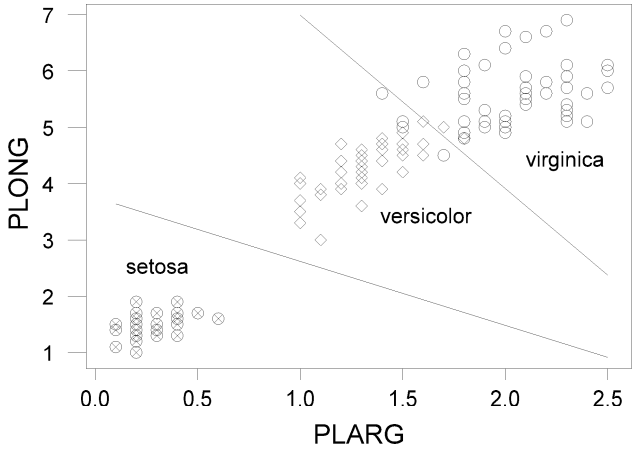

Figure 5. Représentation des trois espèces dans l'espace des deux variables, PLONG et PLARG, et division de l'espace en trois régions.

contrainte de non-corrélation à la première variable déjà construite, et ainsi de suite pour les autres variables canoniques. Le nombre total de variables canoniques qui peuvent être calculées ne peut dépasser ni p ni $g-1$. Géométriquement, les variables canoniques correspondent aux projections des points-individus sur une série d'axes, qui représentent les directions dans lesquelles les différences entre les groupes sont les plus marquées.

Le calcul de ces variables canoniques repose sur la décomposition des sommes des carrés et des produits des écarts totales en deux parties, l'une liée aux différences entre les échantillons, l'autre liée aux différences dans les échantillons. Cette décomposition conduit à l'équation fondamentale de l'analyse de la variance multivariée à un critère :

$$\mathbf{T} = \mathbf{H} + \mathbf{E}\,,$$

où $\mathbf{T}$, $\mathbf{H}$ et $\mathbf{E}$, représentent les matrices des sommes des carrés et des produits des écarts totales, factorielles (entre échantillons) et résiduelles (dans les échantillons).

Les coefficients des variables canoniques, c'est-à-dire les poids des variables initiales (réduites) dans les combinaisons linéaires sont les vecteurs propres associés aux valeurs propres de la matrice $\mathbf{E}^{-1}\mathbf{H}$. Ces vecteurs propres sont généralement standardisés de manière à ce que les variables canoniques soient de variance résiduelle unitaire.

Exprimées en pour cent de leur somme, les valeurs propres de $\mathbf{E}^{-1}\mathbf{H}$ donnent une idée de la qualité de la représentation sur les différents axes. Ainsi, si la première valeur propre correspond à un pourcentage important, cela signifie que, dans l'espace initial à p dimensions, les g moyennes se trouvent à proximité

d'une droite. Si la somme des deux premières valeurs propres correspond à un pourcentage important, les g moyennes se situent à proximité d'un plan. Et ainsi de suite, pour trois valeurs propres, quatre valeurs propres, etc.

Lorsque les conditions d'application de l'analyse de la variance multivariée à un critère sont remplies (populations normales à p dimensions de même matrice de variances et covariances et échantillons aléatoires, simples et indépendants), on peut tester la signification des valeurs propres afin de déterminer la dimension de l'espace le plus petit dans lequel se situent les g moyennes. Des informations à ce sujet sont données dans DAGNELIE [1975] et PALM [1990].

On peut montrer que la règle de classement établie sur les variables initiales est identique à la règle de classement établie sur l'ensemble des variables canoniques. Cette substitution ne présente, en général, guère d'intérêt, compte tenu des moyens de calculs actuellement disponibles. Par contre, lorsque la première variable canonique contient l'essentiel de l'information, la procédure de classement peut être basée uniquement sur cette variable canonique, cette simplification étant alors sans préjudice important pour la qualité des résultats.

Pour les trois espèces d'iris et les deux caractéristiques des pétales, une analyse des variables canoniques a été réalisée par la procédure CANDISC de SAS et la figure 6 reprend une partie des résultats obtenus. On constate tout d'abord que la première valeur propre de $\mathbf{E}^{-1}\mathbf{H}$, qui vaut 19,68, correspond à 99,5 % de la somme des deux valeurs propres. Les moyennes des trois groupes se situent donc approximativement sur une droite, ce qu'on pouvait d'ailleurs déjà observer sur la figure 5.

La figure 6 donne également les coefficients des deux combinaisons linéaires qui permettent de calculer les valeurs des variables canoniques. Ces coefficients doivent être appliqués aux variables initiales standardisées, la standardisation se faisant par rapport à l'écart-type résiduel. Ainsi, l'équation du premier axe canonique, dénommé CAN1, est égale à :

$$\text{CAN1} = 0,66460\left(\frac{\text{PLONG} - 3,7580}{0,4303}\right) + 0,49165\left(\frac{\text{PLARG} - 1,1993}{0,2047}\right)$$

ou encore :

$$\text{CAN1} = -8,685 + 1,545\text{PLONG} + 2,402\text{PLARG}\,.$$

De façon analogue, on pourrait déterminer le deuxième axe canonique, CAN2.

Les moyennes de CAN1 et CAN2 pour les trois espèces d'iris sont également données dans la figure 6. Ces moyennes s'obtiennent en remplaçant, dans les équations de CAN1 et de CAN2, les valeurs de PLONG et PLARG par les moyennes des groupes. Pour le premier axe et pour *Iris setosa* par exemple, on a :

$$-8,685 + (1,545)(1,462) + (2,402)(0,246) = -5,836\,.$$

Si on souhaite établir la règle de classement sur base de la première variable canonique, la séparation d'*Iris setosa* et d'*Iris versicolor* se situe en :

$$(-5,836 + 1,080)/2 = -2,38\,,$$

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	19.6773	19.5726	0.9947	0.9947
2	0.1047	.	0.0053	1.0000

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2	
PLONG	0.664595872	-0.930048472	LONGUEUR DES PETALES (CM)
PLARG	0.491650070	1.031968043	LARGEUR DES PETALES (CM)

Class Means on Canonical Variables

ESPECE	CAN1	CAN2
1	-5.836157383	0.154888649
2	1.079577295	-0.446204392
3	4.756580088	0.291315743

Figure 6. Analyse des variables canoniques des trois espèces caractérisées par deux variables : résultats fournis par SAS.

et la séparation d'*Iris versicolor* et d'*Iris virginica* se situe en :

$$(1,0796 + 4,757)/2 = 2,92\,.$$

L'iris d'origine inconnue caractérisé par une longueur des pétales de 4,7 et une largeur des pétales de 1,6 (paragraphe 4.2) a comme valeur de la première variable canonique :

$$-8,685 + (1,5445)(4,7) + (2,4024)(1,6) = 2,42\,;$$

il serait donc classé dans l'espèce *versicolor*.

Géométriquement, les relations :

$$-8,685 + 1,5445\mathrm{PLONG} + 2,4024\mathrm{PLARG} = -2,38$$

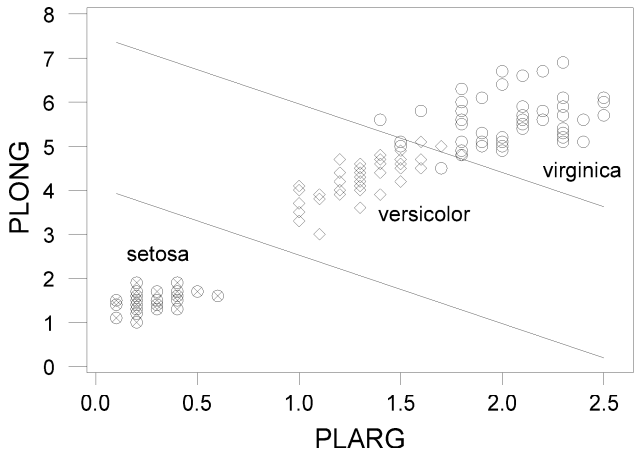

Figure 7. Représentation des trois espèces dans l'espace des deux variables PLONG et PLARG et division de l'espace en trois régions d'après la première variable canonique.

et $$-8,685 + 1,5445\text{PLONG} + 2,4024\text{PLARG} = 2,92$$

sont les équations des deux droites parallèles qui partagent le plan des variables en trois régions, une région étant affectée à chacune des espèces. Ces droites ont été matérialisées sur la figure 7. Un axe perpendiculaire à ces droites représente l'axe canonique et la projection des observations sur cet axe donne, pour chaque observation, la valeur de la première variable canonique.

La comparaison des figures 5 et 7 montre que les sous-espaces affectés à chacune des trois espèces sont pratiquement identiques. La simplification résultant de l'utilisation d'une seule variable canonique ne conduit pas à une diminution de la qualité de la règle de classement, puisque les moyennes des trois groupes sont pratiquement sur une droite.

La figure 8 donne une représentation graphique des 150 iris dans le plan des deux variables canoniques, dénommées CAN1 et CAN2. On voit immédiatement que les différences entre les trois espèces se marquent sur la variable CAN1. Les deux droites divisant l'espace en trois régions sur base de la première variable canonique, sont également reprises sur le graphique.

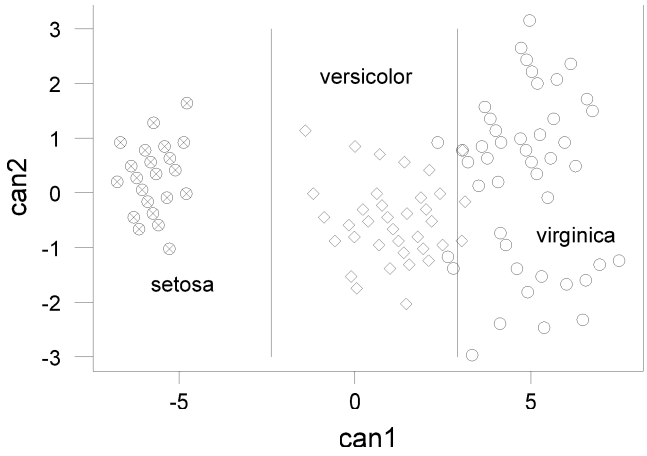

Figure 8. Représentation des trois espèces d'iris dans le plan des deux variables canoniques et division de l'espace en trois régions d'après la première variable canonique.

5. AUTRES MÉTHODES D'ANALYSE DISCRIMINANTE

5.1. Analyse discriminante quadratique

Dans les paragraphes précédents, nous avons considéré le cas de populations normales à p dimensions de même matrice de variances et covariances. La méthode peut cependant être étendue sans difficulté à des populations normales n'ayant pas des matrices identiques. Dans ce cas et pour des probabilités d'appartenance *a priori* égales, le rapport des fonctions de densité de probabilité des populations h et l s'écrit :

$$\frac{f_h(\boldsymbol{x})}{f_l(\boldsymbol{x})} = \frac{\left(1/\sqrt{(2\pi)^p|\hat{\boldsymbol{\Sigma}}_h|}\right) \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_h\right)' \hat{\boldsymbol{\Sigma}}_h^{-1}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_h\right)\right]}{\left(1/\sqrt{(2\pi)^p|\hat{\boldsymbol{\Sigma}}_l|}\right) \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_l\right)' \hat{\boldsymbol{\Sigma}}_l^{-1}\left(\boldsymbol{x}-\hat{\boldsymbol{m}}_l\right)\right]} .$$

Dans cette relation, $\hat{\boldsymbol{\Sigma}}_h$ et $\hat{\boldsymbol{\Sigma}}_l$ représentent les matrices de variances et covariances estimées à partir des données des échantillons h et l.

Après simplification et transformation logarithmique, on conclut que, pour un vecteur $\boldsymbol{x}_i$ fixé, $f_h(\boldsymbol{x}_i)$ est supérieure à $f_l(\boldsymbol{x}_i)$ si :

$$d_{hi}^{'2} = d_{hi}^2 + \log_e|\hat{\boldsymbol{\Sigma}}_h| < d_{li}^{'2} = d_{li}^2 + \log_e|\hat{\boldsymbol{\Sigma}}_l|$$

avec :

$$d_{hi}^2 = \left(\boldsymbol{x}_i-\hat{\boldsymbol{m}}_h\right)' \hat{\boldsymbol{\Sigma}}_h^{-1}\left(\boldsymbol{x}_i-\hat{\boldsymbol{m}}_h\right) \quad \text{et} \quad d_{li}^2 = \left(\boldsymbol{x}_i-\hat{\boldsymbol{m}}_l\right)' \hat{\boldsymbol{\Sigma}}_l^{-1}\left(\boldsymbol{x}_i-\hat{\boldsymbol{m}}_l\right) .$$

Tableau 4. Caractéristiques relatives au classement d'un iris donné dans le cas de matrices de variances et covariances inégales.

Espèces	*Iris setosa*	*Iris versicolor*	*Iris virginica*
d^2_{hi}	395,149	0,148	4,866
$\log_e\lvert\hat{\boldsymbol{\Sigma}}_h\rvert$	–5,716	–3,883	–5,127
d'^2_{hi}	389,433	–3,735	–0,261
$\mathrm{P}(\mathrm{A}h\lvert\boldsymbol{x}_i)$	0,00	0,85	0,15

Il en résulte que la probabilité *a posteriori* d'appartenance à la population h s'écrit :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{\exp\left(-\frac{1}{2}d'^2_{hi}\right)}{\sum_{k=1}^{g}\exp\left(-\frac{1}{2}d'^2_{ki}\right)}.$$

Quant à la fonction permettant de séparer le sous-espace relatif à la population h du sous-espace relatif à la population l, elle s'obtient en égalant à zéro le logarithme du rapport des densités de probabilité relatives aux deux populations. Il ne s'agit cependant plus d'une fonction linéaire mais bien d'une fonction quadratique. C'est la raison pour laquelle on parle, dans ce cas, d'analyse discriminante quadratique.

Pour les trois espèces d'iris et les deux variables relatives aux pétales, l'examen des matrices des variances et covariances montre que les valeurs sont sensiblement plus faibles pour l'espèce *setosa* que pour les deux autres espèces (figure 3). Le test d'égalité de ces trois matrices, dont nous parlerons au paragraphe 8.1, conduit d'ailleurs au rejet de l'hypothèse d'égalité de ces matrices. L'analyse discriminante quadratique constituerait donc une méthode théoriquement plus correcte.

A titre d'illustration, nous avons examiné le classement d'un iris caractérisé par une longueur de pétales de 4,7 cm et une largeur des pétales de 1,6 cm en considérant des probabilités d'appartenance *a priori* égales. Le tableau 4 reprend les éléments permettant de calculer les probabilités d'appartenance *a posteriori*. L'iris est classé dans l'espèce *versicolor*, la probabilité étant égale à 0,85. En considérant des matrices de variances et covariances égales, la probabilité d'appartenance *a posteriori* à l'espèce *versicolor* est de 0,86. Pour cet exemple, la prise en compte des matrices de variances et covariances propres à chacune des populations ne modifie pas les résultats de façon importante.

5.2. L'analyse discriminante logistique

Pour g populations normales multivariées, de même matrice de variances et covariances, nous avons vu, au paragraphe 4.2, que le rapport de vraisemblance

de la population h par rapport à une population de référence, par exemple la population g, s'écrit :

$$L_{hg}(\boldsymbol{x}_i) = \frac{f_h(\boldsymbol{x}_i)}{f_g(\boldsymbol{x}_i)} = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{m}_h\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{m}_h\right)\right]}{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{m}_g\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{m}_g\right)\right]}$$

$$= \exp\left[-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{m}_h\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{m}_h\right) + \frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{m}_g\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{m}_g\right)\right].$$

Après simplification, l'expression peut s'écrire :

$$L_{hg}(\boldsymbol{x}_i) = \exp\left(a_{0h} + \boldsymbol{a}_h' \boldsymbol{x}_i\right)$$

avec $$a_{0h} = -\tfrac{1}{2}\left(\boldsymbol{m}_h - \boldsymbol{m}_g\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m}_h + \boldsymbol{m}_g\right)$$

et $$\boldsymbol{a}_h' = \left(\boldsymbol{m}_h - \boldsymbol{m}_g\right)' \boldsymbol{\Sigma}^{-1}.$$

L'expression :

$$a_{0h} + \boldsymbol{a}_h' \boldsymbol{x}_i,$$

est donc une fonction linéaire de $\boldsymbol{x}_i$.

La densité *a posteriori* peut être exprimée en fonction des rapports de vraisemblance. Pour des probabilités *a priori* égales, on a :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{L_{hg}(\boldsymbol{x}_i)}{\sum_{k=1}^{g} L_{kg}(\boldsymbol{x}_i)} = \frac{\exp\left(a_{0h} + \boldsymbol{a}_h' \boldsymbol{x}_i\right)}{\sum_{k=1}^{g} \exp\left(a_{0k} + \boldsymbol{a}_k' \boldsymbol{x}_i\right)}.$$

L'analyse discriminante logistique a pour principe de considérer l'expression ci-dessus pour la densité de probabilité *a posteriori* sans émettre comme condition d'application que le modèle est multinormal, car la relation donnant les probabilités *a posteriori*, est vraie, non seulement pour les distributions multinormales, mais aussi pour de nombreuses autres distributions multivariées.

La différence entre l'analyse discriminante logistique et l'analyse discriminante linéaire classique porte sur l'estimation des paramètres des fonctions linéaires.

Pour l'analyse discriminante linéaire classique, nous venons de présenter les relations qui existent entre, d'une part, les a_{0h} et $\boldsymbol{a}_h'$ des fonctions linéaires et, d'autre part, les vecteurs des moyennes et la matrice de variances et covariances. Pour estimer les coefficients, il suffit de remplacer, dans les relations ci-dessus, les paramètres des distributions normales par leurs estimations.

Dans le cas de l'analyse discriminante logistique, on recherche des estimateurs du maximum de vraisemblance des coefficients, en déterminant les valeurs qui rendent maximum l'expression :

$$\lambda = \prod_{h=1}^{g} \prod_{j=1}^{n_h} \frac{\exp\left(a_{0h} + \boldsymbol{a}_h^{'} \boldsymbol{x}_{hj}\right)}{\sum_{k=1}^{g} \exp\left(a_{0k} + \boldsymbol{a}_k^{'} \boldsymbol{x}_{kj}\right)} .$$

Des informations relatives à ce sujet sont données par ALBERT et HARRIS [1987].

5.3. Méthodes des plus proches voisins

Les *méthodes des plus proches voisins*[9] sont des méthodes non paramétriques, qui ne reposent sur aucune hypothèse particulière à propos de la forme des populations multivariées.

Pour expliquer le principe à la base de ces méthodes, considérons d'abord le cas particulier du plus proche voisin. Dans ce cas, si on souhaite classer un individu caractérisé par un vecteur d'observations $\boldsymbol{x}_i$, on détermine à quel groupe appartient l'individu de l'échantillon qui est le plus proche de $\boldsymbol{x}_i$. L'application de cette règle implique que la distance entre deux individus ait été préalablement définie. On retiendra, par exemple, la distance euclidienne ou la distance au sens de MAHALANOBIS.

Au lieu de considérer uniquement le voisin le plus proche, on peut aussi tenir compte des r voisins les plus proches. Si r_h est le nombre d'individus appartenant au groupe h parmi ces r voisins, on affecte l'individu $\boldsymbol{x}_i$ à la population pour laquelle le rapport r_h/n_h est le plus grand, n_h étant l'effectif de l'échantillon du groupe h.

A partir du rapport r_h/n_h, on peut estimer la densité de probabilité relative au groupe h au point $\boldsymbol{x}_i$. Il suffit pour cela de diviser la fréquence relative par le volume $v_h(\boldsymbol{x}_i)$ qui contient les r plus proches voisins :

$$f_h(\boldsymbol{x}_i) = \frac{r_h}{n_h v_h(\boldsymbol{x}_i)} .$$

Ce volume, centré sur $\boldsymbol{x}_i$, est limité par l'ellipsoïde d'équation :

$$(\boldsymbol{x} - \boldsymbol{x}_i)^{'} \boldsymbol{\Sigma}_h^{-1} (\boldsymbol{x} - \boldsymbol{x}_i) = d^2(\boldsymbol{x}_r, \boldsymbol{x}_i)$$

$d^2(\boldsymbol{x}_r, \boldsymbol{x}_i)$ étant la distance, au sens de MAHALANOBIS du point $\boldsymbol{x}_i$ au $r^{ième}$ plus proche voisin. Si, comme dans le logiciel SAS, on considère une matrice de variances et covariances commune aux différentes populations, alors le volume de l'ellipsoïde est identique pour chaque population et est fonction du point $\boldsymbol{x}_i$ et du nombre r de voisins considérés. L'indice h peut donc être supprimé.

9. En anglais : *nearest neighbour method.*

L'utilisateur doit fixer la valeur de r. Une approche pratique consiste à répéter l'analyse avec diverses valeurs de r et à retenir la valeur qui conduit aux meilleurs résultats par la validation croisée, dont nous parlerons au paragraphe 7.2. HUBERTY [1994] propose, par exemple, de comparer les résultats pour $r = 1$, $r = 3$ et $r = 5$.

5.4. Méthodes du noyau

Les *méthodes du noyau*[10] reposent sur une estimation des fonctions de densité de probabilité, qui ne suppose pas *a priori* une forme particulière de celles-ci. Elles font également partie des méthodes non paramétriques.

Pour comprendre le principe d'estimation de la densité de probabilité, considérons d'abord le cas d'une seule variable, x, pour laquelle on disposerait de n observations. Une estimation de la densité de probabilité en un point x_0 donné peut être obtenue en définissant un intervalle de longueur $\Delta x = 2r$, centré sur x_0, et en calculant la fréquence unitaire dans cet intervalle :

$$f(x_0) = \frac{n_{x_0}}{2rn} ,$$

n_{x_0} représentant le nombre d'observations situées dans l'intervalle $(x_0 - r, x_0 + r)$. Le calcul peut être fait pour toute valeur de x, en considérant un intervalle de classe mobile, mais de longueur constante. L'estimation ci-dessus peut encore s'écrire, pour tout x :

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) ,$$

$K(x, x_i)$ étant la fonction suivante :

$$K(x, x_i) = \begin{cases} \frac{1}{2r} & \text{si} \quad |x - x_i| \leq r \\ 0 & \text{sinon.} \end{cases}$$

La fonction $K(x, x_i)$ est appelée fonction du noyau. Pour l'exemple considéré, cette fonction de lissage est une fonction de densité de probabilité uniforme. Le degré de lissage dépend de la valeur de r qui est la dimension de la fenêtre de lissage. Une valeur grande de r conduit à une fonction de densité de probabilité estimée variant lentement tandis qu'une valeur faible de r conduit à un résultat plus irrégulier.

La méthode peut être étendue à d'autres fonctions de lissage que la fonction uniforme, par exemple la distribution normale, et peut être généralisée au cas de plusieurs variables. Dans ce cas, la fonction de densité de probabilité du noyau est une fonction multivariée. Ainsi, pour une fonction de lissage de densité uniforme, on a :

$$K(x, x_i) = \begin{cases} \frac{1}{v} & \text{si} \quad (\boldsymbol{x} - \boldsymbol{x}_i)' \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{x} - \boldsymbol{x}_i) \leq r^2 \\ 0 \text{ sinon.} \end{cases}$$

10. En anglais : *kernel methods.*

La constante v intervenant dans cette expression est le volume de l'ellipsoïde à p dimensions dont la surface externe est le lieu des points situés à une distance constante, au sens de MAHALANOBIS, du point $\boldsymbol{x}_i$ et $\boldsymbol{\Sigma}^{-1}$ est la matrice de variances et covariances. Pour cette fonction de lissage, le principe de l'estimation de la densité de probabilité est donc tout à fait semblable à celle utilisée pour la méthode des r plus proches voisins. La différence réside simplement dans le choix du rayon de l'ellipsoïde. Il est fonction de $\boldsymbol{x}_i$ dans la méthode des r voisins les plus proches alors qu'il est constant dans la méthode du noyau.

Pour l'affectation d'un individu $\boldsymbol{x}_i$ à une population donnée, on estime les densités de probabilité relatives à chaque population en $\boldsymbol{x}_i$ et pour des probabilités *a priori* égales, on affecte l'individu à la population pour laquelle la densité de probabilité estimée est maximum ou, ce qui revient au même, pour laquelle la probabilité *a posteriori* :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{f_h(x_i)}{\sum_{k=1}^{g} f_k(\boldsymbol{x}_i)} ,$$

est la plus grande.

Comme pour la méthode des plus proches voisins, se pose ici le problème du choix de la fonction utilisée pour le lissage et de la dimension de la fenêtre de lissage. Ici aussi, on peut tester différentes solutions et retenir celle qui conduit aux meilleurs résultats. Différentes solutions peuvent aussi être retenues, dans le cas multivarié, pour la matrice de variances et covariances : on peut, par exemple, considérer une matrice commune pour les différentes populations, ou, au contraire, des matrices spécifiques aux g populations.

5.5. Méthodes pour données qualitatives et données mixtes

Jusqu'à présent, nous nous sommes limités au cas où les variables sont quantitatives. Dans certaines situations cependant, les variables ou du moins certaines variables servant à la discrimination sont qualitatives.

Si les variables qualitatives sont binaires, on peut les remplacer par des variables de type $0/1$ et utiliser les méthodes pour données quantitatives présentées dans les paragraphes précédents.

Pour des données ordinales, on peut aussi procéder à un codage préliminaire. Ainsi, une variable présentant les modalités "faible", "moyen" et "élevé" peut être remplacée par une variable quantitative en donnant à ces trois modalités respectivement la valeur 0, 1 et 2.

Les variables nominales à q modalités peuvent également être remplacées par $q-1$ variables binaires, mais on risque alors de se trouver en présence d'un trop grand nombre de variables.

Une solution est de procéder à une analyse factorielle des correspondances de ce tableau disjonctif complet. On retient alors les scores des individus sur les

facteurs qui sont les plus discriminants et on réalise une analyse discriminante sur ces nouvelles variables, qui sont continues. Cette méthode est connue sous le nom de méthode DISQUAL [SAPORTA, 1990].

De nombreuses autres approches sont encore proposées dans la littérature. Nous ne les présenterons pas ici. Des informations à ce sujet sont données par CELEUX et NAKACHE [1994], HAND [1981], HUBERTY [1994] et MCLACHLAN [1992], notamment.

6. PROBABILITÉS A PRIORI INÉGALES

6.1. Estimation des probabilités *a priori*

Dans les paragraphes précédents, nous avons toujours considéré que les probabilités *a priori* sont identiques pour tous les groupes. Rappelons que ces probabilités *a priori* correspondent aux probabilités d'appartenance aux différents groupes pour un individu donné, avant que les observations relatives à cet individu, $\boldsymbol{x}_i$, ne soient connues. Elles ont été notées $\mathrm{P}(\mathrm{A}h)$ au paragraphe 2.3. Pour plus de concision, nous les noterons par la suite p_h.

Si les probabilités *a priori* sont inégales, les règles d'affectation des individus aux différentes populations peuvent être adaptées en conséquence (paragraphe 6.2).

Ces probabilités *a priori* peuvent être connues par des études antérieures ou bien sont estimées à partir des proportions observées dans l'échantillon des différentes populations. Cette dernière façon de procéder n'est évidemment valable que si l'échantillon a été prélevé de manière aléatoire et simple dans le mélange des populations.

6.2. Règles de classement

Lorsque les probabilités *a priori* varient avec le groupe considéré, les probabilités *a posteriori* s'écrivent :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{p_h f_h(\boldsymbol{x}_i)}{\sum_{k=1}^{g} p_k f_k(\boldsymbol{x}_i)} ,$$

et on affecte l'individu à la population pour laquelle cette valeur est maximum. Cette procédure résulte de la densité de probabilité du mélange des g populations qui est égale à [DAGNELIE, 1998] :

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{g} p_k f_k(\boldsymbol{x}_i) ,$$

et $p_h f_h(\boldsymbol{x}_i)$ est, pour un vecteur $\boldsymbol{x}_i$ fixé, la proportion de la densité de probabilité qui est liée à la population h.

Pour des populations normales de matrices de variances et covariances identiques, le rapport des densités de probabilité relatives à deux populations, h et l, multipliées par les probabilités *a priori* est égal à :

$$\frac{p_h f_h(\boldsymbol{x}_i)}{p_l f_l(\boldsymbol{x}_i)} = \frac{p_h \exp\left[-\frac{1}{2}d_{hi}^2\right]}{p_l \exp\left[-\frac{1}{2}d_{li}^2\right]} .$$

En prenant le logarithme de ce rapport, on trouve:

$$\log_e\left(\frac{p_h f_h(\boldsymbol{x}_i)}{p_l f_l(\boldsymbol{x}_i)}\right) = \left(\log_e p_h - \frac{1}{2}d_{hi}^2\right) - \left(\log_e p_l - \frac{1}{2}d_{li}^2\right) .$$

Il en résulte que $p_h f_h(\boldsymbol{x}_i)$ est supérieur à $p_l f_l(\boldsymbol{x}_i)$ si :

$$d_{hi}^{'2} < d_{li}^{'2}$$

avec : $\quad d_{hi}^{'2} = d_{hi}^2 - 2\log_e p_h \quad$ et $\quad d_{li}^{'2} = d_{li}^2 - 2\log_e p_l$.

D'autre part, $p_h f_h(\boldsymbol{x}_i)$ est proportionnel à $d_{hi}^{'2}$ et, par conséquent, la probabilité *a posteriori* est égale à :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{p_h f_h(\boldsymbol{x}_i)}{\sum_{k=1}^{g} p_k f_k(\boldsymbol{x}_i)} = \frac{\exp\left[-\frac{1}{2}d_{hi}^{'2}\right]}{\sum_{k=1}^{g} \exp\left[-\frac{1}{2}d_{ki}^{'2}\right]} .$$

Le résultat ci-dessus peut encore être étendu au cas de matrices de variances et covariances inégales. On calcule alors les distances :

$$d_{hi}^{''2} = (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_h)' \, \hat{\boldsymbol{\Sigma}}_h^{-1} \, (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_h) + \log_e|\hat{\boldsymbol{\Sigma}}_h| - 2\log_e p_h$$

et on affecte l'individu i à la population pour laquelle $d_{hi}^{''2}$ est le plus petit, les probabilités d'appartenance *a posteriori* étant égales à :

$$p(\mathrm{A}h|\boldsymbol{x}_i) = \frac{\exp\left[-\frac{1}{2}d_{hi}^{''2}\right]}{\sum_{k=1}^{g} \exp\left[-\frac{1}{2}d_{ki}^{''2}\right]} .$$

De même, pour l'analyse discriminante logistique, on a :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{p_h L_{hg}}{\sum_{k=1}^{g} p_k L_{kg}} .$$

Tableau 5. Caractéristiques relatives au classement d'un iris donné dans le cas de matrices de variances et covariances et probabilités *a priori* inégales.

Espèces	*Iris setosa*	*Iris versicolor*	*Iris virginica*
p_h	0,20	0,20	0,60
d^2_{hi}	395,149	0,148	4,866
$\log_e\lvert\hat{\boldsymbol{\Sigma}}_h\rvert$	–5,716	–3,883	–5,127
$-2\log_e p_h$	3,219	3,219	1,022
d'^2_{hi}	389,433	–3,735	–0,261
d''^2_{hi}	392,652	0,516	0,761
$\mathrm{P}(\mathrm{A}h\lvert\boldsymbol{x}_i)$	0,00	0,65	0,35

Enfin, pour les méthodes non paramétriques, l'extension au cas de probabilités *a priori* inégales est immédiate également :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{p_h f_h(\boldsymbol{x}_i)}{\sum_{k=1}^{g} p_k f_k(\boldsymbol{x}_i)} .$$

En particulier, pour la méthode des r plus proches voisins, si les probabilités *a priori*, p_h, sont proportionnelles aux effectifs des échantillons :

$$p_h = \frac{n_h}{n} ,$$

et si les matrices de variances et covariances sont égales, les probabilités *a posteriori* s'estimeront de façon particulièrement simple :

$$\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) = \frac{r_h}{r} .$$

A titre d'exemple, reprenons l'iris dont la longueur et la largeur des pétales valent respectivement 4,7 et 1,6. En considérant des probabilités *a priori* identiques et des matrices de variances et covariances inégales, nous avons vu que cet iris est classé dans l'espèce *versicolor* (paragraphe 5.1). Supposons maintenant que des études antérieures aient montré que l'espèce *virginica* est trois fois plus abondante que les espèces *versicolor* et *setosa*. On en déduirait les résultats repris dans le tableau 5 et l'iris serait toujours classé dans l'espèce versicolor, mais la probabilité *a posteriori* ne serait que de 0,65, alors que pour des probabilités *a priori* égales, elle était de 0,85.

7. VALIDATION DES RÉSULTATS

7.1. Définition des taux d'erreur

A l'issue de l'établissement d'une règle de classement, on souhaite en général disposer d'informations relatives à la qualité de la règle qui vient d'être établie.

Nous avons, dans les paragraphes précédents, fait allusion à la notion de probabilité *a posteriori* d'appartenance à chacune des populations d'un individu particulier, caractérisé par un vecteur d'observations. Ces probabilités mesurent le risque de classement erroné de cet individu et ne donnent pas d'information concernant le risque de classement erroné pour un individu quelconque.

En relation avec le problème de classement erroné, plusieurs taux d'erreurs correspondant à des concepts différents peuvent être définis.

Le premier taux est le *taux d'erreur optimal*[11]. Il correspond au taux d'erreur quand une règle d'affectation basée sur les paramètres réels des populations est appliquée à ces populations. Ainsi, pour deux populations normales de même matrice de variances et covariances et pour des probabilités d'appartenance *a priori* égales, ce taux est fonction de la distance de MAHALANOBIS qui sépare les centres de gravité des deux populations (paragraphe 3.2).

En pratique, les vecteurs des moyennes et la matrice des variances et covariances sont inconnus et la distance doit être estimée. L'utilisation de la relation suivante :

$$\hat{\Delta}^2 = (\hat{\boldsymbol{m}}_1 - \hat{\boldsymbol{m}}_2)\, \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{m}}_1 - \hat{\boldsymbol{m}}_2)\,,$$

conduit à une surestimation de Δ^2 et donc aussi à une sous-estimation du taux optimal. On peut cependant éliminer cette sous-estimation en utilisant une estimation à peu près non biaisée de la distance [DAGNELIE, 1975; HUBERTY, 1994] :

$$\hat{\Delta}_c^2 = \frac{n_1 + n_2 - p - 3}{n_1 + n_2 - 2}\, \hat{\Delta}^2 - \frac{(n_1 + n_2)p}{n_1\, n_2}\,.$$

Des extensions existent pour le cas où les probabilités *a priori* sont inégales. Par contre, aucune formule n'a été proposée pour le cas de plus de deux populations [HUBERTY, 1994].

Un deuxième taux d'erreur est le *taux d'erreur réel*[12] qui est le taux d'erreur obtenu en appliquant une règle d'affectation, calculée à partir de g échantillons particuliers à d'autres individus provenant du même mélange de populations.

Ce deuxième type de taux d'erreur est le plus utile en pratique, puisqu'il donne le taux d'erreur auquel on doit s'attendre lors de l'utilisation d'une règle préalablement définie.

Une formule de calcul de ce taux d'erreur existe pour le seul cas de deux populations normales [HUBERTY, 1994] et nous examinerons, au paragraphe 7.2, quelques méthodes non paramétriques d'estimation de ce taux d'erreur.

11. En anglais : *optimal error rate.*
12. En anglais : *actual error rate.*

Le troisième taux d'erreur est le *taux d'erreur réel attendu*[13]. Il s'agit de l'espérance mathématique du taux réel, pour tous les échantillons qui auraient pu être prélevés dans les populations. La détermination de ce taux d'erreur peut présenter un intérêt avant le prélèvement d'un échantillon particulier dans chaque population.

Il relie en effet la performance moyenne d'une règle à la taille des échantillons servant à la construire et permet par conséquent de déterminer l'ordre de grandeur de l'effectif à prélever. Pour deux populations normales de même matrice de variances et covariances, et pour des probabilités *a priori* égales, une formule donnant le taux d'erreur réel attendu est proposée par McLachlan [1992].

7.2. Estimations du taux d'erreur

Dans le paragraphe précédent, nous avons défini trois taux d'erreur et signalé que pour deux populations normales de même matrice de variances et covariances, les deux premiers taux pouvaient être estimés par des formules. Nous examinons ci-dessous des solutions plus générales pour estimer le taux d'erreur lié à une règle de décision qui a été établie à partir de g échantillons donnés.

De façon générale, on peut définir le taux de classement erroné à partir des probabilités de classement erroné relatives à chacune des populations.

Si $\mathrm{P}(\mathrm{C}\bar{h}|\mathrm{A}h)$ représente la probabilité de classer un individu appartenant à la population h dans une autre population et si $\mathrm{P}(\mathrm{C}l|\mathrm{A}h)$ représente la probabilité de classer un individu de la population h dans la population l, on a :

$$\mathrm{P}(\mathrm{C}\bar{h}|\mathrm{A}h) = \sum_{l \neq h}^{g-1} \mathrm{P}(\mathrm{C}l|\mathrm{A}h)\,,$$

la sommation des $\mathrm{P}(\mathrm{C}l|\mathrm{A}h)$ étant étendue à toutes les populations, à l'exclusion de la population h.

Le taux de classement erroné est alors donné par :

$$\sum_{k=1}^{g} p_h \, \mathrm{P}(\mathrm{C}\bar{h}|\mathrm{A}h)\,,$$

p_h étant la probabilité *a priori* d'appartenance d'un individu à la population h.

En pratique, les probabilités $\mathrm{P}(\mathrm{C}l|\mathrm{A}h)$ doivent être estimées et plusieurs solutions sont possibles.

Une première solution est basée sur la méthode de resubstitution. Elle consiste à réaffecter les n observations selon la règle de classement définie et à dénombrer les classements erronés. On peut alors estimer les probabilités $\mathrm{P}(\mathrm{C}l|\mathrm{A}h)$ à partir des fréquences relatives des observations de la population h reclassées dans la population l. Si les probabilités *a priori*, p_h, sont proportionnelles aux

13. En anglais : *expected actual error rate*.

effectifs des échantillons, le taux de classement erroné est directement donné par la proportion globale des individus mal classés.

Le taux d'erreur ainsi estimé est parfois appelé taux d'*erreur apparent* ou encore *taux d'erreur de resubstitution*[14].

Cette méthode est très simple mais conduit à des résultats trop optimistes, car ce sont les mêmes données qui sont utilisées pour l'établissement de la règle de classement et pour la validation [Albert et Harris, 1987; Kendall *et al.*, 1983].

Pour pallier cet inconvénient, on peut établir la règle de classement sur un premier ensemble de données et valider la méthode sur un autre ensemble de données. Le taux d'erreur mesuré sur l'échantillon-test sera alors une estimation sans biais du taux réel. Cette technique suppose cependant qu'on dispose de données suffisamment nombreuses afin de pouvoir en réserver à la validation.

Lorsqu'on ne dispose pas de données assez nombreuses, on peut estimer les taux d'erreur par la méthode d'*extraction et insertion*[15], qui consiste à effectuer n fois l'analyse discriminante sur $n-1$ observations en mettant de côté tour à tour chacune des observations et en reclassant cette observation sur la base de la règle établie sur les $n-1$ autres données. Cette méthode est pratiquement non biaisée mais la variance des estimations est plus grande. On peut affirmer, en outre, que la proportion d'individus bien classés par la méthode d'extraction et insertion est toujours plus petite ou égale que la proportion d'individus bien classés par la méthode de resubstitution [Albert et Harris, 1987].

La mesure de la qualité d'une règle de classement peut également être basée sur les probabilités *a posteriori*. L'idée est de ne pas simplement tenir compte du fait que les individus sont bien ou mal classés, mais de tenir compte de la fiabilité du classement, mesurée par les probabilités *a posteriori*.

Ainsi, on peut estimer le taux d'erreur relatif à la population h par la relation :

$$\hat{e}_h = 1 - \frac{1}{n\,p_h} \sum_{i=1}^{n_h^{'}} [\mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i)] \ .$$

Dans cette relation, l'indice i, qui varie de 1 à $n_h^{'}$, signifie que l'on somme les probabilités *a posteriori* des $n_h^{'}$ individus qui sont classés dans la population h.

Le produit $n\,p_h$ est une estimation de la fréquence attendue d'individus appartenant à la population h si on prélève un échantillon d'effectif n dans le mélange des populations.

Une estimation globale du taux d'erreur est alors obtenue en calculant une moyenne des taux d'erreur par population, ces taux étant pondérés par les probabilités *a priori* :

$$\hat{e} = \sum_{h=1}^{g} p_h \hat{e}_h = 1 - \frac{1}{n} \sum_{h=1}^{g} \sum_{i=1}^{n_h^{'}} \mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i) \, .$$

14. En anglais : *apparent error rate* ou *resubstitution error rate*.
15. En anglais : *leaving one out*.

Il s'agit donc de retrancher de l'unité la moyenne, calculée sur les n individus, des probabilités d'appartenance au groupe dans lesquels ils sont classés. Ces probabilités d'appartenance peuvent être calculées, soit par la méthode de resubstitution, soit par la méthode d'extraction et insertion, soit encore sur un échantillon-test. On notera que le calcul de ce paramètre ne nécessite pas la connaissance de l'appartenance exacte des individus à l'une des populations et il peut donc se faire sur un échantillon d'individus d'origine inconnue, n représentant alors le nombre d'individus à classer. Ce dernier point est particulièrement intéressant si on dispose d'un grand nombre d'individus d'origine inconnue, comme c'est le cas, par exemple, dans le domaine de la télédétection.

Une autre estimation du taux d'erreur tient compte en outre de l'appartenance exacte des individus aux différents groupes :

$$(\hat{e}_h)_s = 1 - \frac{1}{p_h} \sum_{k=1}^{g} p_k \frac{1}{n_k} \sum_{i=1}^{n'_{kh}} \mathrm{P}(\mathrm{A}h|\boldsymbol{x}_i)\,,$$

n'_{kh} étant l'ensemble des individus appartenant à la population k qui sont classés dans la population h. L'estimation du taux d'erreur global s'obtient alors en calculant la moyenne des taux d'erreur par population, pondérée par les probabilités *a priori* :

$$(\hat{e})_s = \sum_{h=1}^{g} p_h (\hat{e}_h)_s\,.$$

Le logiciel SAS donne les différents taux basés sur les probabilités *a posteriori* qui viennent d'être définis. Les premiers, $\hat{e}_h$ et $\hat{e}$, sont appelés estimations non stratifiées et les seconds, $(\hat{e}_h)_s$ et $(\hat{e})_s$, sont appelés estimations stratifiées [SAS, 1989]. On notera que ces différents taux sont identiques lorsque les probabilités *a priori* d'appartenance à un groupe sont proportionnelles aux tailles des échantillons.

On notera aussi que les taux d'erreur estimés par groupe peuvent être négatifs, notamment quand il y a de fortes discordances entre les probabilités *a priori* et les tailles relatives des échantillons. Pour avoir des estimations fiables des taux d'erreur par groupe, les tailles relatives des groupes doivent être, au moins de façon approximative, proportionnelles aux probabilités *a priori* [SAS, 1989].

La figure 9 donne les résultats du classement par la méthode de resubstitution des 150 iris sur base de la longueur et de la largeur des pétales, l'analyse étant réalisée sur la base de matrices de variances et covariances égales.

Les taux d'erreur estimés à partir des comptages des iris mal classés sont égaux à 4 % pour *Iris versicolor*, à 8 % pour *Iris virginica* et à 0 % pour *Iris setosa*, soit, en moyenne, un taux d'erreur de 4 %.

Si on tient compte des probabilités *a posteriori*, on retrouve les taux d'erreur donnés dans la figure. Ainsi, pour *Iris virginica*, par exemple, on a, en l'absence de stratification :

$$\hat{e}_{virg} = 1 - \frac{1}{(150)(0,3333)}\,[(2)(0,7098) + (46)(0,9739] = 0,0757 \text{ ou } 7,57\ \%\,.$$

Number of Observations and Average Posterior Probabilities
Classified into ESPECE:

From ESPECE	1	2	3
1	50	0	0
	1.0000	.	.
2	0	48	2
	.	0.9765	0.7098
3	0	4	46
	.	0.7784	0.9739
Total	50	52	48
	1.0000	0.9613	0.9629
Priors	0.3333	0.3333	0.3333

Posterior Probability Error Rate Estimates for ESPECE:

Estimate	1	2	3	Total
Stratified	0.0000	0.0002	0.0757	0.0253
Unstratified	0.0000	0.0002	0.0757	0.0253
Priors	0.3333	0.3333	0.3333	

Figure 9. Résultat du classement (par resubstitution) sur la base de la longueur et de la largeur des pétales (analyse discriminante linéaire): résultats obtenus par le logiciel SAS.

et le taux d'erreur global est égal à 0,0253 ou 2,53 %.

Pour cet exemple, les taux obtenus avec et sans stratification sont égaux car les proportions d'individus dans les groupes sont égales aux probabilités *a priori*.

D'autre part, les résultats du reclassement des iris par la méthode d'extraction-insertion sont identiques aux résultats de la figure 9 en ce qui concerne la répartition des iris, mais les probabilités moyennes *a posteriori* dans les différentes cellules du tableau sont légèrement différentes. Il en résulte que les taux d'erreur estimés à partir des comptages sont les mêmes que ceux donnés ci-dessus. Par contre, de petites différences s'observent pour les taux d'erreurs calculés à partir des probabilités *a posteriori* et le taux d'erreur global est égal à 2,41 %.

Si on réalise l'analyse discriminante linéaire à partir des quatre variables, les taux d'erreurs sont légèrement plus faibles. En effet, seuls trois iris sont mal classés, à la fois pour la méthode de resubstitution et pour la méthode d'extraction et insertion et les taux d'erreur calculés à partir des probabilités *a posteriori* sont respectivement égaux à 1,62 et à 1,79 pour la méthode de resubstitution et pour la méthode d'extraction-insertion. La prise en considération des deux variables relatives aux sépales n'améliore en définitive guère le pouvoir discriminant, par rapport à la prise en compte des seules caractéristiques des pétales.

8. QUELQUES RECOMMANDATIONS PRATIQUES

8.1. Choix d'une règle d'affectation

Nous avons vu, dans les paragraphes précédents, que l'analyse discriminante linéaire repose sur la normalité à p dimensions des g populations et sur l'égalité des matrices de variances et covariances de ces g populations. Lorsque ces conditions d'application sont remplies, l'analyse discriminante linéaire est la méthode à préconiser, car elle présente le taux d'erreur optimal le plus faible.

L'utilisateur peut vérifier ces conditions d'application. Ainsi, si les échantillons de chacune des populations sont de taille suffisante, le caractère multinormal de la population peut être apprécié par l'examen des distances de MAHALANOBIS des individus d'un échantillon au centre de gravité de cet échantillon. En effet, pour des échantillons suffisamment grands et pour des nombres de variables suffisamment élevés, la distribution des carrés des distances est approximativement une distribution χ^2 et la distribution des distances est donc approximativement normale [DAGNELIE, 1975]. Cette normalité des distances peut être testée par les méthodes univariées classiques, comme par exemple le test de SHAPIRO et WILK proposé par SAS [1989] ou le test d'ANDERSON et DARLING, proposé par Minitab [X, 1994].

Pour des populations multinormales, l'égalité des matrices de variances et covariances peut être vérifiée par une généralisation du test de BARTLETT [DAGNELIE, 1975]. Ce test est proposé dans la procédure DISCRIM de SAS [SAS, 1989].

Si les populations sont normales mais n'ont pas des matrices de variances et covariances identiques, le taux d'erreur optimal le plus faible est obtenu par l'analyse discriminante quadratique. Cela ne signifie cependant pas que le taux d'erreur réel soit plus faible. En effet, la synthèse des diverses études comparatives montre que la supériorité de l'analyse discriminante quadratique n'est manifeste que si les matrices de variances et covariances sont très différentes et que les échantillons sont de taille élevée par rapport au nombre de variables. SEBER [1984] cite, par exemple, les valeurs de $n_h \geq 25$ pour $p = 4$ et $g = 2$ et HUBERTY [1994] recommande que n_h soit supérieur à $5p$.

De façon plus générale, diverses études ont été réalisées pour vérifier la robustesse de l'analyse discriminante à la non-normalité et à l'inégalité des matrices de variances et covariances. Une synthèse est proposée par LAZAAR [1993]. De ces études, il ressort que l'analyse linéaire discriminante est relativement robuste vis-à-vis de violations modérées à la fois des conditions de normalité et d'inégalité de matrices de variances et covariances. L'étude de LAZAAR et PALM [1994] confirme ces conclusions.

Pour des populations nettement non normales ou pour des variables qualitatives, l'utilisateur peut se tourner vers d'autres méthodes de discrimination qui ont été présentées sommairement au paragraphe 5.

Une autre solution consiste à réaliser des transformations de variables permettant de se rapprocher de la conditions de normalité. Dans la même optique, pour des variables continues, on peut remplacer les observations par leur rang et utiliser les méthodes de discrimination relatives aux variables normales [HUBERTY, 1994].

Pour les exemples considérés précédemment, la normalité à p dimensions n'a pas été vérifiée. Toutefois, nous avons réalisé des tests de normalité à une dimension sur chacune des variables au sein de chacune des trois espèces d'iris. Les douze tests de SHAPIRO et WILK ainsi réalisés conduisent à l'acceptation de l'hypothèse de normalité, au niveau 0,05, à deux exceptions près, qui concernent la largeur des pétales pour les espèces *setosa* et *versicolor*. La non-normalité est surtout marquée pour *Iris setosa*, la probabilité associée au test étant de 0,0001. Comme cette espèce se distingue de toute manière très nettement des deux autres, la non-normalité reste assez secondaire. Pour *Iris versicolor*, la non-normalité est moins marquée, la probabilité associée au test étant égale à 0,02. Bien que la normalité des p variables n'implique pas la normalité à p dimensions, on a de bonnes raisons de croire que la non-normalité n'est pas particulièrement accentuée et ne constitue certainement pas un problème majeur pour l'exemple considéré.

Quant à l'égalité des matrices de variances et covariances, les tests réalisés dans le cas de deux populations et deux variables (exemple présenté au paragraphe 3) de trois populations et deux ou quatre variables (exemples présentés aux paragraphes 4 et 7) conduisent tous à la conclusion qu'il existe des différences très hautement significatives entre les matrices de variances et covariances. L'utilisation de l'analyse discriminante quadratique pourrait donc se justifier, compte tenu notamment du nombre réduit de variables utilisées et du nombre élevé d'individus dans les groupes.

Une analyse plus détaillée montre cependant que, dans le cas de deux variables (longueur et largeur des pétales), l'analyse discriminante quadratique ne présente guère d'avantages par rapport à l'analyse discriminante linéaire. En effet, l'analyse discriminante quadratique conduit à 3 individus mal classés par la méthode de resubstitution, contre 6 individus mal classés pour l'analyse discriminante linéaire. Mais par extraction-insertion, ce nombre est de 5, contre 6 pour l'analyse discriminante linéaire. D'autre part, le taux d'erreur calculé à partir des probabilités *a posteriori* dans le cas de l'extraction-insertion est plus faible pour l'analyse discriminante linéaire (2,42 %) que pour l'analyse discriminante quadratique (3,01 %).

8.2. Choix des probabilités *a priori*

Rappelons que la probabilité *a priori* d'un groupe donné correspond à la probabilité qu'a un individu particulier d'appartenir à ce groupe, avant qu'on ait pris connaissance des observations réalisées sur cet individu.

Comme signalé au paragraphe 6, les probabilités *a priori* interviennent dans le classement des individus et l'utilisateur doit décider s'il accepte des probabilités *a priori* égales ou, au contraire, s'il souhaite utiliser des probabilités différentes.

Lorsque l'utilisateur sait que les différents groupes sont inégalement représentés, il devrait prendre en considération des probabilités *a priori* reflétant l'inégale représentation de ces groupes.

La proportion de chacun des groupes dans la population globale peut être connue par des études antérieures, ou être estimée par les proportions observées dans l'échantillon, à condition que cet échantillon ait été prélevé au hasard dans la population globale.

8.3. Choix d'une mesure de la validité de la règle

Différentes approches ont été proposées au paragraphe 7. Rappelons que les méthodes basées sur la resubstitution sont trop optimistes et doivent donc être évitées, au profit des méthodes basées sur le principe de l'extraction et insertion.

L'estimation du taux d'erreur à partir du comptage des individus mal classés par la technique d'extraction et insertion est non biaisée, mais possède une grande variance. La prise en compte des probabilités *a posteriori* permet de réduire cette variance, pour autant que ces probabilités soient correctes.

On peut donc recommander l'estimation du taux d'erreur par la prise en compte des probabilités *a posteriori*, calculées par extration et insertion sauf si les conditions d'application de la méthode d'analyse discriminante utilisée sont manifestement non remplies (analyse discriminante linéaire sur données nettement non normales, par exemple). Dans ce dernier cas, il est préférable d'estimer le taux d'erreur à partir du comptage des individus mal classés par extraction et insertion.

Enfin, lors de la prise en compte des probabilités *a posteriori* pour estimer le taux d'erreur global, se présente encore le choix entre estimation non stratifiée et estimation stratifiée. A ce sujet HUBERTY [1994] signale que ce choix dépend de la confiance qu'a l'utilisateur dans les probabilités *a priori* qui sont prises en compte : si celles-ci sont basées sur une bonne connaissance de l'importance relative des groupes, les estimations stratifiées sont préférables. Rappelons également que les deux méthodes donnent les mêmes résultats lorsque les probabilités *a priori* sont proportionnelles à la taille des échantillons.

BIBLIOGRAPHIE

ALBERT A., HARRIS E.K. [1987]. *Multivariate interpretation of clinical laboratory.* New York, Dekker, 312 p.

CELEUX G., NAKACHE J.P. [1994]. *Analyse discriminante sur variables qualitatives.* Paris, Polytechnica, 270 p.

DAGNELIE P. [1975]. *Analyse statistique à plusieurs variables.* Gembloux, Presses Agronomiques, 362 p.

DAGNELIE P. [1998]. *Statistique théorique et appliquée.* Tome 1 : *Statistique descriptive et bases de l'inférence statistique.* Bruxelles, De Boeck Université, 508 p.

DEBOUCHE C., MARQUET J.P., TEUGELS H. [1979]. Détermination de poissons du genre Clarias par une méthode généralisable aux poissons non écailleux. *Bull. Inst. Fr. Afr. Noire* 41, 844-862.

FISHER R.A. [1936]. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179-188.

HAND D.J. [1981]. *Discrimination and classifiction.* New York, Wiley, 209 p.

HAND D.J., DALY F., LUNN A.D., MCCONWAY K.J., OSTROWSKI E. [1994]. (eds). *A handbook of small data sets.* London, Chapman and Hall, 458 p.

HUBERTY C.J. [1994]. *Applied discriminant analysis.* New York, Wiley, 466 p.

KENDALL M.G., STUART A., ORD J.K. [1983]. *The advanced theory of statistics* (vol. 3). London, Griffin, 780 p.

LAZAAR N. [1993]. *Etude comparative de méthodes d'analyse discriminante* (travail de fin d'études). Gembloux, Faculté des Sciences agronomiques, 84 p.

LAZAAR N., PALM R. [1994]. Comparaison pratique de quelques méthodes d'analyse discriminante. *Biom. Praxim.* 34, 3-4, 207-222.

MCLACHLAN G.J. [1992]. *Discriminant analysis and statistical pattern recognition.* New York, Wiley, 544 p.

PALM R. [1990]. La corrélation canonique: principes et application. *Notes stat. Inform.* (Gembloux) 90/1, 28 p.

SAPORTA G. [1990]. *Probabilités, analyse des données et statistique.* Paris, Technip, 493 p.

SAS INSTITUTE INC [1989]. *SAS/STAT User's guide, version 6*, Fourth edition (2 volumes). Cary NC: SAS Institute Inc. 943 + 946 p.

SEBER G.A.F. [1984]. *Multivariate observations.* New York, Wiley, 686 p.

X [1994]. *Minitab reference manual, release 10 for Windows.* PA State College, Minitab 984 p.