Méthodes statistiques de discrimination

Gilbert Saporta

CHAPITRE 2

MÉTHODES STATISTIQUES DE DISCRIMINATION

Gilbert Saporta

2.1 Introduction

Le but des méthodes de discrimination consiste à prédire une variable qualitative à k catégories à l'aide de $p$prédicteurs, généralement numériques.

La plupart du temps, le problème se pose de la manière suivante. On connaît sur un échantillon de n observations les groupes d'appartenance ainsi que les variables explicatives. L'exemple des infarctus (Nakache repris par Saporta, 1990) en est une bonne illustration : 101 malades sont décrits par 6 variables médicales et divisés en deux groupes, les survivants et les décédés.

À partir de cet échantillon, on cherche des fonctions des variables explicatives permettant d'affecter, avec une probabilité d'erreur minimale, les observations dans les groupes. Ces fonctions de classement peuvent être explicites (linéaires, quadratiques ou logiques) ou implicites comme en estimation de densité.

De très nombreuses méthodes existent, tant géométriques que probabilistes. Nous nous limiterons ici aux méthodes utilisées dans les logiciels les plus courants, renvoyant le lecteur à la bibliographie pour la régression logistique ou la discrimination par arbre.

2.2 Méthodes géométriques de discrimination sur données numériques

2.2.1 Données et notation

Les n individus $\mathbf{e}_i$ de l'échantillon constituent un nuage E, de R^p partagé en k sous-nuages $E_1, E_2, \ldots, E_k$ de centres de gravité $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ de matrices de variances $\mathbf{V_1}, \mathbf{V_2}, \ldots, \mathbf{V_k}$ (figure 1). Soit g le centre de gravité et $\mathbf{V}$ la matrice de variance de E tout entier. Si les n individus e_i sont affectés des poids $p_1, p_2, \ldots, p_n$, les poids $q_1, q_2, \ldots, q_k$ de chaque sous-nuage sont alors :

$$q_j = \sum_{\mathbf{e}_i \in E_j} p_i$$

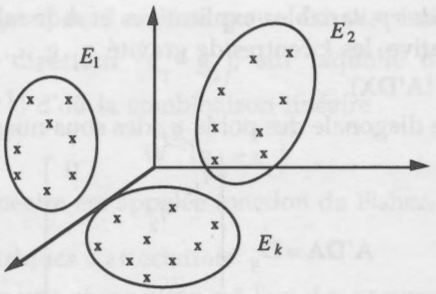

Figure 1. Espace des données

On a :
$$\mathbf{g}_j = \frac{1}{q_j} \sum_{\mathbf{e}_i \in E_j} p_i \mathbf{e}_i \text{ pour } \mathbf{e}_i \in E$$

$$\mathbf{g} = \sum_{j=1}^{k} q_j \mathbf{g}_j \text{ et } \mathbf{V}_j = \frac{1}{q_j} \sum_{\mathbf{e}_i \in E_j} p_i \left(\mathbf{e}_i - \mathbf{g}_j\right)\left(\mathbf{e}_i - \mathbf{g}_j\right)'$$

Appelons matrice de variance interclasse, la matrice de variance **B** des k centres de gravité affectés des poids q_j :

$$\mathbf{B} = \sum_{j=i}^{k} q_j \left(\mathbf{g}_j - \mathbf{g}\right)\left(\mathbf{g}_j - \mathbf{g}\right)'$$

et matrice de variance intraclasse **W** la moyenne des matrices $\mathbf{V}_j$:

$$\mathbf{W} = \sum_{j=1}^{k} q_j \mathbf{V}_j$$

En règle générale, **W** est inversible tandis que **B** ne l'est pas, car les k centres de gravité sont dans un sous-espace de dimension k - 1 de R^p (si p>k-1 ce qui est généralement le cas), alors que la matrice **B** est de taille p.

On a alors la relation suivante :

$$\mathbf{V} = \mathbf{W} + \mathbf{B}$$

qui se démontre aisément et constitue une généralisation de la relation classique : variance totale = moyenne des variances + variance des moyennes.

Nous supposerons désormais que **g**=0, c'est-à-dire que les variables explicatives sont centrées.

Si on considère que le tableau de données à étudier se met sous la forme :

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \ldots & k \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ & & & \\ & & \mathbf{A} & \\ & & & \\ 0 & 0 & \ldots & 1 \end{array}\right. \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \ldots & p \end{array} \\ \left.\begin{array}{ccc} & & \\ & & \\ & & \\ & \mathbf{X} & \\ & & \\ & & \end{array}\right] \end{array}$$

où **X** est la matrice des p variables explicatives et **A** le tableau logique associé à la variable qualitative, les k centres de gravité $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ sont les lignes de la matrice $(\mathbf{A'DA})^{-1}(\mathbf{A'DX})$.

A'DA est la matrice diagonale des poids q_j des sous-nuages :

$$\mathbf{A'DA} = \mathbf{D}_q = \begin{bmatrix} q_1 & & & 0 \\ & q_2 & & \\ & & \ddots & \\ 0 & & & q_k \end{bmatrix}$$

La matrice de variance interclasse s'écrit alors, si $\mathbf{g} = 0$:

$$\mathbf{B} = ((\mathbf{A'DA})^{-1}\mathbf{A'DX})'\mathbf{A'DA}((\mathbf{A'DA})^{-1}\mathbf{A'DX})$$
$$= \mathbf{X'DA}(\mathbf{A'DA}) - 1\mathbf{A'DX} = (\mathbf{X'DA})\mathbf{D}_q^{-1}(\mathbf{A'DX})$$

Dans le cas où $p_i = 1/n$ les expressions précédentes se simplifient et en introduisant les effectifs $n_1, n_2, \ldots, n_k$ des k sous-nuages, on a :

$$\mathbf{B} = \frac{1}{n}\sum_j n_j \mathbf{g}_j \mathbf{g}_j' \quad ; \quad \mathbf{g}_j = \frac{1}{n_j}\sum_{E_j} \mathbf{e}_i \quad ; \quad \mathbf{W} = \frac{1}{n}\sum_j n_j \mathbf{V}_j$$

Nous supposerons désormais être dans ce cas.

2.2.2 Axes discriminants

L'analyse factorielle discriminante, appelée en anglais *canonical discriminant analysis* consiste à chercher des axes sur lesquels on projette les observations de telle sorte que :

- les centres des groupes soient projetés avec la dispersion maximale,
- les projections des observations de chaque groupe soient en moyenne peu dispersées.

Ceci revient donc à chercher une combinaison linéaire **u** telle que :

u'Bu soit maximal

et **u'Wu** soit minimal

Cette simultanéité étant impossible, on choisit comme critère de maximiser **u'Bu/u'Wu** ou, ce qui est équivalent, **u'Bu/u'Vu**.

Les combinaisons linéaires associées aux axes sont alors les vecteurs propres de $\mathbf{W}^{-1}\mathbf{B}$ ou de $\mathbf{V}^{-1}\mathbf{B}$.

On voit alors que l'analyse factorielle discriminante est équivalente à une ACP des centres de gravité avec soit la métrique $\mathbf{V}^{-1}$, soit la métrique $\mathbf{W}^{-1}$. C'est cette dernière que l'on appelle métrique de Mahalanobis.

S'il n'y a que deux groupes, il n'existe qu'un axe discriminant, la droites des centres, de vecteur directeur $(\mathbf{g}_1 - \mathbf{g}_2)$, sur laquelle on projette avec la métrique $\mathbf{W}^{-1}$(ou $\mathbf{V}^{-1}$) d'où la combinaison linéaire :

$$\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

cette combinaison linéaire est appelée fonction de Fisher.

2.2.3 Règles géométriques d'affectation

On cherche à affecter une observation **e** à l'un des groupes.

La règle naturelle consiste à calculer les distances de l'observation à classer à chacun des *k* centres de gravité et à affecter selon la distance la plus faible. Encore faut-il définir la métrique à utiliser.

Règle de Mahalanobis-Fisher

Elle consiste à utiliser la métrique $\mathbf{W}^{-1}$ (ou $\mathbf{V}^{-1}$ ce qui est équivalent) :

$$d^2(\mathbf{e},\mathbf{g}_i) = (\mathbf{e}-\mathbf{g}_i)' \mathbf{W}^{-1}(\mathbf{e}-\mathbf{g}_i)$$

En développant cette expression on trouve :

$$d^2(\mathbf{e},\mathbf{g}_i) = \mathbf{e}'\mathbf{W}^{-1}\mathbf{e} + \mathbf{g}_i' \mathbf{W}^{-1}\mathbf{g}_i - 2\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_i$$

Comme $\mathbf{e}'\mathbf{W}^{-1}\mathbf{e}$ ne dépend pas du groupe *i*, la règle consiste donc à chercher le minimum de $\mathbf{g}_i' \mathbf{W}^{-1}\mathbf{g}_i - 2\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_i$

ou le maximum de $\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_i - (\mathbf{g}_i'\mathbf{W}^{-1}\mathbf{g}_i)/2$.

On voit que cette règle est linéaire par rapport aux coordonnées de **e**.

Il faut donc calculer pour chaque individu *k* fonctions linéaires de ces coordonnées et en chercher la valeur maximale.

Cas de deux groupes

Pour deux groupes la règle devient :

$$\mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_1 - \frac{1}{2}(\mathbf{g}_1' \mathbf{W}^{-1}\mathbf{g}_1) > \mathbf{e}'\mathbf{W}^{-1}\mathbf{g}_2 - \frac{1}{2}\mathbf{g}_2' \mathbf{W}^{-1}\mathbf{g}_1$$

soit : $$\mathbf{e}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2) > \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)' \mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$$

On retrouve en $\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ la fonction de Fisher.

Lorsque les deux groupes sont de même effectif $\mathbf{g}_1 + \mathbf{g}_2 = 0$; on affectera alors au groupe 1 si la fonction $\mathbf{e}'\mathbf{W}^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$ est positive.

Équivalence entre régression multiple et discrimination entre deux groupes

Soit $\mathbf{y}$, un vecteur à n composantes tel que $y_i = a$ pour un individu de groupe 1 et $y_i = b$ pour un individu de groupe 2.

Si l'on effectue une régression multiple de $\mathbf{y}$ sur $\mathbf{x}$, on obtient alors un vecteur des coefficients de régression proportionnel à la fonction de Fisher pour un choix quelconque de a. Le choix $a = \frac{n}{n_1}$, $b = -\frac{n}{n_2}$ conduit alors à

$$b = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{V}^{-1}\left(\mathbf{g}_1 - \mathbf{g}_2\right)$$

La qualité de cette régression est mesurée par le coefficient R :

$$R^2 = \frac{D_p^2}{\frac{n(n-2)}{n_1 n_2} + D_p^2} = \frac{\text{variance de la régression}}{\text{variance de } \mathbf{y}}$$

où D_p^2 est la distance de Mahalanobis estimée par :

$$D_p^2 = \frac{n-2}{n}\left(\mathbf{g}_1 - \mathbf{g}_2\right)' \mathbf{W}^{-1}\left(\mathbf{g}_1 - \mathbf{g}_2\right)$$

On prendra garde au fait que les hypothèses habituelles de la régression ne sont pas vérifiées bien au contraire : ici $\mathbf{y}$ est non aléatoire et $\mathbf{x}$ l'est. Il ne faudra donc pas utiliser, autrement qu'à titre indicatif, les statistiques usuelles fournies par un programme de régression.

2.2.4 Insuffisance des règles géométriques

L'utilisation de la règle précédente conduit à des affectations incorrectes lorsque les dispersions des groupes sont très différentes entre elles : rien ne justifie alors l'usage de la même métrique pour les différents groupes.

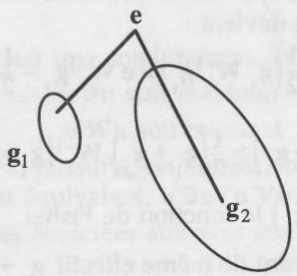

Figure 2. Problème de l'affectation

En effet, si on considère la figure 2, bien que $\mathbf{e}$ soit plus proche de $\mathbf{g}_1$ que de $\mathbf{g}_2$ au sens habituel il est plus naturel d'affecter $\mathbf{e}$ à la deuxième classe qu'à la première dont le « pouvoir d'attraction » est moindre.

Diverses solutions utilisant des métriques locales $\mathbf{M_i}$ telles que :

$$d^2(\mathbf{e}, \mathbf{g}_i) = (\mathbf{e} - \mathbf{g}_i)' \mathbf{M}_i (\mathbf{e} - \mathbf{g}_i)$$

ont été proposées, la plupart prenant $\mathbf{M_i}$ proportionnel à $\mathbf{V}_i^{-1}$. La question de l'optimalité d'une règle de décision géométrique ne peut cependant être résolue sans référence à un modèle probabiliste. En effet le problème est de savoir comment cette règle se comportera pour de nouvelles observations ce qui impose de faire des hypothèses distributionnelles sur la répartition dans l'espace de ces nouvelles observations. On atteint donc ici les limites des méthodes descriptives. Nous verrons plus loin dans quelles conditions elles conduisent à des règles optimales.

2.3 Une méthode de discrimination sur variables qualitatives : la méthode DISQUAL

Lorsque les prédicteurs sont p variables qualitatives $x_1, x_2, \ldots, x_p$ à $m_1, m_2, \ldots, m_p$ modalités respectivement, on peut utiliser la procédure suivante : on effectue dans un premier temps l'analyse des correspondances multiples des variables $x_1, x_2, \ldots, x_p$ c'est-à-dire l'analyse des correspondances du tableau disjonctif $\mathbf{X} = (\mathbf{X}_1 | \mathbf{X}_2 | \ldots | \mathbf{X}_p)$.

On remplace alors les p variables qualitatives par les q coordonnées sur les axes factoriels et on effectue ensuite une analyse discriminante sur ces q variables numériques $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_q$.

Une fonction discriminante **d** est une combinaison linéaire des z_j qui sont des combinaisons linéaires des indicatrices des x_i. On exprime alors directement **d** comme combinaison linéaire des indicatrices des x_i ce qui revient à attribuer à chaque catégorie de chaque variable une valeur numérique ou *score*. **d** est alors simplement égal à l'addition des *scores* obtenus dans les catégories des p variables. Ceci revient donc à transformer chaque variable qualitative x_i en une variable discrète à mj valeurs.

Lorsque k=2 cette méthode est optimale au sens suivant : en conservant tous les facteurs de l'analyse des correspondances multiples la quantification des variables x_i est celle qui donne la distance de Mahalanobis la plus grande entre les deux groupes.

En pratique, cependant, on n'utilisera que les facteurs présentant à la fois une inertie et un pouvoir séparateur entre les classes suffisant. L'avantage de l'analyse des correspondances est de fournir outre une description des liaisons entres les variables explicatives, des composantes orthogonales.

2.4 Méthodes probabilistes

2.4.1 La règle bayesienne

On suppose que les k groupes sont en proportion $p_1, p_2, \ldots, p_k$ dans la population totale et que la distribution de probabilité du vecteur observation $\mathbf{x} = (x_1, \ldots, x_p)$ est donnée pour chaque groupe j par une densité (ou loi discrète) $f_j(\mathbf{x})$.

Observant un point de coordonnées $(x_1, x_2 \ldots, x_p)$ la probabilité qu'il provienne du groupe j est donnée par la formule de Bayes :

$$P(G_j / \mathbf{x}) = \frac{p_j f_j(\mathbf{x})}{\sum_{j=1}^{k} p_j f_j(\mathbf{x})}$$

La règle bayesienne consiste alors à affecter l'observation x au groupe qui a la probabilité *a posteriori* maximale.

Les dénominateurs étant les mêmes pour les k groupes on doit donc chercher le maximum de : $p_j f_j(\mathbf{x})$.

Il est donc nécessaire de connaître ou d'estimer $f_j(\mathbf{x})$. Diverses possibilités existent alors.

2.4.2 Méthodes non paramétriques

On ne fait pas d'hypothèse spécifique sur la famille de loi de probabilité. Des variantes multidimensionnelles de la méthode du noyau permettent d'estimer $f_j(\mathbf{x})$.

$$\hat{f}_j(\mathbf{x}) = \frac{1}{n_j h} \sum_{i=1}^{n_j} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

où K est une densité multidimensionnelle.

La discrimination « par boules » en est un cas particulier : on trace autour de x une boule de rayon ρ donné dans R^p et on compte le nombre d'observations k_j du groupe j dans cette boule.

On estimera alors directement $P(G_j / \mathbf{x})$ par : $k_j / \sum_j k_j$.

Remarque. La boule peut être vide si ρ est trop petit.

Une des méthodes les plus utilisées est cependant la méthode des k plus proches voisins. On cherche les k points les plus proches de x au sens d'une métrique à préciser et on classe x dans le groupe le plus représenté : la probabilité *a posteriori* s'obtient comme pour la discrimination par boules mais n'a pas grand sens si k est faible.

Le choix de k ou de h est crucial. En l'absence de résultats donnant les valeurs optimales de k ou de h, on choisit les paramètres de telle sorte qu'ils donnent de bons résultats sur un échantillon test, mais il faut prendre garde

qu'alors l'échantillon test participe maintenant à l'apprentissage ! La qualité de la méthode doit alors se juger sur un autre échantillon test.

2.4.3 Le modèle normal multidimensionnel

On supposera que x suit une loi $N_p(\mu, \Sigma_j)$ pour chaque groupe :

$$f_j(x) = \frac{1}{(2\pi)^{p/2}\left(\det \Sigma_j\right)^{1/2}} \exp\left[-\frac{1}{2}\left(x-\mu_j\right)' \Sigma_j^{-1}\left(x-\mu_j\right)\right]$$

Le cas général

La règle bayesienne max $p_j f_j(x)$ revient donc en passant en logarithmes à minimiser :

$$\left(x-\mu_j\right)' \Sigma_j^{-1}\left(x-\mu_j\right) - 2\ln p_j + \ln\left(\det\Sigma_j\right)$$

Lorsque les Σ_j sont différents cette règle est donc *quadratique* et il faut comparer k fonctions quadratiques de x.

Σ_j est en général estimé par $\frac{n}{n-1}\mathbf{V}_j$ et μ_j par $\mathbf{g}_j$.

Le cas d'égalité des matrices de variance covariance

Si $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k = \Sigma$, la règle devient linéaire.

En effet $\ln\left(\det\Sigma_j\right)$ est une constante et $\left(x-\mu_j\right)\Sigma^{-1}\left(x-\mu_j\right)$ est alors égale à la distance de Mahalanobis théorique $\Delta^2\left(x, \mu_j\right)$ entre x et μ_j.

En développant et en éliminant $x'\Sigma^{-1}x$ qui ne dépend pas du groupe on a :

$$\max\left\{x'\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j' \Sigma^{-1}\mu_j + \ln p_j\right\}$$

Si Σ est estimé par $\frac{n}{n-k}\mathbf{W}$, la règle bayesienne correspond à la règle géométrique lorsqu'il y a égalité des probabilités *a priori*. Alors la règle géométrique est optimale.

La probabilité *a posteriori* d'appartenance au groupe j est proportionnelle à :

$$p_j \exp\left(-\frac{1}{2}\Delta^2\left(x,\mu_j\right)\right)$$

Deux groupes avec égalité des matrices de variance

On affectera x au groupe 1 si :

$$x'\Sigma^{-1}\left(\mu_1-\mu_2\right) > \frac{1}{2}\left(\mu_1+\mu_2\right)\Sigma^{-1}\left(\mu_1-\mu_2\right) + \ln\frac{p_2}{p_1}$$

Si $p_1 = p_2 = 0.5$ on trouve la règle de Fisher en estimant Σ par $\frac{n}{n-2}\mathbf{W}$.

Soit : $$S(x) = x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + \ln\frac{p_2}{p_1}$$

alors on affectera x au groupe 1 si $S(x) > 0$ et au groupe 2 si $S(x) > 0$.

La fonction $S(x)$ appelée *score* ou statistique d'Anderson est liée simplement à la probabilité *a posteriori* d'appartenance au groupe 1.

On a en effet :

$$P = P(G_1 / x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}$$

d'où : $$\frac{1}{P} = 1 + \frac{p_2 f_2(x)}{p_1 f_1(x)} = 1 + \frac{p_2}{p_1}\exp\left[-\frac{1}{2}(x-\mu_2)\Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)\right]$$

$$\frac{1}{P} - 1 = \frac{p_2}{p_1}\exp\left[\frac{1}{2}\Delta^2(x,\mu_1) - \frac{1}{2}\Delta^2(x,\mu_2)\right]$$

d'où : $\ln\left(\frac{1}{P} - 1\right) = -S(x)$.

Soit : $$P = \frac{1}{1+\exp(-S(x))} = \frac{\exp(S(x))}{1+\exp(S(x))}$$

On dit que P est fonction logistique du *score*.

Lorsque $p_1 = p_2 = 1/2$:

$$P = \frac{1}{1+\exp\left(-\frac{1}{2}\left(\Delta^2(x,\mu_1) - \Delta^2(x,\mu_2)\right)\right)}$$

2.5 Mesures d'efficacité des règles de classement

Le critère usuel est la probabilité de bien classer une observation quelconque. On comparera l'efficacité des diverses méthodes de classification en termes de taux d'erreur.

2.5.1 Taux d'erreur théorique pour deux groupes avec $\Sigma_1 = \Sigma_2$ et distribution normale

Quand $p_1 = p_2$, la règle de classement théorique est d'affecter au groupe 1 si :

$$S(x) = x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) > 0$$

La probabilité d'erreur de classement est donc :

$$P\left(S(x) > 0 / x \in N_p(\mu_2; \Sigma)\right)$$

La loi de $S(x)$ est une loi de Gauss à 1 dimension comme combinaison linéaire des composantes de x.

$$E(S(\mathbf{x})) = \mu_2' \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)$$
$$= -\frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) = -\frac{1}{2}\Delta^2(\mu_1, \mu_2) = -\frac{1}{2}\Delta_p^2$$
$$V(S(\mathbf{x})) = (\mu_1 - \mu_2)' \Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1 - \mu_2) = \Delta_p^2$$

d'où : $S(\mathbf{x})$ suit une loi normale $N\left(-\frac{1}{2}\Delta_p^2; \Delta_p\right)$ si $\mathbf{x} \in G_2$

La probabilité de classer dans le groupe 1 une observation du groupe 2 est :

$$P(1/2) = P\left(U > \frac{\Delta_p}{2}\right)$$

Elle est égale à $P(2/1)$. Cette relation donne une interprétation concrète à la distance de Mahalanobis.

Si $p_1 \neq p_2$ on trouve :

$$P(1/2) = P\left(U > \frac{\Delta_p}{2} + \frac{1}{\Delta_p}\ln\frac{p_2}{p_1}\right)$$
$$P(2/1) = P\left(U > \frac{\Delta_p}{2} - \frac{1}{\Delta_p}\ln\frac{p_2}{p_1}\right)$$

Lorsque μ_1, μ_2, Σ sont estimés, $S(\mathbf{x})$ ne suit plus une loi normale et utiliser D_p comme estimation de Δ_p conduit à une estimation biaisée des probabilités d'erreur de classement : il y a en moyenne, sous-estimation de la probabilité globale d'erreur $p_1 P(2/1) + p_2 P(1/2)$, due entre autres raisons au fait que D_p^2 surestime Δ_p^2.

2.5.2 La méthode de resubstitution

Elle consiste à réaffecter les n observations selon les fonctions discriminantes trouvées.

Cette méthode possède un grave défaut : elle sous-estime systématiquement le taux d'erreur puisqu'on utilise les mêmes observations que celles qui ont servi à trouver les fonctions discriminantes. La règle étant optimale pour l'échantillon ne peut donner que de bons résultats si on la réapplique sur lui.

2.5.3 Les méthodes de validation croisée

Pour échapper au défaut de la méthode de resubstitution on conseille de partager l'échantillon en deux sous-échantillons : l'un servira à l'élaboration des règles de classement (échantillon de base ou d'apprentissage) l'autre pour l'application des règles de classement (échantillon-test).

Le taux d'erreur mesuré sur l'échantillon-test sera alors une estimation sans biais du taux vrai. Ceci suppose cependant de disposer de données nombreuses pour pouvoir soustraire, sans risque, une partie notable des données (25 % est conseillé).

Dans le cas d'échantillons de taille faible la technique suivante due à Lachenbruch et Mickey (1968), comparable au *press* en régression, permet d'obtenir une estimation réaliste du taux d'erreur : on effectue n analyses discriminantes sur chacun des n échantillons de $n - 1$ observations obtenus en mettant de côté tour à tour chacune des observations. On classe alors l'observation mise à part et on compte le pourcentage d'erreur de classement.

2.5.4 Conclusion

Les procédures usuelles de sélection de variables optimisent des critères probabilistes : (lois du Λ de Wilks ou distance de Mahalanobis), elles n'optimisent pas nécessairement le pourcentage de bien classés.

La règle bayesienne donne la solution optimale ; cependant, même dans le cas de lois normales avec matrices de variance connues et différentes, il peut être préférable pour des raisons de robustesse d'utiliser une règle linéaire non optimale, si l'échantillon d'apprentissage est trop petit ou si les variables explicatives sont trop corrélées.

La recherche de règles interprétables favorise les classifieurs linéaires ou par arbres, par opposition aux autres méthodes qui sont plutôt des boîtes noires.

Les hypothèses de distribution jouent un rôle déterminant dans certains cas. En l'absence de telles informations, il faut s'adapter aux données sans chercher à les modéliser. C'est ce que réalise la discrimination non paramétrique par estimation de densité.

Si des comparaisons sont à faire entre les algorithmes neuronaux et les méthodes statistiques classiques, il faut en fait comparer méthodes neuronales et méthodes non paramétriques, car l'intérêt des algorithmes neuronaux est de pouvoir utiliser des classifieurs non-linéaires.