Une introduction à l'analyse discriminante avec *SPSS pour Windows*

Dominique DESBOIS

INRA-ESR Nancy et SCEES
251 rue de Vaugirard, 75732 Paris Cedex 15.
Fax : +33 1 49 55 85 00 Mél :desbois@jouy.inra.fr

RÉSUMÉ : *cette note initie l'utilisateur débutant à la mise en œuvre de* ***l'analyse discriminante*** *par l'intermédiaire de la procédure* ***DISCRIM*** *du logiciel* ***SPSS pour Windows****. Cette mise en œuvre concerne le classement d'individus caractérisés par des variables quantitatives, les affectant à des groupes a priori au moyen de* ***fonctions discriminantes.*** *Les sorties du logiciel sont commentées par la présentation du formulaire de l'analyse discriminante associé à chacun des résultats obtenus.*

MOTS-CLÉS : analyse discriminante linéaire, analyse factorielle discriminante, fonctions linéaires discriminantes, logiciel *SPSS 11.0.1*, mise en œuvre.

I) Problématiques de discrimination

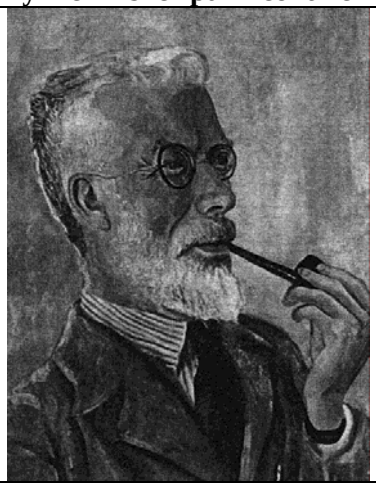
Dans beaucoup de domaines, les professionnels sont amenés à prévoir les comportements sur la base de certains critères : c'est le cas par exemple d'un médecin établissant un diagnostic pour prescrire un traitement, ou d'un banquier accordant un crédit à un particulier ou une entreprise.

Une stratégie très empirique mais suffisamment éprouvée consiste à comparer ces caractéristiques aux relevés effectués sur un ensemble d'observations où le comportement des individus qu'ils soient malades ou emprunteurs, entreprises ou particuliers, est connu. Diagnostic ou prévision s'effectuent alors sur la base des corrélations observées entre les critères retenus et la situation des individus. Cette stratégie est d'ailleurs appliquée de façon plus ou moins implicite par la plupart des décideurs sur la base de leurs expériences antérieures dans des circonstances similaires.

Dans certains cas, les enjeux financiers ou humains sont suffisamment importants pour qu'on envisage de formaliser ce processus décisionnel en proposant une aide au diagnostic ou au jugement de l'expert. Par exemple dans le secteur bancaire, les techniques de *credit-scoring* cherchent à distinguer les bons des mauvais payeurs sur la base d'un ensemble de variables appelés **descripteurs** comme l'âge, le revenu annuel et le profil des transactions bancaires. Disposant de comptabilités d'entreprises, on peut également chercher à minimiser le risque de défaillance ou à prévenir les défaillances des emprunteurs, en classant selon une échelle de risque fixée a priori les agents économiques sur la base des descripteurs que sont les ratios financiers (cf. [Bardos, 2001]). **L'affectation à une classe** de risque décide alors de leur accès aux ressources financières ou de leur éligibilité à des mesures d'aide publique.

Consistant à prévoir l'appartenance du sujet aux groupes d'une partition a priori, la plupart de ces applications fondent leurs prédictions sur une technique statistique multidimensionnelle, **l'analyse discriminante linéaire**, introduite par **Ronald Fisher**.

Figure 1 : Portrait de Sir Ronald Aylmer Fisher par Leontine Tinter (Biometrika 50, 1963).°



Sir Ronald Aylmer FISHER
1890-1962

Droits réservés

La solution proposée dès 1936 par Fisher consiste à chercher des **combinaisons linéaires de descripteurs quantitatifs,** indicateurs synthétiques qui permettent de classer les individus correctement dans chacun des groupes. Cette méthode, essentiellement analytique, est basée sur des concepts géométriques, cependant pour que ces combinaisons linéaires puissent être optimales (au sens où le risque d'erreur de classement serait alors minimal), nous verrons que les données doivent vérifier certaines hypothèses de nature probabiliste.

II) Données : les Iris de Fisher

Les données utilisées dans cet exemple furent publiées initialement par Fisher dans son article originel présentant le concept de fonction linéaire discriminante [Fisher, 1936]. L'échantillon étudié par Fisher comporte cent cinquante iris provenant de trois espèces distinctes (*Iris Setosa*, *Iris Versicolor* et *Iris Virginica*) à raison de cinquante iris par espèce qui constituent ainsi notre **échantillon d'apprentissage** :

Figure 2 : Fleurs des Iris Setosa, Versicolor et Virginica

Iris Setosa	***Iris Versicolor***	***Iris Virginica***

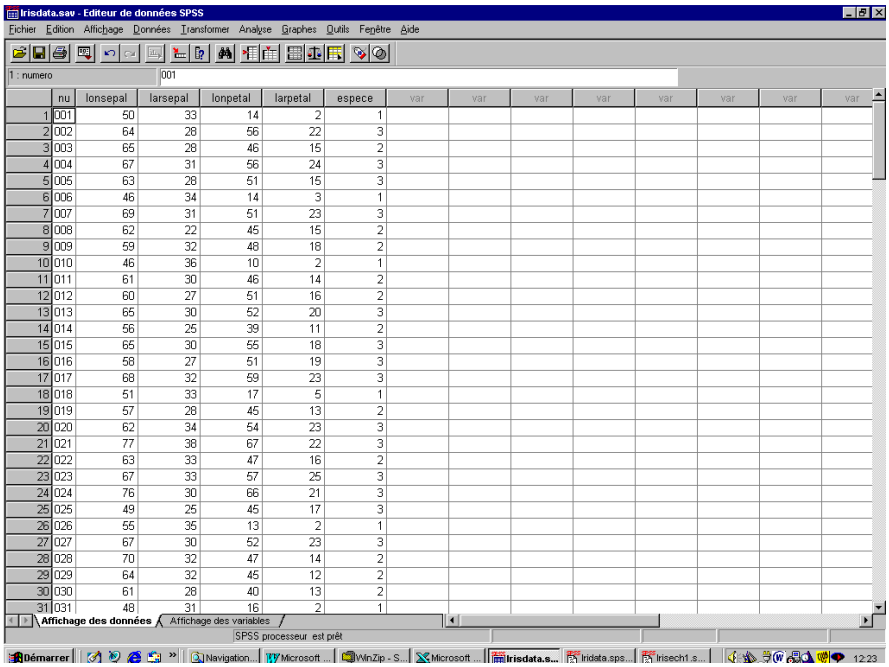
Droits réservés, avec l'aimable autorisation de Marc-Michel Corsini (http://www.sm.u-bordeaux2.fr/~corsini/Pedagogie/)

Chaque individu est identifié par un numéro de séquence (**numero**) au sein de l'échantillon d'apprentissage et son appartenance à l'une des trois populations est renseignée par un code d'espèce (**1** pour *Setosa*, **2** pour *Versicolor* et **3** pour *Virginica*) . Ces données constituent l'un des échantillons les plus utilisés pour les méthodes de discrimination : on les retrouve dans de nombreux ouvrages et répertoires, notamment *http://www.ics.uci.edu/~mlearn/MLRepository.html* [Blake & Mertz, 1998].

Parmi les mesures effectuées, quatre d'entre elles caractérisent la fleur : longueur du sépale (**lonsepal**), largeur du sépale (**larsepal**), longueur du pétale (**lonpetal**), largeur du pétale (**larpetal**) exprimées en millimètres. Le problème de discrimination se pose ainsi : à partir de ces quatre mesures quantitatives donnant une indication sur la morphologie globale de la fleur, peut-on décider de l'espèce à laquelle appartient l'individu ?

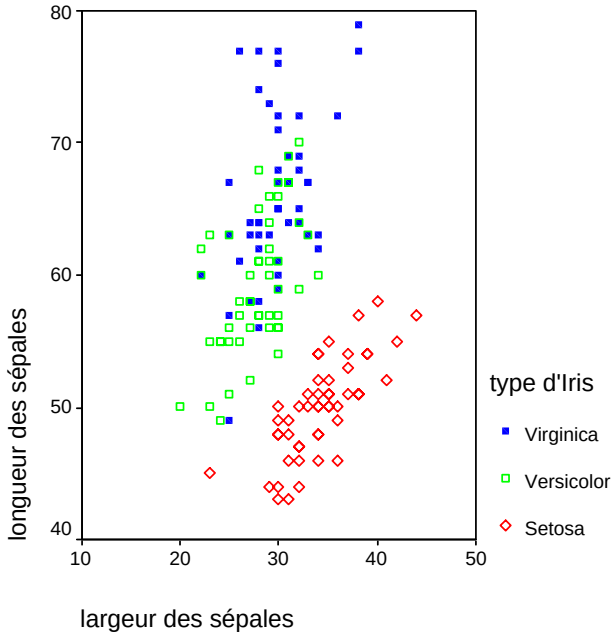
La variable *y* à prédire est donc une variable qualitative (**espece**) à $k = 3$ modalités. Cette prédiction s'effectue à partir d'un tableau X de $p = 4$ variables quantitatives observées sur un échantillon d'apprentissage de $n = 150$ individus.

Figure 3 : extrait du fichier des données.



Les options de la procédure DISCRIM de *SPSS*, version *11.0.1*, permettant d'effectuer ce type d'analyse sont décrites de façon exhaustive dans la quatrième partie de cette note. On y retrouvera les spécifications permettant d'obtenir les résultats interprétés dans la troisième partie. Par exemple, dans le logiciel *SPSS*, les observations de l'échantillon d'apprentissage retenues pour l'analyse discriminante ne doivent pas comporter de valeurs manquantes (option retenue par défaut), à moins de spécifier leur remplacement par la moyenne (cf. infra *IV.v*, dernier §).

Figure 4 : les observations dans l'espace des descripteurs des sépales.



III) Interprétation des résultats de l'analyse discriminante

L'analyse des liaisons entre le tableau quantitatif *X* des descripteurs et la variable qualitative ***y*** à *k* modalités, codant la partition a priori en *k* groupes de l'ensemble des individus, peut être menée selon deux points de vue : le premier à orientation descriptive est centré sur la décomposition de la variance en s'appuyant sur des notions géométriques ; le second à orientation décisionnelle se focalise sur le risque d'erreur en faisant intervenir une modélisation probabiliste.

III.1) Statistiques descriptives

i) Résumé statistique univarié

Bien que des mesures multiples impliquent une certaine redondance dans l'information apportée par l'échantillon et se traduisent par des corrélations entre variables quantitatives, il est toujours intéressant dans un but descriptif de disposer de statistiques univariées pour chacun des groupes étudiés sur les variables de l'analyse, ne serait ce que pour en contrôler les paramètres.

Tableau 1 : statistiques univariées pour chacun des groupes.

Statistiques de groupe

type d'Iris		Moyenne	Ecart-type	N valide (liste) Non pondérées	N valide (liste) Pondérées
Setosa	longueur des sépales	49,94	3,66	50	50,000
	largeur des sépales	34,28	3,79	50	50,000
	longueur des pétales	14,62	1,74	50	50,000
	largeur des pétales	2,46	1,05	50	50,000
Versicolor	longueur des sépales	59,36	5,16	50	50,000
	largeur des sépales	27,70	3,14	50	50,000
	longueur des pétales	42,60	4,70	50	50,000
	largeur des pétales	13,20	1,97	50	50,000
Virginica	longueur des sépales	65,88	6,36	50	50,000
	largeur des sépales	29,74	3,22	50	50,000
	longueur des pétales	55,52	5,52	50	50,000
	largeur des pétales	20,26	2,75	50	50,000
Total	longueur des sépales	58,39	8,34	150	150,000
	largeur des sépales	30,57	4,36	150	150,000
	longueur des pétales	37,58	17,65	150	150,000
	largeur des pétales	11,97	7,62	150	150,000

Afin de suivre les calculs effectués par la procédure d'analyse discriminante, fixons les notations adoptées : soit x_{ijl}, **l'observation** selon le descripteur *j* pour l'individu *i* du groupe G_l. À chaque individu *i* du groupe G_l, est associée une **pondération** m_{il}. La somme des masses m_{il} pour l'ensemble des n_l individus du groupe G_l donne la masse du groupe $m_l = \sum_{i=1}^{n_l} m_{il}$. La **masse totale** de l'échantillon des *n* individus s'obtient par sommation des masses des différents groupes : $m = \sum_{l=1}^{k} m_l$.

Dans notre échantillon, la pondération étant uniforme et égale à 1 pour chaque individu ($m_{il} = 1 \ \forall i,l$), la **masse locale** pour chaque groupe est égale au nombre d'individus ($m_l = n_l = 50 \quad \forall l$).

La **moyenne locale** $\bar{x}_{.jl}$ de la variable *j* pour le groupe G_l est définie par :

$$\bar{x}_{.jl} = \frac{1}{m_l} \sum_{i=1}^{n_l} m_{il} x_{ijl}$$

Cette valeur définit la j^e coordonnée du barycentre g_l du groupe *l*. Soit $\bar{x}_{.12} = 59{,}36$ moyenne locale de la variable longueur des sépales pour le groupe *Versicolor*.

L'**écart-type local** $s_{.jl}$ de la variable *j* pour le groupe G_l est défini par la racine carrée de la variance locale : $s^2_{.jl} = \frac{1}{(m_l - 1)}\left(\sum_{i=1}^{n_l} m_{il} x^2_{ijl} - m_l \bar{x}^2_{.jl}\right)$

Soit $s_{.12} = 5{,}16$ écart-type local de la variable longueur des sépales.

La **moyenne globale** $\bar{x}_{.j.}$ de la variable *j* est définie par : $\bar{x}_{.j.} = \frac{1}{m}\sum_{l=1}^{k} m_l \bar{x}_{.jl}$ masse totale de l'échantillon . Soit $\bar{x}_{.1.} = 58{,}39$ moyenne globale de la variable longueur des sépales.

L'**écart-type global** $s_{.j.}$ de la variable *j* est défini par la racine carrée de la variance globale :

$$s^2_{.j.} = \frac{1}{(m-1)}\left(\sum_{l=1}^{k}\sum_{i=1}^{n_l} m_{il} x^2_{ijl} - m\bar{x}^2_{.j.}\right)$$

Soit $s_{.1.} = 8{,}34$ écart-type local de la variable longueur des sépales.

On observe pour chacune des mesures effectuées des différences notables entre les moyennes des trois groupes mais également des valeurs sensiblement distinctes des écarts-types. Ces remarques peuvent être confirmées ou invalidées par des tests statistiques.

L'usage des statistiques univariées est fortement recommandé dans l'étape de constitution de l'échantillon et de la sélection des descripteurs pour étudier la nature des distributions des valeurs observées au sein des différents groupes.

ii) Composantes de la variabilité

L'analyse discriminante généralise à plusieurs descripteurs quantitatifs (soit ***p*** variables explicatives) la question à laquelle l'analyse de la variance permet de répondre : comment à partir des valeurs d'une variable quantitative prédire le classement des observations parmi ***k*** groupes distincts. Dans le contexte d'un seul descripteur quantitatif *x*, d'un facteur à *k* modalités et d'une pondération uniforme, l'analyse de la variance conduit à décomposer la somme des carrés des écarts à la moyenne globale de l'échantillon entre la somme des carrés des écarts à la moyenne locale pour les observations de chacun des groupes (intraclasses, ou ***W*** comme *Within*) et la somme des carrés des écarts des moyennes locales à la moyenne globale pour chacun des groupes (interclasses, ou ***B*** comme *Between*).

$$[1] \qquad SCE_T = \sum_{i=1}^{n}\left(x_i - \bar{x}_{.j.}\right)^2 = \sum_{l=1}^{k}\sum_{i=1}^{n_l}\left(x_{il} - \bar{x}_l\right)^2 + \sum_{l=1}^{k} n_l\left(\bar{x}_l - \bar{x}\right)^2 = SCE_W + SCE_B$$

En divisant chacun des termes de cette équation par les degrés de libertés correspondant au nombre de valeurs indépendantes dans les sommations effectuées, on aboutit à la notion de carré moyen intraclasses CM_W et carré moyen interclasses CM_B permettant une comparaison de la variance intraclasses et de la variance interclasses :

$$CM_W = SCE_W/(n-k) \qquad CM_B = SCE_B/(k-1)$$

En situation d'inférence par rapport à une population dont *n* observations constitueraient un échantillon aléatoire, pour une variable normalement distribuée, la statistique ***F*** du rapport de variance définie par :

$$F = \frac{CM_B}{CM_W}$$

suit une distribution théorique $F[(k-1);(n-k)]$ de Fisher-Snedecor à $(k-1)$ et $(n-k)$ degrés de liberté sous l'hypothèse nulle H_0 d'égalité des moyennes entre les *k* groupes. Au seuil de risque choisi α, l'hypothèse nulle H_0 sera rejetée si *F* est supérieure au $(1-\alpha)^e$ quantile $F_{(1-\alpha)}[(k-1);(n-k)]$ d'une distribution de Fisher-Snedecor.

Dans le cas de plusieurs descripteurs $\{X_1, X_2, ..., X_p\}$, la variabilité totale du tableau ***X*** des variables explicatives du classement est exprimée par la **matrice des sommes de carrés et de coproduits totaux *T,*** de terme général ***:***

$$t_{jj'} = \sum_{l=1}^{k} \sum_{i=1}^{n_k} m_{il} x_{ijl} x_{ij'l} - m\bar{x}_{.j.} \bar{x}_{.j'.}$$

Soit, par exemple pour le premier élément diagonal t_{11}, somme totale des carrés des écarts pour la longueur des sépales : $t_{11} = 10365{,}793 = SCE_T$

À l'instar de l'analyse de la variance univariée, la matrice des sommes de carrés et de coproduits totaux ***T*** se décompose en deux matrices : ***W***, matrice des sommes de carrés et de coproduits intragroupes, et ***B***, matrice des sommes de carrés et de coproduits intergroupes. L'équation matricielle d'analyse de la variance s'écrit alors :

[2] $T = W + B$.

La variabilité intraclasses du tableau *X* est exprimée par la **matrice des sommes de carrés et de coproduits intraclasses W**, de terme général :

$$w_{jj'} = \sum_{l=1}^{k} \sum_{i=1}^{n_k} m_{il} x_{ijl} x_{ij'l} - \sum_{l=1}^{k} m_l \bar{x}_{.jl} \bar{x}_{.j'l}$$

Soit, par exemple pour le premier élément diagonal w_{11}, somme intraclasses des carrés des écarts pour la longueur des sépales : $w_{11} = 3943{,}620 = SCE_W$

La matrice ***B*** des **sommes de carrés et de produits interclasses**, définie par $B = T - W$, s'obtient par différence terme à terme :

[3] $b_{jj'} = t_{jj'} - w_{jj'}$

Soit, par exemple pour le premier élément diagonal b_{11}, somme des carrés des écarts interclasses pour la longueur des sépales : $b_{11} = 6422{,}173 = SCE_B$

Le calcul du *F* univarié pour chaque descripteur *j* s'effectue alors directement à partir des termes diagonaux des matrices *T* et *W* :

$$F_j = \frac{(t_{jj} - w_{jj})(m-k)}{w_{jj}(k-1)} = \frac{b_{jj}(m-k)}{w_{jj}(k-1)}$$

<u>Tableau 2 :</u> tests statistiques d'égalité des moyennes

Tests d'égalité des moyennes des groupes

	Lambda de Wilks	F	ddl1	ddl2	Signification
longueur des sépales	,380	119,695	2	147	,000
largeur des sépales	,599	49,160	2	147	,000
longueur des pétales	,059	1180,161	2	147	,000
largeur des pétales	,071	961,645	2	147	,000

Ainsi pour une valeur $F \approx 119{,}7$ et un risque $\alpha < 0{,}0005$, nous sommes conduits à rejeter l'hypothèse nulle d'égalité des moyennes des espèces d'iris pour la variable longueur des sépales.

Dans le tableau 2, figure une autre statistique, le ***lambda de Wilks univarié.*** Pour chacune des variables, le lambda de Wilks univarié est constitué par le rapport de la somme des carrés des écarts intraclasses à la somme totale des carrés des écarts : $\Lambda_j = \frac{w_{jj}}{t_{jj}}$

Tableau 3 : analyse de la variance à un facteur pour la longueur des sépales.

ANOVA

longueur des sépales

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	6422,173	2	3211,087	119,695	,000
Intra-groupes	3943,620	147	26,827		
Total	10365,793	149			

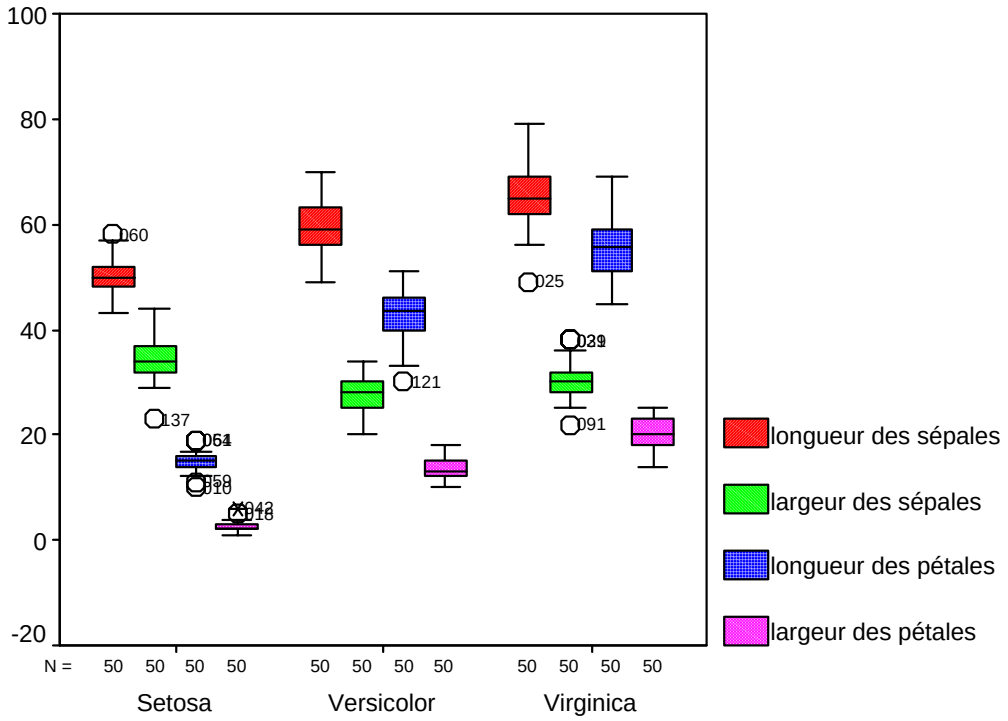
On peut vérifier cette égalité d'après les valeurs inscrites dans le tableau 3 pour la longueur des sépales :

$$\Lambda_1 = \frac{w_{11}}{t_{11}} = \frac{SCE_W}{SCE_T} \approx \frac{3943{,}6}{10365{,}8} \approx 0{,}380$$

La valeur du lambda de Wilks univarié varie entre 0 et 1. La valeur 1 signifie l'égalité des moyennes pour l'ensemble des groupes. Une valeur quasi-nulle est associée à de très faibles variabilités intraclasses donc à de très fortes variabilités interclasses et des moyennes de groupes manifestement différentes.

Cependant, l'étude graphique de la distribution des variables au sein de chacun des groupes et leur comparaison montrent qu'un certain nombre de recouvrements peuvent s'opérer entre les trois populations pour les mesures effectuées, ainsi qu'en témoignent les quatre séries de boîtes à moustaches juxtaposées en figure 5.

Figure 5 : boîte à moustaches multiple pour chacun des groupes.



Il convient donc d'étudier globalement pour l'ensemble des descripteurs la décomposition de cette variabilité afin d'obtenir une vision synthétique des différences intergroupes et des différences interindividuelles (ou intraclasses), ce qui nécessite de passer de l'étude partielle de chacun des descripteurs à celle globale des matrices.

iii) Analyse bivariée de la variabilité

Dans la plupart des analyses multivariées, on observe des liaisons parmi les descripteurs retenus. Cependant, de fortes corrélations entre les variables explicatives peuvent conduire à une forte variabilité dans les estimations des coefficients de la fonction discriminante. Il est donc indispensable d'examiner soit les matrices de variance-covariance si les variables ont des dispersions équivalentes, sinon les matrices de corrélation entre variables afin de détecter de semblables situations.

En outre, que le point de vue théorique adopté soit géométrique ou probabiliste, le fait que les groupes présentent ou non une dispersion des mesures individuelles à peu près similaire peut avoir, selon la règle de classement alors adoptée, une certaine influence sur les résultats.

Tableau 4 : estimations des matrices communes de variance-covariance et corrélation intraclasses

Matrices intra-groupes combinées[a]

		longueur des sépales	largeur des sépales	longueur des pétales	largeur des pétales
Covariance	longueur des sépales	26,827	9,406	16,736	3,805
	largeur des sépales	9,406	11,539	5,524	3,183
	longueur des pétales	16,736	5,524	18,519	4,218
	largeur des pétales	3,805	3,183	4,218	4,177
Corrélation	longueur des sépales	1,000	,535	,751	,359
	largeur des sépales	,535	1,000	,378	,459
	longueur des pétales	,751	,378	1,000	,480
	largeur des pétales	,359	,459	,480	1,000

a. La matrice de covariance a 147 degré(s) de liberté

La **matrice *C* de variance-covariance intraclasses combinée** (cf. tableau 4) est la moyenne pondérée des **matrices C_l de variance-covariance locale** à chacun des *k* groupes (cf. tableau 5) :

$$C = \frac{W}{(m-k)} = \frac{\sum_{l=1}^{k} W_l}{(m-k)} = \frac{1}{(m-k)} \sum_{l=1}^{k} (m_l - 1) C_l$$

soit, pour la variance intraclasses de la longueur des sépales :

$$c_{11} = \frac{3943,62}{147} = \frac{1}{147}(49 \times 13,404 + 49 \times 26,643 + 49 \times 40,434) = 26,827 .$$

La **matrice de corrélation commune *R*** est calculée à partir de la matrice *W* des sommes de carrés et de produits intraclasses combinée : $R_{jj'} = \frac{w_{jj'}}{\sqrt{w_{jj} w_{j'j'}}} = \frac{c_{jj'}}{\sqrt{c_{jj} c_{j'j'}}}$

Soit, pour la corrélation entre la longueur des sépales et la largeur des sépales :

$$r_{12} = \frac{1382,682}{\sqrt{3943,620 \times 1696,200}} = \frac{9,406}{\sqrt{26,827 \times 11,539}} \approx 0,535$$

La **matrice de variance-covariance totale *T'*** (cf. tableau 5) est calculée à partir de la matrice *T* des sommes de carrés et de produits globale à l'échantillon : $T' = \frac{T}{(m-1)}$

Ainsi, la variance globale de la longueur des sépales est égale à :

$$t'_{11} = \frac{t_{11}}{(m-1)} = \frac{10365,793}{149} \approx 69,569$$

Tableau 5 : matrices locales et totale de variance-covariance

Matrices de covariances[a]

type d'Iris		longueur des sépales	largeur des sépales	longueur des pétales	largeur des pétales
Setosa	longueur des sépales	13,404	10,323	1,589	1,212
	largeur des sépales	10,323	14,369	1,170	,930
	longueur des pétales	1,589	1,170	3,016	,607
	largeur des pétales	1,212	,930	,607	1,111
Versicolor	longueur des sépales	26,643	8,518	18,290	5,294
	largeur des sépales	8,518	9,847	8,265	3,857
	longueur des pétales	18,290	8,265	22,082	7,163
	largeur des pétales	5,294	3,857	7,163	3,878
Virginica	longueur des sépales	40,434	9,376	30,329	4,909
	largeur des sépales	9,376	10,400	7,138	4,763
	longueur des pétales	30,329	7,138	30,459	4,882
	largeur des pétales	4,909	4,763	4,882	7,543
Total	longueur des sépales	69,569	-4,261	128,341	51,957
	largeur des sépales	-4,261	18,998	-32,966	-12,193
	longueur des pétales	128,341	-32,966	311,628	129,412
	largeur des pétales	51,957	-12,193	129,412	58,040

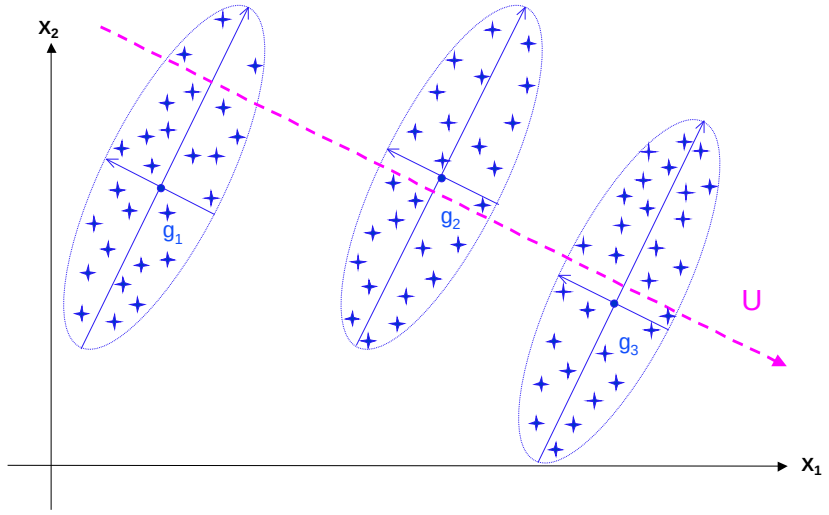
a. La matrice de covariance totale a 149 degré(s) de liberté.

iv) Homogénéité des variances

Les variances locales sont dites homogènes si on observe une dispersion équivalente des valeurs observées au sein de chaque groupe autour de la moyenne locale. L'hypothèse d'homogénéité des variances joue un rôle important dans le modèle probabiliste de l'analyse discriminante car elle correspond à une situation théorique intéressante où le nombre de paramètres à estimer est plus faible : *k* moyennes et une seule matrice de variance-covariance commune à l'ensemble des classes contre *k* moyennes et *k* matrices de variance-covariance locales distinctes si on a hétérogénéité des variances.

En analyse multivariée, l'homogénéité des variances se traduit géométriquement par des nuages de points aux formes similaires, c'est à dire avec une dispersion et des orientations semblables :

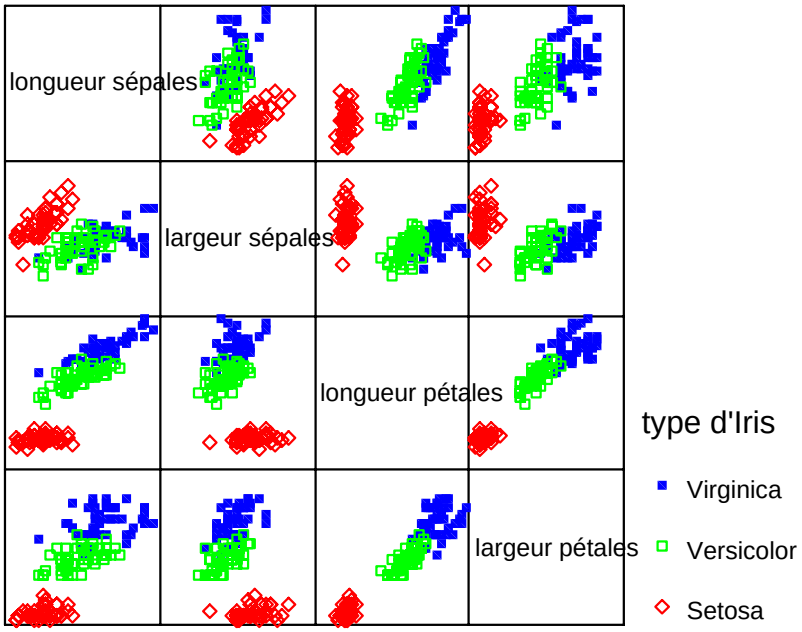
Figure 6 : dispersion et orientation des classes dans une situation idéale pour la discrimination.



En langage algébrique, l'homogénéité des variances se traduit dans l'espace multidimensionnel des descripteurs par l'égalité des matrices de variance-covariance de chacun des groupes.

Si le nombre des descripteurs n'est pas trop élevé, une matrice de nuages de points peut aider à visualiser la forme des distributions de valeur au sein de chaque groupe.

Figure 7 : matrice de graphiques de dispersion.



On constate que les différences de dispersion et d'orientation les plus notables sont celles existant entre le nuage des Setosa et ceux des Virginica et des Versicolor. Chacune des mesures permet de sélectionner facilement les Setosa parmi l es autres Iris, par contre aucune des mesures ne permet de distinguer les Virginica des Versicolor sans risque de confusion.

v) Définition des fonctions linéaires discriminantes

Le principe de l'analyse discriminante linéaire est de former des combinaisons linéaires des variables explicatives permettant d'affecter les individus à leur groupe d'origine avec un minimum d'erreur de classement. La ***fonction linéaire discriminante*** u qui minimise le taux de mal-classés s'écrit comme une moyenne pondérée des variables explicatives, résumant ainsi en un seul indicateur l'information apportée par les ***q* variables explicatives** sélectionnées parmi les descripteurs :

$$u = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q$$

Dans cette équation similaire à celle de la régression multiple, u est la variable à prédire dont les valeurs doivent permettent de distinguer les trois espèces d'Iris, les X_j sont les variables explicatives et les β_j figurent les coefficients à estimer d'après les valeurs observées sur l'échantillon d'apprentissage. Les **estimations des coefficients** $b_j = \hat{\beta}_j$ de cette équation sont choisies de manière à minimiser les différences interindividuelles au sein de chaque groupe (groupes homogènes) pour les valeurs de u_h tout en maximisant les différences entre les groupes (groupes bien séparés). Ces estimations maximisent donc le rapport des deux sommes de carrés (intergroupes sur intragroupes) :

$$[3] \qquad \left[\hat{\beta}_j\right]_J = \left\{ \left[\beta_j\right]_J \; tel \; que \; \frac{SCE_B}{SCE_W} \; soit \; \max imum \right\} = Arg \max \left\{ \frac{SCE_B}{SCE_W} \right\}$$

III.2) Approche géométrique

i) Analyse factorielle discriminante

Les **méthodes factorielles** d'analyse multivariée reposent sur les concepts de la géométrie euclidienne faisant intervenir la notion d'inertie, produit de la masse (somme des pondérations des observations) par la distance au carré (variance des descripteurs), correspondant à une somme pondérée des écarts au carré.

Cherchant des **combinaisons linéaires** $u = \sum_{j=1}^{q} \beta_j X_j$ des descripteurs susceptibles de séparer du mieux possible les *k* groupes, on sélectionne celles présentant le meilleur compromis entre deux objectifs distincts : représenter les groupes à la fois comme homogènes (minimiser l'inertie intraclasses) et comme bien séparés (maximiser leur inertie interclasses). D'après l'équation matricielle [2], l'inertie du nuage de points se décompose en inertie intraclasses et en inertie interclasses :

[4] $u'Tu = u'Wu + u'Bu$.

La recherche de variables discriminantes revient à trouver une combinaison linéaire *u* qui maximise le rapport de ces deux inerties : $\frac{u'Bu}{u'Wu}$

ou, ce qui revient au même compte tenu de l'équation [4] : $f(u) = \frac{u'Bu}{u'Tu}$.

On montre que ce problème est équivalent à la recherche du maximum de la forme quadratique $u'Bu$ sous la contrainte $u'Tu = 1$ (cf. [Lebart, Morineau & Piron, 1995]) . En utilisant la technique des multiplicateurs de Lagrange pour résoudre ce problème de recherche d'un extremum sous contrainte, on obtient comme solution le vecteur *u* défini par : $Bu = \lambda Tu$.

En supposant que la matrice *T* de variance-covariance totale est inversible, le vecteur *u*, solution de cette équation :

[5] $T^{-1}Bu = \lambda u$

est le vecteur propre associé à la plus grande **valeur propre λ** de l'opérateur $T^{-1}B$.

Cette valeur propre $\lambda = u'Bu$ réalise le maximum recherché de la variance interclasses $u'Bu$ de la variable *u* et constitue le **pouvoir discriminant** de la variable ***u***, appelée alors **fonction linéaire discriminante** ou **facteur discriminant**.

L'axe factoriel discriminant *a*, associé au facteur discriminant *u* tel que $u = T^{-1}a$, est le vecteur propre associé à λ pour l'opérateur BT^{-1} :

$BT^{-1}a = \lambda a$.

Le nombre d'axes factoriels distincts que l'on peut extraire est au plus égal au minimum du nombre ***q*** de **variables explicatives** et du nombre de groupes minoré de 1 : $\min\{q, k-1\}$.

Le pouvoir discriminant d'un axe factoriel varie entre 0 et 1 :

- le cas $\lambda = 1$ correspond à une dispersion intraclasses nulle (les *k* sous-nuages de points correspondant aux groupes se situent dans un hyperplan orthogonal à l'axe factoriel discriminant *a*) et à une discrimination parfaite si les *k* centres de gravités g_l se projettent sur *a* en des points distincts ;
- le cas $\lambda = 0$, les projections des centres de gravités g_l sur l'axe factoriel discriminant *a* sont confondues.

Ainsi définie par la recherche d'axes factoriels discriminants orthonormés pour la métrique T^{-1}, **l'analyse factorielle discriminante** (**AFD**) n'est rien d'autre qu'une analyse en composantes principales du nuage des *k* centres de gravité g_L , doté de la métrique T^{-1}.

Au lieu de la métrique T^{-1}, l'utilisation par la procédure DISCRIM de la métrique W^{-1}, connue sous le nom de « **métrique de Mahanalobis** »[1], conduit à des résultats équivalents. La

[1] Mahanalobis en proposant la même année que Fisher une notion de distance généralisée entre groupes, peut également être considéré comme un des pères de l'analyse discriminante.

recherche des combinaisons linéaires maximisant $\frac{u'Bu}{u'Wu}$ selon un raisonnement similaire conduit à la solution donnée par l'équation : [6] $W^{-1}Bu = \mu u$

On montre aisément que les vecteurs propres de $W^{-1}B$ solutions de l'équation [6] sont les mêmes que ceux de $T^{-1}B$ solutions de l'équation [5].

ii) Pouvoir discriminant des facteurs

Les **valeurs propres** μ à valeurs sur l'intervalle $[0;+\infty[$ sont définies par les valeurs propres λ à valeurs sur l'intervalle $[0;1]$ et réciproquement :

$$\mu = \frac{\lambda}{1-\lambda} \qquad \text{et} \qquad \lambda = \frac{\mu}{1+\mu}$$

Tableau 6 : valeurs propres associées aux fonctions linéaires discriminantes

Valeurs propres

Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	31,857[a]	99,1	99,1	,985
2	,297[a]	,9	100,0	,478

a. Les 2 premières fonctions discriminantes canoniques ont été utilisées pour l'analyse.

Les valeurs propres associées aux fonctions linéaires discriminantes permettent de juger du pouvoir discriminant respectif de ces fonctions, en effet chaque **valeur propre** μ_h de **rang *h*** est égale à la variance interclasses de la fonction linéaire discriminante de même rang. Ainsi, la première valeur propre est égale à : $\mu_1 = 31{,}857$ et la seconde valeur propre $\mu_2 = 0{,}297$.

Le pourcentage de variance expliquée rapporte la valeur propre à la somme totale des valeurs propres : le pourcentage de la variance intergroupe expliquée par la première fonction discriminante est de $\tau_1 = \frac{\mu_1}{\sum_{h=1}^{2}\mu_h} * 100 = \frac{31{,}857}{32{,}154} * 100 = 99{,}076 \approx 99{,}1$ tandis que la seconde fonction discriminante, orthogonale à la première, n'explique que $\tau_2 = 0{,}9\ \%$ de la variabilité interclasse.

Tableau 7 : analyse de variance des coordonnées factorielles sur le premier axe discriminant.

ANOVA

Scores discriminants de la fonction 1 pour l'analyse 1

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	4683,032	2	2341,516	2341,516	,000
Intra-groupes	147,000	147	1,000		
Total	4830,032	149			

Pour chaque fonction linéaire discriminante, la valeur propre est égale au rapport de la sommes des carrés des écarts interclasses sur la somme de carrés des écarts intraclasses :

$$\mu_1 = \frac{SCE_B}{SCE_W} = \frac{4683{,}032}{147} \approx 31{,}857$$

Le ***coefficient de corrélation canonique*** est une mesure de la liaison entre les coordonnées factorielles discriminantes et la variable qualitative codant l'appartenance aux groupes. Plus précisément, le coefficient de corrélation canonique entre la fonction linéaire discriminante et le sous-espace engendré par les variables logiques y_l, indicatrices de codage des groupes G_l est donné par : $\rho_h = \sqrt{\mu_h/(1+\mu_h)} = \sqrt{\lambda_h}$.

Soit $\rho_1 = \sqrt{31{,}857/32{,}857} = 0{,}9847 \approx 0{,}985$ et $\rho_2 = \sqrt{0{,}297/1{,}297} = 0{,}4785 \approx 0{,}478$.

Le carré de la corrélation canonique est égal au rapport de corrélation $\eta_{u/y}$ de la variable dépendante u que constitue la fonction linéaire discriminante avec le facteur y codant l'appartenance au groupe, soit pour la première fonction discriminante u_1.

$$\eta_{u_1/y} = \sqrt{SCE_B/SCE_T} = \sqrt{4683{,}032/4830{,}032} \approx 0{,}9847 = \rho_1$$

Rappelons que le carré du coefficient η représente la part de variance expliquée par les différences entre groupes, soit le rapport entre la somme des carrés des écarts intergroupes et la somme totale des carrés des écarts, ce qui le rend complémentaire du lambda de Wilks univarié (cf. tableau 2) : $\Lambda + \eta^2 = 1$

L'analogie pourrait être poursuivie en montrant que l'analyse factorielle discriminante correspond à l'analyse canonique entre le tableau (non centré) Y des variables logiques indicatrices des groupes de la typologie a priori et le tableau X des descripteurs quantitatifs (voir [Saporta, 1990]).

Dans le cas particulier de deux groupes, on peut montrer que l'analyse discriminante est équivalente à la régression multiple, à une transformation linéaire près. La variable qualitative à expliquer y ne possédant que deux modalités, il suffit de prendre des valeurs c_1 pour le groupe G_1 et c_2 pour le groupe G_2 tel que $n_1 c_1 + n_2 c_2 = 0$; par exemple, un vecteur $\boldsymbol{y}$ à n composantes y_i, recodé comme suit: $y_i = \begin{cases} c_1 = \sqrt{n_2/n_1} = 1 & si \quad i \in G_1 \\ c_2 = -\sqrt{n_1/n_2} = -1 & si \quad i \in G_2 \end{cases}$

Le vecteur y, le vecteur b des **coefficients de la fonction linéaire discriminante** u est alors proportionnel au vecteur $\tilde{b}$ des coefficients de la régression multiple**, estimation des paramètres de la régression,** défini par : $\tilde{b} = (X'X)^{-1} X'y$. Cette propriété peut être vérifiée, à l'exception des constantes, sur les coefficients non standardisés par une analyse discriminante entre le groupe des *Setosa* et celui des *Versicolor* :

Tableau 8 : équivalence entre régression linéaire et analyse discriminante pour deux groupes.

Analyse discriminante linéaire (Setosa / Versicolor)		**Régression linéaire**		
Coefficients des fonctions discriminantes canoniques		***Modèle Variable dépendante : type d'Iris***		***Rapport***
Coefficients non standardisés	*F 1*	*Coefficients non standardisés*	*B*	*F1 / B*
longueur des sépales	-0,025756	longueur des sépales	-0,002454	10,495
largeur des sépales	-0,175685	largeur des sépales	-0,016741	10,495
longueur des pétales	0,217398	longueur des pétales	0,020715	10,495
largeur des pétales	0,289617	largeur des pétales	0,027597	10,495
(Constante)	-1,635420	(constante)	1,344165	

Dans le cas de deux groupes (e.g. *Setosa* contre *Versicolor*), le coefficient de corrélation canonique de l'analyse discriminante (tableau 9) :

Tableau 9 : corrélation canonique de l'analyse discriminante pour deux groupes.

Valeurs propres

Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	26,057[a]	100,0	100,0	,981

a. Les 1 premières fonctions discriminantes canoniques ont été utilisées pour l'analyse.

est égal au coefficient de corrélation multiple (tableau 10) de la régression entre *y* et *X*.

Tableau 10 : coefficient de corrélation et R^2 du modèle de régression (deux groupes).

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,981[a]	,963	,961	,099

a. Valeurs prédites : (constantes), largeur des pétales, largeur des sépales, longueur des sépales, longueur des pétales

Les autres indicateurs statistiques basés sur le modèle probabiliste de la régression ne sont guère utilisables puisque dans le contexte de l'analyse discriminante, *y* n'est pas aléatoire tandis que les vecteurs X_j le sont.

iii) Coefficients et interprétation des fonctions discriminantes

Les coefficients non standardisés b_{jh} de chaque fonction discriminante de rang *h* sont les estimations $b_{jh} = \hat{\beta}_{jh}$ des coefficients de l'équation : $u_h = \beta_{0h} + \beta_{1h}X_1 + \beta_{2h}X_2 + \cdots + \beta_{qh}X_q$

La constante b_{0h} associée à la h^e fonction discriminante est égale à :

$$b_{0h} = -\sum_{j=1}^{q} b_{jh}\overline{X}_{.j.}$$

Tableau 11 : estimation des coefficients non standardisés des deux fonctions linéaires discriminantes

Coefficients des fonctions discriminantes canoniques

	Fonction	
	1	2
longueur des sépales	-,079	-,002
largeur des sépales	-,152	,216
longueur des pétales	,218	-,095
largeur des pétales	,279	,292
(Constante)	-2,274	-6,447

Coefficients non standardisés

Les valeurs des coefficients non standardisés de la fonction linéaire discriminante permettent d'utiliser directement les valeurs des variables explicatives pour calculer la **coordonnée factorielle** $u_h(i)$, valeur[2] de la fonction linéaire discriminante u_h d'ordre *h* pour l'individu *i* :

$$u_1(i) \approx -2{,}274 - 0{,}079 \times (lonsepal) - 0{,}152 \times (larsepal) + 0{,}218 \times (lonpetal) + 0{,}279 \times (larpetal)$$

Ainsi la valeur de la première fonction linéaire discriminante u_1 pour le second individu (iris n°2) de notre échantillon d'apprentissage est égale à :

$$u_1(2) \approx -2{,}274 - 0{,}079 \times (64) - 0{,}152 \times (28) + 0{,}218 \times (56) + 0{,}279 \times (22) \approx 6{,}760$$

De même, la seconde fonction linéaire discriminante u_2 s'écrit :

$$u_2(i) \approx -6{,}447 - 0{,}002 \times (lonsepal) + 0{,}216 \times (larsepal) - 0{,}095 \times (lonpetal) + 0{,}292 \times (larpetal)$$

soit pour l'iris n°2, la coordonnée factorielle

$$u_2(2) \approx -6{,}447 - 0{,}002 \times (64) + 0{,}216 \times (28) - 0{,}095 \times (56) + 0{,}292 \times (22) \approx 0{,}577$$

[2] Dans la terminologie adoptée par SPSS, ces coordonnées factorielles sont appelées « **scores discriminants** ».

Les coordonnées factorielles discriminantes sont centrées (de moyenne nulle) relativement à l'ensemble de l'échantillon d'apprentissage.

Comme en régression, les valeurs et les signes des coefficients non standardisés ne sont pas toujours directement interprétables. Pour interpréter une fonction linéaire discriminante, l'analyse des coefficients standardisés est plus pertinente.

Tableau 12 : estimation des coefficients standardisés des deux fonctions linéaires discriminantes.

Coefficients des fonctions discriminantes canoniques standardisées

	Fonction	
	1	2
longueur des sépales	-,409	-,008
largeur des sépales	-,517	,735
longueur des pétales	,939	-,409
largeur des pétales	,571	,597

Pour la première fonction linéaire discriminante, les mesures sur les pétales s'opposent aux mesures effectuées sur les sépales, tandis que longueur et largeur contribuent dans le même sens que ce soit pour les pétales ou les sépales. Pour la seconde fonction linéaire discriminante, largeur des pétales et largeurs des sépales s'opposent à la longueur des pétales, alors que la contribution de la longueur des sépales est quasi-nulle. Notons que les signes sont arbitraires : seules les oppositions de signes ont un sens.

Une autre façon d'interpréter les contributions des variables prédictives aux fonctions linéaires discriminantes est d'étudier la ***matrice de structure*** donnant les corrélations intragroupes combinées entre les variables explicatives et les fonctions discriminantes. Bien que les valeurs des corrélations totales soient plus importantes, les deux types de corrélation sont du même ordre de grandeur.

Tableau 13 : corrélations entre variables prédictives et fonctions linéaires discriminantes.

Matrice de structure

	Fonction	
	1	2
longueur des pétales	,710*	,149
largeur des sépales	-,119	,850*
largeur des pétales	,637	,735*
longueur des sépales	,224	,292*

Les corrélations intra-groupes combinés entre variables discriminantes et les variables des fonctions discriminantes canoniques standardisées sont ordonnées par tailles absolues des corrélations à l'intérieur de la fonction.

*. Plus grande corrélation absolue entre chaque variable et une fonction discriminante quelconque.

On note que la variable la plus corrélée avec la première fonction est la longueur des pétales suivie par la largeur des pétales, tandis que les variables les plus corrélées avec la seconde fonction sont les largeurs, des pétales comme des sépales.

iv) Critères d'affectation géométriques

Les coordonnées factorielles des barycentres de groupe sur les axes discriminants sont évaluées comme valeurs moyennes des groupes : $$g_{kh} = b_{0h} + \sum_{j=1}^{q} b_{jh} \overline{X}_{.j}$$

Tableau 14 : estimation des valeurs moyennes des groupes

Fonctions aux barycentres des groupes

	Fonction	
type d'Iris	1	2
Setosa	-7,563	,221
Versicolor	1,799	-,743
Virginica	5,764	,522

Fonctions discriminantes canoniques non standardisées évaluées aux moyennes des groupes

D'après les valeurs de la première fonction discriminante estimée aux barycentres de chacun des groupes (tableau 14), la projection de l'iris n°2 sur la première fonction linéaire discriminante le classe parmi les *Virginica*, groupe auquel il appartient effectivement. Cependant, le classement n'est pas toujours aussi aisé et l'information complémentaire apportée par la seconde fonction linéaire discriminante peut alors être utile.

Les axes factoriels discriminants fournissent un système de représentation maximisant la variance interclasses et minimisant la variance intraclasses de la partition des *n* individus en *k* groupes. Pour classer une observation x_0 parmi les *k* groupes, l'affectation géométrique consiste à projeter cette observation dans l'espace défini par les axes factoriels discriminants puis à calculer la distance de cette observation à chacun des *k* centres de gravité des groupes. La règle d'affectation est alors définie par la métrique utilisée, soit dans notre cas W^{-1} la métrique de Mahanalobis.

Pour *q* variables explicatives, la ***distance de Mahanalobis*** entre deux groupes G_1 et G_2, de barycentres respectifs μ_{l_1} et μ_{l_2}, de même matrice de variance-covariance Σ est définie par : $\Delta_q(g_1, g_2) = \sqrt{(\mu_{l_1} - \mu_{l_2})' \Sigma^{-1} (\mu_{l_1} - \mu_{l_2})}$.

Le carré de la distance entre les deux groupes est appelé le D^2 de Mahanalobis $D_q^2(g_1, g_2) = (m-k)\sum_{j=1}^{q}\sum_{j'=1}^{q} \widetilde{w}_{jj'} (\overline{X}_{.j1} - \overline{X}_{.j2})(\overline{X}_{.j'1} - \overline{X}_{.j'2})$ où $\widetilde{w}_{jj'}$ est l'élément *(j,j')* de la matrice W^{-1}.

Pour trouver la classe G_l qui minimise la distance de Mahanalobis au carré $d_q^2(x_0; g_l) = (x_0 - g_l)' W^{-1} (x_0 - g_l)$ entre l'observation à classer x_0 et son barycentre g_l, il faut chercher le minimum de la fonction $g_l' W^{-1} g_l - 2 x_0' W^{-1} g_l$

ou encore le maximum de la fonction [7] $x_0' W^{-1} g_l - \left(g_l' W^{-1} g_l \right) \Big/ 2$,

ce qui nous donne une **fonction de classement** linéaire par rapport aux coordonnées de x_0. On calcule donc la valeur en x_0 de ces *k* fonctions de classement et l'observation x_0 est affectée au groupe dont la fonction de classement est maximale en x_0. Cette règle géométrique d'affectation est appelée la **règle de Mahanalobis-Fisher** (cf. [Saporta, 1990]).

L'équation de l'hyperplan médiateur entre les groupes G_l et $G_{l'}$ est le lieu des points qui annulent la **fonction score** $f_{l/l'}(x)$: $f_{l/l'}(x) = (\mu_l - \mu_{l'}) W^{-1} \left(x - \frac{\mu_l + \mu_{l'}}{2} \right) = 0$.

Dans SPSS, les fonctions de classement pour chaque groupe $l = 1,2,\cdots,k$ sont définies par le jeu de coefficients applicables aux valeurs observées, soit :

$b_{lj} = (m-k)\sum_{j'=1}^{q} \widetilde{w}_{jj'} \overline{X}_{.j'l}$ pour $j = 1,\cdots,q$

et avec pour constante $b_{l0} = \ln\left(\frac{m_l}{m}\right) - \frac{1}{2}\sum_{j'=1}^{q} b_{lj'} \overline{X}_{.j'l}$

À une constante près, qui n'a pas d'influence si les groupes ont des masses égales, on retrouve bien la règle de Mahanalobis-Fisher.

Tableau 15 : coefficients des fonctions de classement

Coefficients des fonctions de classement

	type d'Iris		
	Setosa	Versicolor	Virginica
longueur des sépales	2,253	1,515	1,200
largeur des sépales	2,344	,710	,380
longueur des pétales	-1,566	,568	1,312
largeur des pétales	-1,669	,666	2,143
(Constante)	-84,044	-72,388	-104,408

Fonctions discriminantes linéaires de Fisher

Si les dispersions des groupes sont très différentes à la fois en taille et en orientation, la règle géométrique de classement peut conduire à des taux de mal-classés importants. Pour pallier ces insuffisances inhérentes au point de vue géométrique, il est parfois nécessaire d'adopter une démarche probabiliste susceptible de fournir des règles de classement optimales.

III.3) Approche probabiliste

i) Règle bayésienne de classement

L'approche décisionnelle en analyse discriminante est construite sur un raisonnement probabiliste qualifié de bayésien car il s'appuie sur le théorème de Bayes utilisant les probabilités conditionnelles et les probabilités a priori pour calculer les probabilités a posteriori.

La **probabilité a priori**, $p_l = P(G_l)$ est la probabilité qu'un individu appartienne au groupe G_l en l'absence de tout autre information. Les proportions observées dans l'échantillon d'apprentissage peuvent fournir une estimation de ces probabilités a priori. Ces probabilités a priori peuvent également être estimées d'après d'autres sources statistiques comme par exemple le recensement de la population pour des classes d'âge ou bien le recensement de l'agriculture pour les catégories d'exploitation agricole. En l'absence de toute information, on choisira les groupes équiprobables.

Tableau 16 : probabilités a priori des groupes

Probabilités à priori des groupes

		Observations utilisées dans l'analyse	
type d'Iris	A priori	Non pondérées	Pondérées
Setosa	,333	50	50,000
Versicolor	,333	50	50,000
Virginica	,333	50	50,000
Total	1,000	150	150,000

La **probabilité conditionnelle**, $P(x/G_l)$ probabilité d'un vecteur $x = \{x_1, \cdots, x_j, \cdots, x_p\}$, descriptif des observations, connaissant le groupe G_l d'un individu *i,* permet de s'appuyer sur l'information auxiliaire que constitue l'appartenance au groupe pour développer une règle de classement basée sur une estimation de cette probabilité conditionnelle. En effet, si l'on suppose que la distribution des valeurs observées au sein de chacun des groupes est normale et si l'on peut estimer les

paramètres (moyenne et écart-type) de chacune de ces distributions, la connaissance pour chaque individu du groupe d'appartenance permet de calculer la probabilité d'observer le descriptif *x* sachant que l'individu appartient au groupe G_l.

Cependant, dans la démarche de classement ou de discrimination le groupe d'appartenance est inconnu. Connaissant par l'observation les valeurs des variables explicatives, on souhaite avoir une estimation de la probabilité d'appartenance au groupe, soit $P(G_l / x)$ **la probabilité a posteriori**. Cette probabilité a posteriori peut être calculée, en utilisant le théorème de Bayes, à partir de la probabilité conditionnelle $P(x / G_l)$ et de la probabilité a priori $P(G_l)$. Une fois ce calcul effectué pour chacun des groupes, l'individu *i* pourra alors être affecté au groupe G_0 pour lequel la probabilité a posteriori $P(G_0 / x)$ sera maximum.

Soit *k* groupes distincts G_l de probabilité a priori $p_l = P(G_l)$. D'après le théorèmes de Bayes, la probabilité d'appartenance d'un individu *i* au groupe *l* connaissant son vecteur des observations x est donnée par la **probabilité a posteriori** :

$$P(G_l / x) = \frac{P(x / G_l) P(G_l)}{\sum_{l=1}^{k} P(x / G_l) P(G_l)}$$

Si la distribution de probabilité du vecteur des observations est connue par l'intermédiaire d'une densité de probabilité ou d'une distribution discrète $f_l(x)$, on peut construire une **règle bayésienne d'affectation** pour les individus en maximisant la probabilité a posteriori :

$$\underset{l}{Max}\{P(G_l / x)\} = \underset{l}{Max}\left\{\frac{p_l f_l(x)}{\sum_{l=1}^{k} p_l f_l(x)}\right\} = \underset{l}{Max}\{p_l f_l(x)\}.$$

Chaque groupe G_l d'observations est supposé extrait d'une population multinormale $N(\mu_l, \Sigma_l)$ d'espérance $\boldsymbol{\mu_l}$, et de matrices de variance-covariance $\boldsymbol{\Sigma_l}$, de densité :

$$f_l(x) = \frac{1}{(2\pi)^{q/2}\sqrt{\det(\Sigma_l)}} \exp\left\{-\frac{1}{2}(x-\mu_l)' \Sigma_l^{-1}(x-\mu_l)\right\}$$

Dans le cas général où les matrices de variance-covariance locales sont différentes, la règle bayésienne d'affectation revient à minimiser la fonction quadratique en *x* :

[8] $$QS_l(x) = (x-\mu_l)' \Sigma_l^{-1}(x-\mu_l) + \ln\{\det(\Sigma_l)\} - 2\ln\{p_l\}$$

Lorsque les matrices de variance-covariance des groupes sont toutes égales ($\Sigma_L = \Sigma \quad \forall l$), la règle bayésienne d'affectation aboutit à maximiser la fonction linéaire :

[9] $$LS_l(x) = x'\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l'\Sigma^{-1}\mu_l + \ln(p_l)$$

Si la matrice Σ est estimée par $\frac{m}{(m-k)}W$ et si les probabilités a priori p_l sont égales, la règle bayésienne correspond à la règle géométrique de classement de Mahanalobis-Fisher qui est alors optimale dans ce cas particulier. L'équation de l'hyperplan médiateur entre les groupes G_l et $G_{l'}$ est le lieu des points qui annule la **fonction score** $f_{l/l'}(x)$:

$$f_{l/l'}(x) = (\mu_l - \mu_{l'})W^{-1}\left(x - \frac{\mu_l + \mu_{l'}}{2}\right) - \ln\left(\frac{p_{l'}}{p_l}\right) = 0.$$

Dans le cas de deux groupes, la règle d'affectation passe par le **score** ou **statistique d'Anderson** : $$S_{1/2}(x) = x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) + \ln\left(\frac{p_2}{p_1}\right)$$

Le « **point pivot** » $P = \dfrac{(\mu_1 + \mu_2)}{2}$ représente le milieu du segment joignant les barycentres des deux groupes. Si la statistique d'Anderson est positive, l'individu est affecté au groupe 1 ; si le score est négatif, l'individu est affecté au groupe 2, une valeur nulle entraîne l'indétermination.

Ainsi, la règle bayésienne d'affectation conduit à une technique paramétrique basée sur le modèle normal multidimensionnel : distributions normales multivariées pour les *q* variables explicatives sur *k* populations distinctes.

On vérifiera donc l'hypothèse de multinormalité des distributions pour chaque groupe et celle d'homogénéité en comparant les matrices de variance-covariance locales à chaque groupe.

ii) Test d'homogénéité des variances

Dans un contexte de multinormalité, le ***test de Box*** fournit une procédure permettant de valider cette hypothèse d'égalité des matrices de variance-covariance. Il est basé sur la **statistique *M* multivariée de Box** qui constitue une adaptation au cas multivarié de la statistique *M* de Bartlett :

$$M = (m-k)\ln|C| - \sum_{l=1}^{k}(m_l - 1)\ln|C_l|$$

Le déterminant des matrices de variance-covariance représente le volume des ellipsoïdes d'inertie des groupes. La monotonie strictement croissante de la fonction logarithmique permet de comparer la matrice commune de variance-covariance intraclasses, moyenne arithmétique pondérée de ces déterminants, et leur moyenne géométrique. Si les matrices de variance-covariance locales sont égales, alors la moyenne géométrique est égale à la moyenne arithmétique et la valeur de *M* est nulle. Sinon, la valeur de *M* mesure l'écart à l'hypothèse d'indépendance.

Sous l'hypothèse nulle d'égalité des matrices de variance-covariance locales $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$, le rapport :

$$F = \begin{cases} M/b & si \quad e_2 > e_1^2 \\ t_2 M / t_1 (b - M) & si \quad e_2 < e_1^2 \end{cases}$$

avec

$$e_1 = \left(\sum_{l=1}^{k} \frac{1}{m_l - 1} - \frac{1}{m-k} \right) \frac{2p^2 + 3p - 1}{6(k-1)(p+1)}$$

$$e_2 = \left(\sum_{l=1}^{k} \frac{1}{(m_l - 1)^2} - \frac{1}{(m-k)^2} \right) \frac{(p-1)(p+2)}{6(k-1)}$$

$$t_1 = (k-1)p(p+1)/2$$

$$t_2 = (t_1 + 2)/|e_2 - e_1^2|$$

$$b = \begin{cases} t_1/(1 - e_1 - t_1/t_2) & si \quad e_2 > e_1^2 \\ t_2/(1 - e_1 + 2/t_2) & si \quad e_2 < e_1^2 \end{cases}$$

suit une distribution de Fisher-Snedecor à t_1 et t_2 degrés de libertés.

Si $e_1^2 - e_2$ est proche de 0, alors SPSS utilise la statistique de Bartlett $(1 - e_1)M$ qui est approximativement distribuée comme un χ^2 à t_1 degrés de libertés (cf. [Saporta, 1990]).

Tableau 17 : Test de Box pour l'égalité des matrices de variance-covariance locales

Déterminants Log

type d'Iris	Rang	Déterminant Log
Setosa	4	5,416
Versicolor	4	7,650
Virginica	4	9,494
Intra-groupes combinés	4	8,503

Les rangs et logarithmes naturels des déterminants imprimés sont ceux des matrices de covariance du groupe.

Résultats du test

M de Box		144,520
F	Approximativement	6,942
	ddl1	20
	ddl2	77566,751
	Signification	,000

Teste l'hypothèse nulle de matrices de covariance à égales populations.

Les valeurs du logarithme des déterminants des matrices de variance-covariance s'interprètent comme le volume de l'ellipsoïde d'inertie relatif à chacun des groupes dans l'espace des variables explicatives de dimension 4. *Iris Virginica* apparaît donc comme l'espèce présentant le plus de variabilité relativement à ces quatre mesures, tandis que *Setosa* apparaît comme la plus homogène.

D'après les résultats du tableau 17, nous sommes conduits à rejeter l'hypothèse nulle d'égalité des matrices de variance-covariance entre les trois espèces d'Iris. Cependant, le test de Box étant réputé sensible au défaut de multinormalité, nous devons rester prudents par rapport à la conclusion du test. D'autre part, vu la taille de nos échantillons relativement au nombre de paramètres à estimer, les fonctions linéaires peuvent s'avérer plus robustes que des fonctions quadratiques.

iii) Affectation des individus selon la règle bayésienne

Le calcul des probabilités sur lequel est basée la règle d'affectation bayésienne s'effectue dans *SPSS* selon la procédure suivante. Soit, pour un individu à classer i_0, le vecteur $\tilde{x}$ des q mesures sélectionnées comme variables explicatives du classement. Le vecteur $\tilde{z}$ des coordonnées factorielles discriminantes de l'individu i_0 est obtenu par projection sur les h fonctions linéaires discriminantes : $\tilde{z} = \tilde{x}b$. Soit $\tilde{g}_l$ est le vecteur des h coordonnées factorielles du groupe G_l. Dans le sous-espace des h premières fonctions linéaires discriminantes, sous l'hypothèse de multinormalité, le carré de la distance généralisée $\delta_l^2 = \left(\tilde{x} - \tilde{g}_l\right)' D_l^{-1} \left(\tilde{x} - \tilde{g}_l\right)$ entre l'individu i_0 et le barycentre du groupe G_l, suit une distribution du χ^2 à h degrés de libertés.

La probabilité conditionnelle $P\left(\tilde{x} \middle| G_l\right)$ d'observer le vecteur $\tilde{x}$ sachant que l'individu appartient au groupe G_l est le seuil de risque associé au quantile $\chi_l^2 = \delta_l^2$.

La probabilité a posteriori d'appartenance $P\left(G_l \middle| \tilde{x}\right)$ au groupe G_l est calculée selon la formule :

$$P\left(G_l \middle| \tilde{x}\right) = \frac{p_l \left|D_l\right|^{-1/2} e^{-\delta_l^2}}{\sum_{j=1}^{h} p_j \left|D_j\right|^{-1/2} e^{-\delta_j^2}}$$

où D_l est la matrice de variance-covariance locale au groupe G_l pour les h premières fonctions discriminantes et $\left|D_l\right|$ son déterminant.

Suivant l'option d'estimation choisie, matrice de variance-covariance intraclasse commune ou matrices de variance-covariance locales à chacun des groupes, le calcul du carré de la distance de Mahanalobis s'effectue soit avec la métrique W^{-1}, soit avec les métriques locales W_l^{-1} à chacun des groupes. La règle d'affectation bayésienne conduit donc respectivement soit à une règle de classement quadratique (cf. la fonction QS en [8]) avec des matrices de variance-covariance D_l distinctes pour chaque groupe, soit à une règle de classement linéaire (cf. la fonction LS en [9]) avec une matrice unique $D = I$, égale à la matrice-identité puisque le calcul s'effectue dans l'espace des fonctions discriminantes.

L'application de la règle bayésienne d'affectation à chaque individu de l'échantillon d'apprentissage selon cette procédure conduit aux résultats listés dans le tableau 18. Dans le cas présent, il s'agit de la règle linéaire correspondant à une métrique unique W^{-1}.

Ce tableau, intitulé *« Diagnostic des observations »* donne les résultats du classement pour chaque individu de l'échantillon :

- La première colonne donne le numéro de séquence i_0 de l'individu permettant son identification ;
- La seconde colonne affiche le « *groupe effectif »* d'affectation de l'individu i_0, c'est à dire l'espèce à laquelle il appartient réellement ;
- La troisième colonne propose le « *groupe prévu »* par la procédure de classement selon la règle bayésienne. L'individu i_0 est affecté au groupe l_{max} possédant la plus forte probabilité

a posteriori $P(G = l_{\max} / D = d)$ connaissant la valeur *d* de la distance de Mahanalobis entre l'individu i_0 et le barycentre g_{max} du groupe d'affectation. Les individus mal classés (e.g. individus n° 5 et 9) sont signalés par une double astérisque ;

Tableau 18 : affectation des individus aux groupes, probabilités et scores (extrait).

Diagnostic des observations

			Plus grand groupe					Deuxième plus grand groupe			Scores discriminants	
	Nombre d'observations	Groupe effectif	Groupe prévu	P(D>d | G=g) p	P(D>d | G=g) ddl	P(G=g | D=d)	Carré de la distance de Mahalanobis au barycentre	Groupe	P(G=g | D=d)	Carré de la distance de Mahalanobis au barycentre	Fonction 1	Fonction 2
Original	1	1	1	,937	2	1,000	,130	2	,000	89,404	-7,637	-,131
	2	3	3	,601	2	1,000	1,017	2	,000	26,527	6,768	,611
	3	2	2	,728	2	,995	,634	3	,005	11,327	2,552	-,484
	4	3	3	,288	2	1,000	2,492	2	,000	30,038	6,633	1,839
	5	3	2**	,132	2	,730	4,054	3	,270	6,039	3,801	-,957
	6	1	1	,922	2	1,000	,163	2	,000	82,131	-7,193	,383
	7	3	3	,262	2	1,000	2,679	2	,000	18,563	5,105	2,020
	8	2	2	,155	2	,959	3,726	3	,041	10,054	3,485	-1,683
	9	2	3**	,100	2	,758	4,603	2	,242	6,892	3,692	1,077
	10	1	1	,439	2	1,000	1,649	2	,000	111,893	-8,650	,904

- La quatrième colonne donne $P(D > d / G = l_{\max})$ la probabilité conditionnelle d'observer une valeur au moins aussi extrême de la distance de Mahanalobis supposant l'appartenance au groupe le plus probable ;
- La cinquième colonne affiche la probabilité a posteriori maximale, $P(G = l_{\max} / D = d)$.
- La sixième colonne indique la valeur $d_q^2(x_0; g_{\max})$ du « *Carré de la distance de Mahanalobis* » au barycentre du groupe le plus probable ;
- La septième colonne donne le numéro du *second groupe le plus probable* ;
- La huitième colonne affiche la *seconde plus forte probabilité a posteriori* ;
- La neuvième colonne indique la valeur du *carré de la distance de Mahanalobis* au barycentre du second groupe le plus probable ;
- La dixième et onzième colonne indiquent les valeurs (« *Scores discriminants* ») des coordonnées factorielles pour les deux axes discriminants.

iv) Validation

Les **taux de bien-classés** constituent une mesure immédiate des performances de ***RC*** la règle de classement élaborée. On attend bien sûr des résultats meilleurs que ceux produit par une règle aléatoire, soit pour 3 groupes supérieurs à 33 %. L'objectif d'un classement quasi-parfait, taux proche de 100 %, est difficile à atteindre dans la plupart des contextes d'application réelle des méthodes de discrimination. Certains contextes supposent également la recherche de taux de bien-classés qui soient équilibrés entre les groupes : une règle optimale pour un groupe et désastreuse pour les autres groupes peut se révéler d'un intérêt limité dans la pratique. D'autre part, la pertinence du classement obtenu doit être contrôlée par des procédures de validation car des taux de bien-classés satisfaisants sur l'échantillon d'apprentissage ne garantissent pas la performance réelle de la règle de classement appliquée à de nouveaux échantillons.

Les taux de bien-classés et de mal-classés calculés sur l'échantillon d'apprentissage sont appelés des **taux de resubstitution** ou **taux apparents**. Le **taux réel de bien-classés *TBC*** est la probabilité que la règle *RC* classe correctement les individus d'un nouvel échantillon extrait de la même population et le **taux réel de mal-classés *TMC*** est la probabilité de classement incorrect. Les taux de resubstitution surestiment les taux réels de bien-classés et sous-estiment les taux réels de mal-classés car les performances de la règle de classement peuvent être influencées par les particularités de l'échantillon d'apprentissage.

Dans le cas de distributions multinormales et **homoscédastiques** (homogénéité des matrices de variance-covariance des groupes), les taux théoriques de mal-classés peuvent être estimés (pour le cas de deux groupes, cf. [Bardos, 2001]). Cependant, dans la plupart des applications pratiques, on ne connaît pas les lois de probabilité conditionnelles de la distribution des valeurs observées au sein des groupes. On est alors contraint d'utiliser des procédures empiriques de validation.

Une première procédure de validation envisageable est celle de **l'échantillon-test**, si l'on dispose d'un échantillon **indépendant** de l'échantillon d'apprentissage. Cette procédure, appelée **échantillon-test**, a l'avantage de fournir des estimations non biaisées un pourcentage de bien-classés et de mal-classés.

Sinon, on utilise la **validation croisée** qui consiste à partitionner l'échantillon de base *E* en *v* sous-ensembles $E_1, \cdots, E_l, \cdots, E_v$ qui serviront tour à tour d'échantillons-tests. À chaque pas *l* de la procédure de validation, le complémentaire $E \cap \overline{E_l}$ sert d'échantillon d'apprentissage. L'estimation du taux de mal-classés est donnée par la moyenne des taux de mal-classés obtenus à chaque étape. La **validation croisée systématique** (« *leave one out* ») est un cas particulier où chacun des sous-ensembles tests E_l est réduit à un seul individu. On effectue alors *n* discriminations portant sur *n-1* observations et excluant tour à tour chacun des individus de l'échantillon de base *E*. Cependant, pour *n* grand, les taux moyens de bien-classés convergent vers les taux de resubstitution.

Tableau 19 : pourcentage de bien-classés et validation croisée

Résultats du classement[b,c]

			Classe(s) d'affectation prévue(s)			
		type d'Iris	Setosa	Versicolor	Virginica	Total
Original	Effectif	Setosa	50	0	0	50
		Versicolor	0	48	2	50
		Virginica	0	1	49	50
	%	Setosa	100,0	,0	,0	100,0
		Versicolor	,0	96,0	4,0	100,0
		Virginica	,0	2,0	98,0	100,0
Validé-croisé[a]	Effectif	Setosa	50	0	0	50
		Versicolor	0	48	2	50
		Virginica	0	1	49	50
	%	Setosa	100,0	,0	,0	100,0
		Versicolor	,0	96,0	4,0	100,0
		Virginica	,0	2,0	98,0	100,0

a. La validation croisées n'est effectuée que pour les observations de l'analyse. Dans la validation croisée, chaque observation est classée par les fonctions dérivées de toutes les autres observations.

b. 98,0% des observations originales classées correctement.

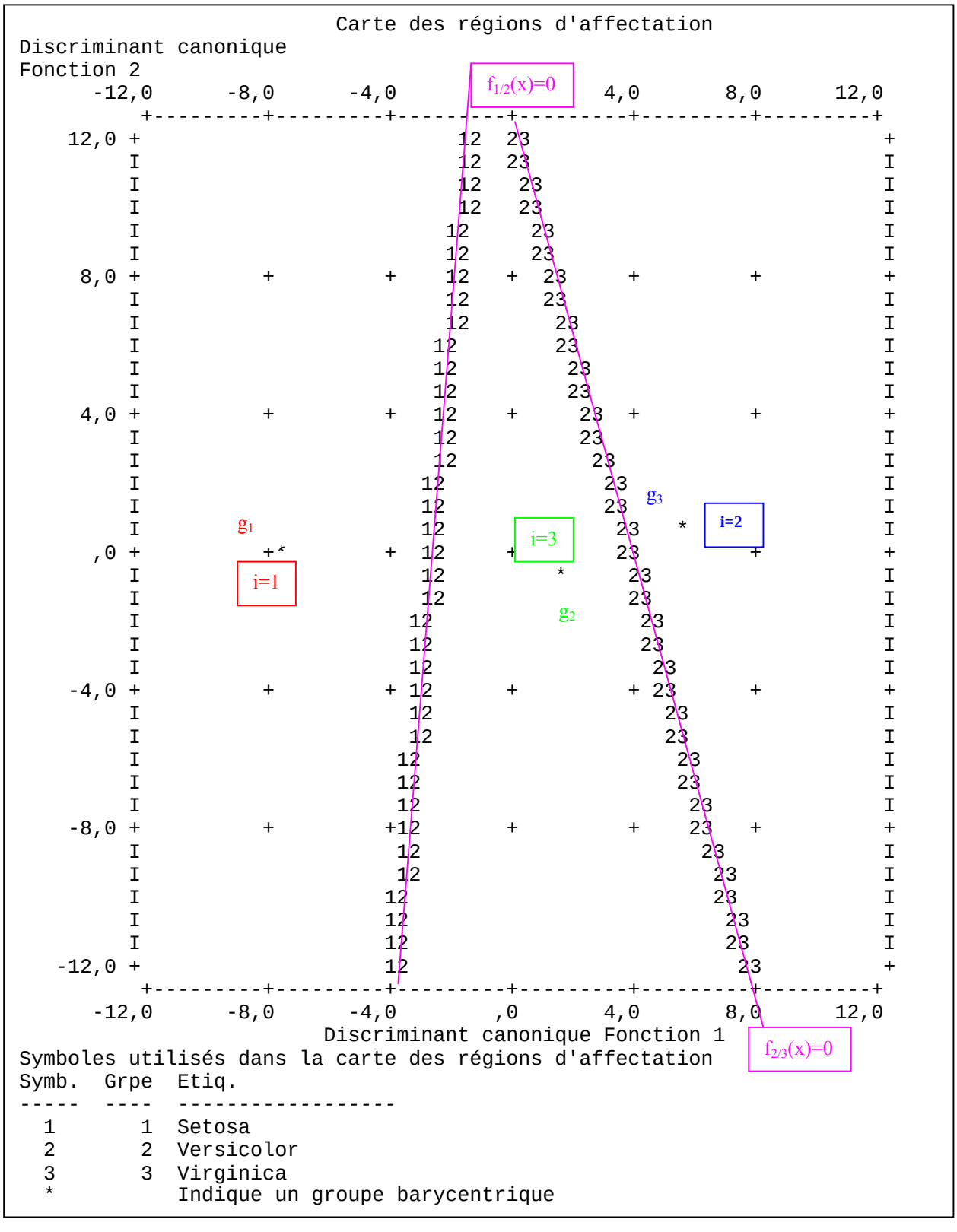
c. 98,0% des observations validées-croisées classées correctement.

Les résultats du classement effectué par la procédure DISCRIM de *SPSS* selon la règle bayésienne d'affectation montrent un taux apparent global de bien-classés élevé (98%), moyenne pondérée du taux apparent de bien-classés pour chacun des groupes qui varie de 100% pour le groupe *Setosa* à 96% pour le groupe *Versicolor*, comportant le plus d'erreur d'affectation.. Le calcul des probabilités a posteriori utilisées dans la règle bayésienne d'affectation n'étant fonction que de la valeur des distances généralisées des individus aux barycentres des groupes, la procédure DISCRIM propose une validation croisée systématique qui ne nécessite pas d'effectuer *n* analyses discriminantes distinctes mais de réajuster le calcul des distances de Mahanalobis. Dans notre cas particulier, la validation croisée systématique aboutit aux mêmes estimations que la resubstitution mais ce n'est pas toujours le cas, a fortiori pour les petits échantillons.

v) Utilisation des graphiques factoriels discriminants

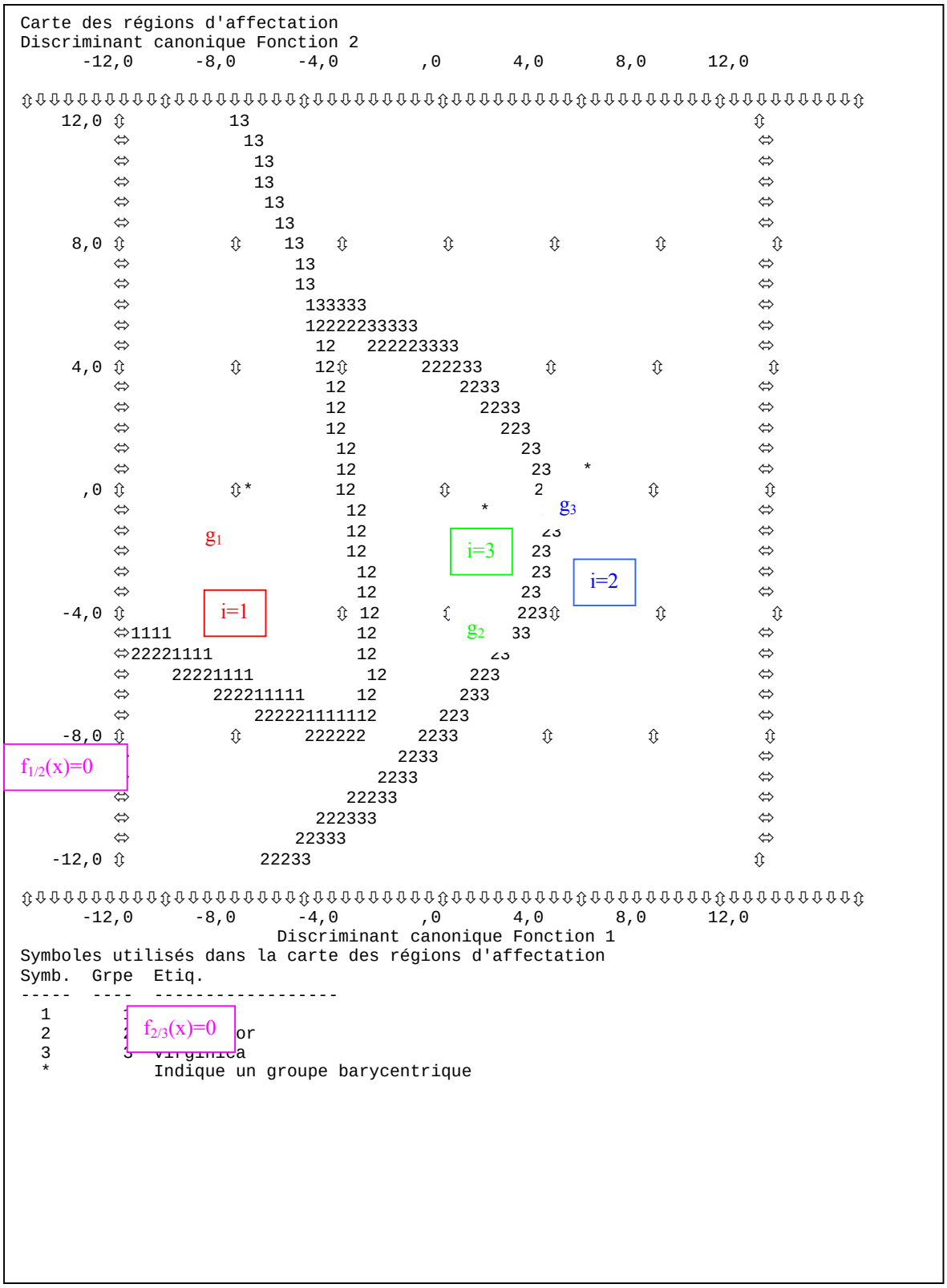
Le graphique de la figure 8 produit par la procédure, intitulé « cartes des régions d'affectation », est une projection dans le premier plan factoriel croisant les deux premiers axes factoriels discriminants correspondant aux deux premières fonctions linéaires discriminantes. Les projections des barycentres des groupes sont indiquées par une étoile et les régions d'affectation résultant de l'application de la règle bayésienne de décision sont délimitées par le tracé des hyperplans $f_{l/l'}(\tilde{x}) = 0$ au moyen des numéros des classes *l* et *l'*.

Figure 8 : carte territoriale des régions d'affectation, discrimination linéaire.



Par contre, pour la figure 9, les limites territoriales ne sont plus fournies par des **hyperplans** médiateurs (par exemple, une droite dans un plan) issus d'une discrimination linéaire mais par des **variétés** (par exemple, une courbe dans le plan) issues de la discrimination quadratique. L'utilisation des matrices locales de variance-covariance n'améliore pas les résultats puisque nous obtenons des pourcentages de bien-classés identiques au niveau de l'échantillon d'apprentissage.

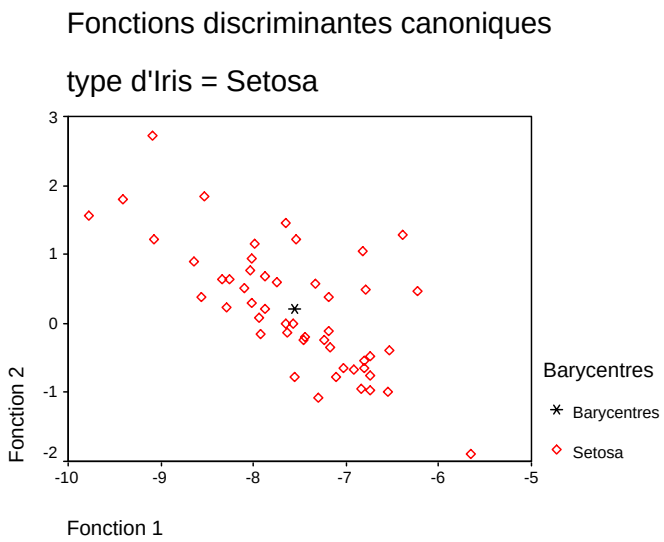
Figure 9 : carte territoriale des régions d'affectation, discrimination quadratique.



La carte territoriale des régions d'affectation de la figure 8 illustre le fait qu'individu ayant une coordonnée factorielle sur le premier axe discriminant inférieure à – 4 sera classé comme *Iris Setosa* (groupe G_1) par la procédure d'affectation. Par contre, si l'individu possède sur le premier axe discriminant une coordonnée factorielle supérieure à 8, il sera classé comme *Iris Virginica* (groupe G_3). Si cette coordonnée est comprise entre – 4 et 8, alors son classement dépend de la valeur de la coordonnée factorielle sur le deuxième axe discriminant.

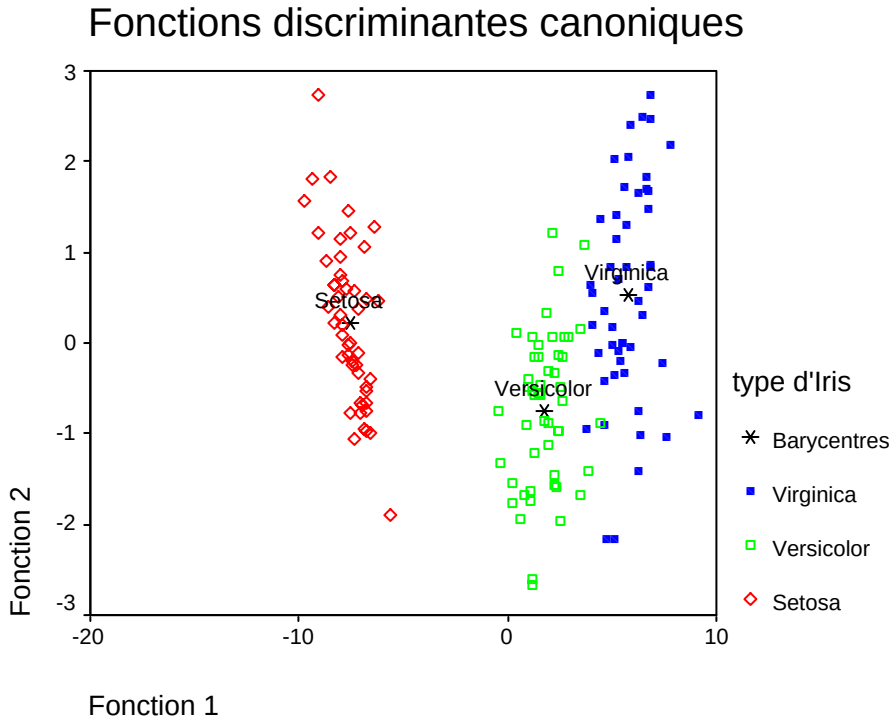
Des projections graphiques des individus et du barycentre peuvent également être obtenues groupe par groupe sur le plan des deux premiers axes discriminants correspondant aux deux premières fonctions linéaires discriminantes, comme pour le groupe des *Iris Setosa* en figure 10 :

<u>Figure 10 :</u> projection locale dans le premier plan factoriel discriminant des individus et du barycentre du groupe des *Iris Setosa*.



On peut également demander un graphique « toutes classes combinées » projetant l'ensemble des individus et des barycentres de groupes dans le plan des deux premiers axes discriminants (cf. figure 11) :

<u>Figure 11 :</u> projection globale dans le premier plan factoriel discriminant des individus et des barycentres des groupes de l'échantillon d'apprentissage.

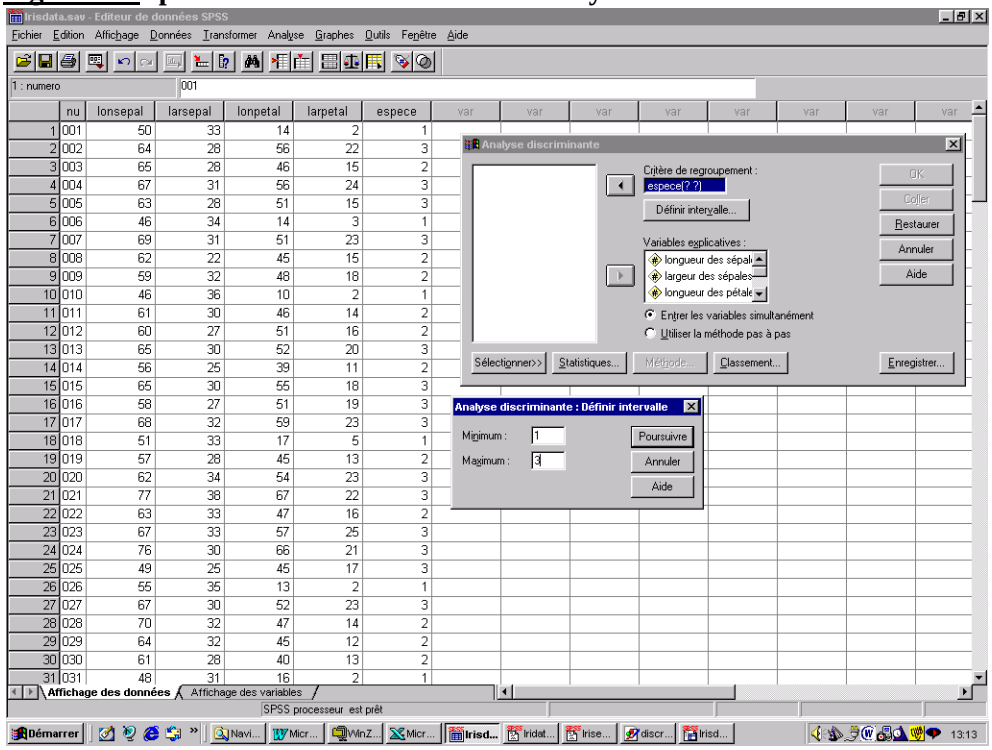


IV) Spécification des paramètres de l'analyse discriminante

i) Spécifications globales

Pour effectuer une analyse discriminante, il convient de sélectionner la procédure Analyse discriminante de l'option Classification du menu Statistiques afin d'obtenir la boîte de dialogue permettant de spécifier les principaux paramètres de l'analyse discriminante :

Figure 12 : spécification des variables de l'analyse discriminante.



Les spécifications requises concernent les descripteurs (Variables explicatives : larpetal, larsepal, lonpetal, lonsepal) et le critère qualitatif à prédire indiquant le groupe d'appartenance des individus (Critère de regroupement : espece). En cliquant sur le bouton Définir intervalle..., on spécifiera les modalités (? ?) qui doivent être codées selon une séquence ordonnée de valeurs numériques (ici, de 1 à 3, soit les valeurs 1, 2 et 3). Les observations présentant des valeurs du critère de regroupement situées en dehors de cet intervalle seront exclues de l'analyse. Il faut au moins deux groupes distincts non vides d'observations et au moins un descripteur pour que la procédure s'exécute.

L'analyse discriminante définie par défaut fournit les coefficients des fonctions discriminantes canoniques standardisées, la matrice de structure des fonctions discriminantes et de tous les descripteurs (qu'ils soient inclues ou non dans l'équation) ainsi que les valeurs moyennes des fonctions discriminantes pour chacun des groupes. Les observations présentant des valeurs manquantes pour les variables explicatives sont exclues des étapes d'estimation de l'analyse (calcul des coefficients et des indicateurs statistiques de base).

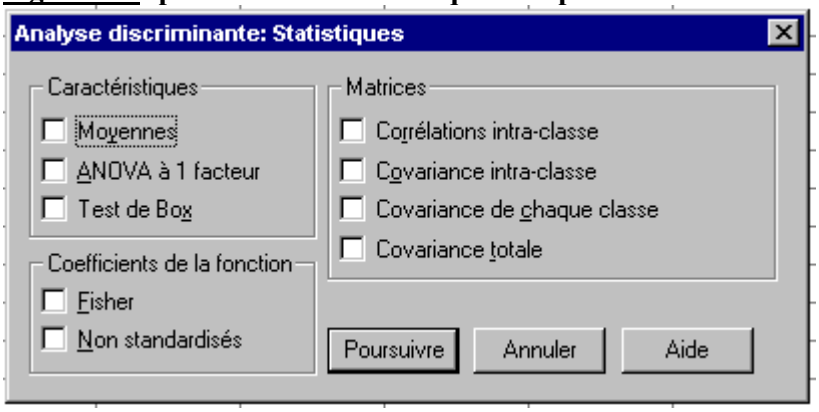
On peut choisir entre les deux méthodes suivantes de sélection des descripteurs comme variables explicatives :

- **Entrer les variables simultanément.** Sélection a priori des descripteurs. C'est l'option par défaut. Toutes les variables satisfaisant le critère de tolérance sont inclues comme variables explicatives dans l'équation ;
- **Utiliser la méthode pas à pas.** Sélection pas à pas des descripteurs. Le critère de sélection du pas à pas minimise le lambda de Wilks global.

ii) Statistiques

Pour obtenir des indicateurs statistiques complémentaires, tels que des statistiques descriptives, les coefficients de certaines fonctions ou la structure des matrices utilisées, il suffit de cliquer sur le bouton Statistiques... pour valider vos choix dans la boîte de dialogue correspondante de l'analyse discriminante :

Figure 13 : spécification des statistiques complémentaires.



Caractéristiques. Choix multiple concernant les indicateurs de statistique descriptive demandés en complément :

- **Moyennes.** Moyennes et écarts-types globaux (ensemble des données) et locales (par groupe) pour chacune des variables explicatives (cf. tableau 1) ;
- **ANOVA à 1 facteur.** Tests d'analyse de la variance à un facteur sur l'égalité des moyennes pour chacune des variables explicatives (cf. tableau 2) ;
- **Test de Box.** Test M de Box sur l'égalité des matrices de variance-covariance locales à chacun des groupes (cf. tableau 17).

Coefficients de la fonction. Choix multiple concernant les coefficients standardisés ou non des fonctions linéaires discriminantes :

- **Fisher.** Coefficients standardisés (cf. tableau 12) des fonctions linéaires discriminantes permettant d'affecter les observations à chacun des groupes dans l'espace des variables centrées réduites au moyen d'une équation sans constante ;
- **Non standardisés.** Coefficients non standardisés (cf. tableau 11) des fonctions linéaires discriminantes pour le classement des observations dans l'espace des variables d'origine au moyen d'une équation avec constante.

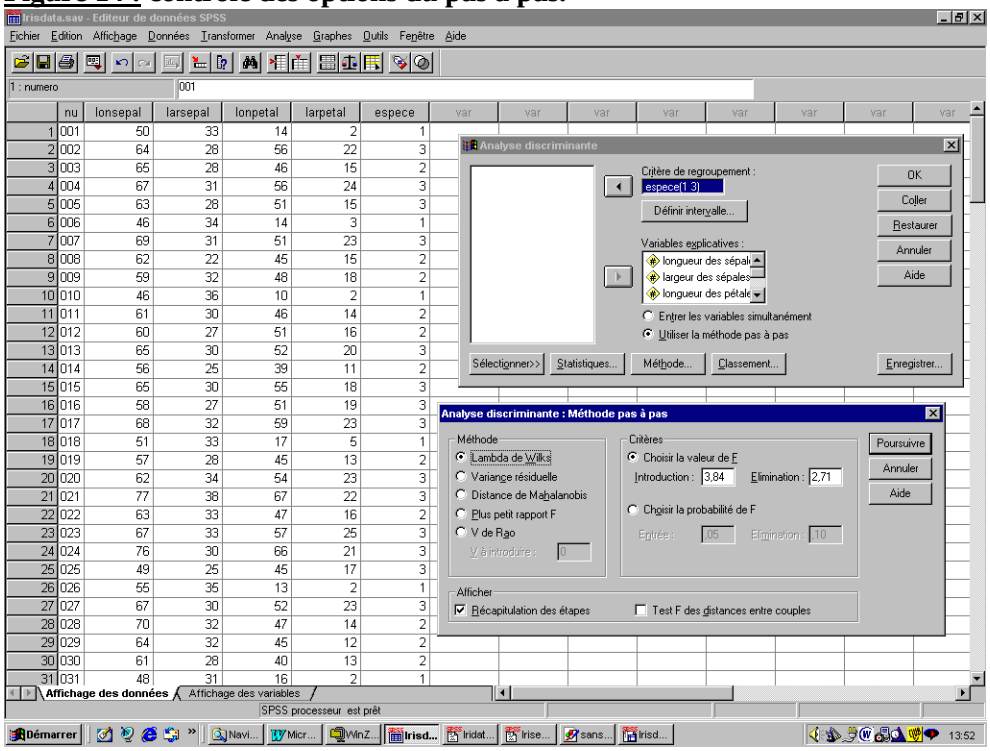
Matrices. Choix multiple concernant les matrices des différents opérateurs d'inertie :

- **Corrélation intraclasse.** Matrice globale de corrélations intraclasses (cf. tableau 4) ;
- **Inertie intraclasse.** Matrice globale d'inertie intraclasses ou de variance-covariance intraclasses, dans le cas particulier de la pondération uniforme (cf. tableau 4) ;
- **Inertie de chaque classe.** Matrices locale d'inertie (ou de variance-covariance, dans le cas particulier de la pondération uniforme) pour chacun des groupes (cf. tableau 5) ;
- **Inertie totale.** Matrice globale d'inertie (ou de variance-covariance totale, dans le cas particulier de la pondération uniforme) calculée sur l'ensemble de l'échantillon (cf. tableau 5).

iii) Sélection des variables

Si l'on souhaite définir les options de la procédure de pas à pas et contrôler l'inclusion des variables exogènes dans l'équation de prédiction, il suffit de sélectionner l'option Utiliser la méthode pas à pas, puis cliquer sur le bouton Méthodes... dans la boîte de dialogue principale :

Figure 14 : contrôle des options du pas à pas.



Méthode. On peut sélectionner l'une des méthodes suivantes :

- **Lambda de Wilks.** Sélection à chaque étape de la variable qui minimise le lambda de Wilks ;
- **Variance résiduelle.** Sélection à chaque étape de la variable qui minimise la variance résiduelle (inexpliquée par les différences entre groupes) ;
- **Distance de Mahanalobis.** Sélection à chaque étape de la variable qui minimise la distance de Mahanalobis entre les deux groupes les plus proches ;
- **Plus petit rapport F.** Sélection à chaque étape de la variable qui minimise le rapport F sur l'ensemble des groupes pris deux à deux ;
- **V de Rao.** Sélection à chaque étape de la variable qui minimise le V de Rao.

V à introduire. Par défaut, la borne minimum du V pour introduire une variable est fixée à 0. Il suffit d'entrer une nouvelle valeur pour changer ce paramétrage.

Critères. On peut sélectionner l'un des critères suivants :

- **Choisir la valeur de F.** Utilise comme critère la valeur du rapport F, avec une valeur par défaut pour la borne d'introduction (3,84) et pour la borne d'élimination (2,71). Ce paramétrage peut être changé en spécifiant de nouvelles valeurs strictement positives, la borne d'introduction devant être supérieure à la borne d'élimination ;
- **Choisir la probabilité de F.** Utilise comme critère la probabilité du rapport F, avec une probabilité par défaut pour la borne d'introduction (0,05) et pour la borne d'élimination (0,10). Ce paramétrage peut être changé en spécifiant de nouvelles valeurs comprises entre 0 et 1, la borne d'introduction devant être supérieure à la borne d'élimination.

iv) Affichage des résultats

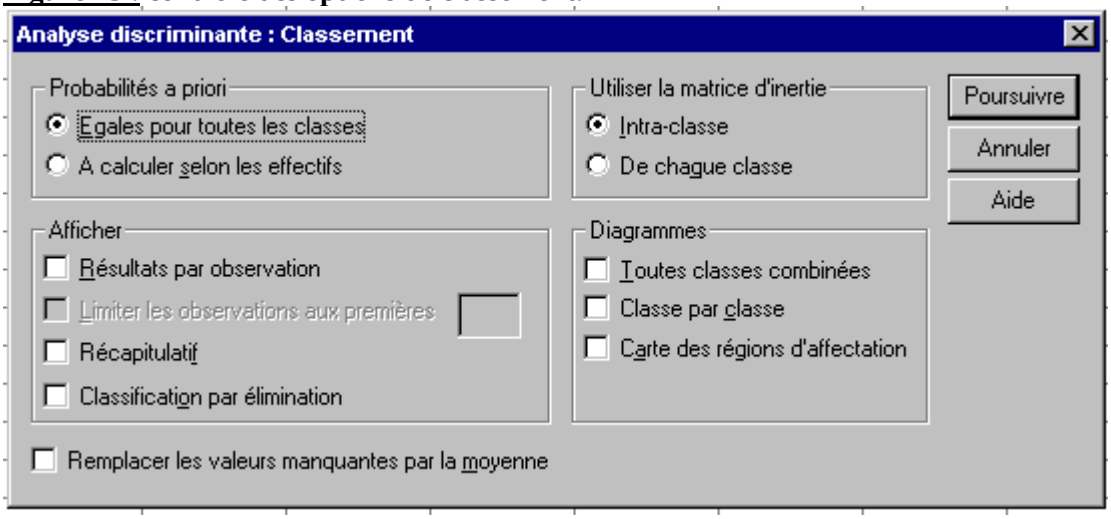
Afficher. On peut choisir l'un des options suivantes :

- **Récapitulation des étapes.** Pour la méthode du pas à pas, des résultats récapitulatifs sont produits à chaque étape. Les statistiques récapitulatives sont le lambda de Wilks, le pseudo-F, les degrés de liberté et le niveau de signification du F. La tolérance, le F d'élimination et la valeur de la statistique utilisée pour sélectionner la variable sont produites pour chaque variable incluse dans l'équation. La tolérance, la tolérance minimum, le F d'introduction et la valeur de la statistique utilisée pour sélectionner la variable sont produites pour chaque variable exclue de l'équation. Pour supprimer ce récapitulatif, il suffit de désélectionner l'option ;
- **Test F des distances entre couples.** Produit une matrice des rapports F pour les groupes pris deux à deux. Ces rapports F permettent de tester la valeur de la distance de Mahanalobis entre deux groupes.

v) Options de classement

Pour définir les options de classement (ou d'affectation) et contrôler l'inclusion des variables exogènes dans l'équation de prédiction, il suffit de cliquer sur le bouton Classement… dans la boîte de dialogue principale :

Figure 15 : contrôle des options de classement.



Probabilités a priori. On peut choisir l'une des options suivantes :

- **Egales pour toutes les classes.** Les probabilités a priori d'appartenance au groupe sont supposées égales, option par défaut ;
- **A calculer selon les effectifs.** Les probabilités a priori d'appartenance au groupe sont estimées d'après les proportions d'individus dans chaque groupe, observées sur l'échantillon.

Afficher. On peut sélectionner l'un des choix d'affichage suivants :

- **Résultats par observation.** Affiche pour chaque observation les codes des groupes observés et estimés, les probabilités a posteriori et les scores discriminants (cf. tableau 18) ;
- **Récapitulatif.** Affiche une table de classement récapitulant les appartenances observées et estimées pour chacun des groupes (cf. tableau 19). Si un échantillon d'apprentissage est sélectionné deux tables seront affichées, l'une pour les individus échantillonnés, l'autre pour le reste des individus ;
- **Classification par élimination.** Affiche les résultats du classement effectué selon la méthode de validation croisée systématique (*leave one out*) : pour chaque individu extrait de l'échantillon, une fonction score est construite sur un échantillon d'apprentissage comportant les *(n-1)* individus restants ; chacun de ces *n* scores est testé sur l'individu éliminé.

Utiliser la matrice d'inertie. On peut choisir l'une des options suivantes :

- **Intraclasse.** On utilise la matrice globale de variance-covariance intraclasses pour classer les individus, option par défaut (cf. règle de classement linéaire [9]) ;
- **De chaque classe.** On utilise la matrice de variance-covariance locale à chaque groupe pour classer les individus (cf. règle de classement quadratique [8]).

Le classement est basé sur les fonctions discriminantes et non sur les variables explicatives.

Diagrammes. Choix multiple parmi les diagrammes suivants :

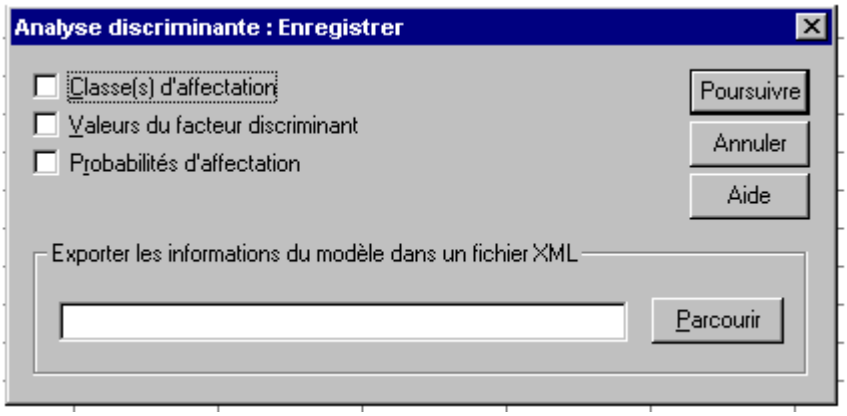
- **Toutes classes combinées.** Projection graphique de l'ensemble des individus et des groupes dans le plan des deux premiers axes discriminants (cf. figure 11). S'il n'existe qu'une seule fonction discriminante, un histogramme remplace le diagramme ;
- **Classe par classe.** Production d'un diagramme distinct pour chaque groupe dans le plan des deux premiers axes discriminants (cf. figure 10). S'il n'existe qu'une seule fonction discriminante, des histogrammes remplacent les diagrammes ;
- **Carte des régions d'affectation.** Une carte territoriale (cf. figure 8,9) délimite les frontières utilisées pour le classement et indique le barycentre de chaque groupe. S'il n'existe qu'une seule fonction discriminante, cette carte n'est pas produite.

- **Remplacer les valeurs manquantes par la moyenne.** Lors du processus de classement, les moyennes sont substituées aux valeurs manquantes pour les descripteurs afin de pouvoir classer les individus avec des valeurs manquantes.

vi) Options d'enregistrement

Pour définir les options de sauvegarde, il suffit de cliquer sur le bouton Enregistrer... dans la boîte de dialogue principale :

Figure 16 : contrôle des options d'enregistrement.



Choix multiples concernant les options d'enregistrement :

- **Classe(s) d'affectation.** Affectation au groupe possédant la plus grande probabilité a posteriori ;
- **Valeur du facteur discriminant.** Sauvegarde des scores pour chacune des fonctions discriminantes estimées ;
- **Probabilités d'affectation.** Création d'une variable par groupe. La k^e variable contient la probabilité a posteriori d'appartenance au k^e groupe.

V) Références bibliographiques

Bardos M. (2001). *Analyse discriminante : application au risque et scoring financier*, Dunod, Paris, 224 p.

Blake, C.L. & Merz, C.J. (1998). *UCI Repository of machine learning databases*. University of California, Department of Information and Computer Science Irvine, CA, USA.

Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. (1989). *Classification automatique des données. Environnement statistique et informatique*, Bordas, Paris, 285 p.

Dagnelie P. (1975) *Analyse statistique à plusieurs variables*, Les Presses agronomiques de Gembloux, Gembloux, 362 p.

Fisher, R. A.. (1936). « The Use of Multiple Measurements in Taxonomic Problems », *Ann. of Eugenics 7*, Part II, pp. 179-188.

Klecka W.R. (1980). *Discriminant Analysis*, Sage Publications, Beverly Hills, 88p.

Lebart L., Morineau A., Piron M. (1995) *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 439 p.

Mahanalobis P.C. (1936) « On the Generalised Distance in Statistics », *Proc. Nat. Inst. Sci. India*, 12, pp. 49-55.

Romeder J.-M. (1973) *Méthodes et programmes d'analyse discriminante*, Dunod, Paris, 274 p.

SPSS Inc. (1994). *SPSS 6.1 Professional Statistics*, SPSS Inc., Chicago, 385 p.

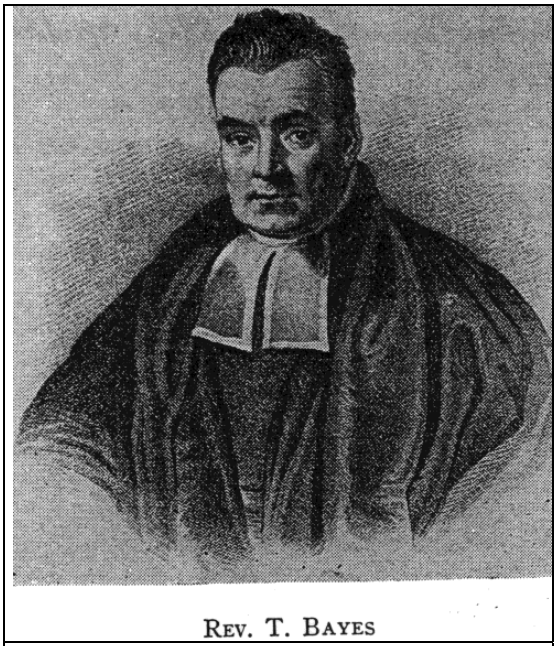
SPSS Inc. (1999). *SPSS 10.0 Applications Guide*, SPSS Inc., Chicago, 426 p.

Saporta G. (1990). *Probabilité, analyse des données et statistique*, Technip, Paris, 493 p.

Tomassone R. 1988, « Comment interpréter les résultats d'une analyse factorielle discriminante ? », ITCF, Paris, 56 p.

Tomassone R., Danzart M., Daudin J.J., Masson J.P. (1988) *Discrimination et classement*, Masson, Paris, 174 p.

<u>Figure 17 :</u> portrait présumé du révérend Thomas Bayes.



Rev. Thomas BAYES
1702-1761

[extrait de *History of Life Insurance in its Formative Years*,
par T O'Donnell, American Conservation Company
Chicago, 1936]