Notions sur l'analyse discriminante

Laurent Carraro, Anca Badea

ENSM.SE

Axe Méthodes Statistiques et Applications

novembre 2007

Table des matières

1 Exemple : les iris de Fisher

Ces données sont issues d'une étude du botaniste Anderson [1] et ont été utilisées en 1937 par le célèbre statisticien Sir Ronald Fisher [2] pour démontrer la pertinence de ses méthodes, dites d'analyse discriminante. Elles sont devenues depuis partie intégrante du folklore statistique et rares sont les ouvrages traitant d'analyse discriminante qui n'évoquent pas cet exemple. Nous ne dérogerons pas à cette règle ici.

Les données sont celles de 150 mesures faites sur des fleurs d'iris de diverses variétés. De manière précise, 4 mesures sont effectuées sur chaque fleur : largeur et longueur du sépale, largeur et longueur du pétale, et on détermine par ailleurs la variété de la fleur : setosa, versicolor, virginica. On cherche alors à identifier les caractères ou les combinaisons de caractères qui permettent de distinguer au mieux les espèces d'iris. Il s'agit donc des quantités qui discriminent le plus les espèces, d'où le terme générique utilisé d'analyse discriminante.

1.1 Les problèmes posés

Dans le contexte de notre exemple, deux types de questions se posent naturellement. Elles peuvent sembler très proches en apparence car elles concourent au même objectif. On va voir pourtant qu'elles sont de nature profondément différente.

1. Analyse discriminante descriptive

Il s'agit ici, comme lorsque l'on réalise une ACP, de représenter les données dans un espace ad hoc, qui permette de bien mettre en évidence les variables liées à l'espèce de la fleur. En d'autres termes, les techniques qui seront mises sur pied vont chercher à répondre aux questions qui suivent.

Quelles variables, quels groupes de variables, quels sous-espaces discriminent-ils au mieux les 3 espèces d'iris ?

On retrouve souvent ce type de problème dans des applications médicales ou financières (recherche de facteurs de risques [3] par exemple), ou socio-économiques. Donnons un exemple simple.

[1] E. Anderson. *The irises of the Gaspé peninsula*, Bulletin of the American Iris Society, 59, p. 2-5 (1935).

[2] R.A. Fisher. *The use of multiple measurements in taxonomic problems*, Annals of Eugenics, 7, p. 179-188 (1936).

[3] La signification du terme risque est évidemment différente en finance et en santé !

Imaginons un responsable politique qui cherche à identifier les facteurs de succès ou d'échec au lycée pour diverses populations d'adolescents. Son souci dans un tel cas n'est pas de prévoir pour un adolescent donné ses chances de succès ou d'échec mais plutôt [4] d'identifier les facteurs, ou les combinaisons de facteurs, les plus influents afin d'orienter ses actions ultérieures.

2. Analyse discriminante décisionnelle

On cherche ici à affecter une nouvelle fleur à une espèce en connaissant les valeurs des 4 variables quantitatives qui la décrivent. On voit ici que l'on est passé d'un objectif de description à un objectif de prévision ; c'est pourquoi on verra de nombreux concepts probabilistes utilisés pour traiter de ces questions.

Les domaines d'application de l'analyse discriminante décisionnelle sont tellement vastes qu'il serait illusoire d'en faire un recensement. Notons en premier lieu la très riche théorie de la reconnaissance des formes qui a pour objectif de reconnaître des objets à partir d'acquisition d'images ou de sons. Un exemple courant de l'utilisation de cette technique est celui du traitement automatisé des codes postaux indiqués sur les lettres. Un système optique et informatique pixellise le code postal, charge à un algorithme de reconnaître les chiffres qui composent le code. Si l'on modélise pour simplifier la pixellisation d'un chiffre par la donnée d'une grille 20*20 de niveaux de gris, on voit qu'il s'agit de prévoir l'appartenance d'un individu à un groupe (chiffre 0, chiffre 1, ...) à partir de l'observations de 400 variables.

Ces précisions étant faites, revenons à nos données avec un premier graphique.

On observe déjà que les variables relatives aux pétales semblent davantage discriminer les espèces que les données de sépales et que l'espèce setosa semble facile à distinguer des deux autres. Par contre, versicolor et virginica paraissent assez mêlées.

2 L'analyse discriminante

2.1 Formalisation

On considère une population de n individus indexés par i, $1 \leq i \leq n$, chaque individu numéro i étant de poids p_i. Ces individus sont caractérisés par deux types de variables :

– p variables X_j, le plus souvent quantitatives

[4] Nous n'en sommes pas certains mais nous l'espérons !

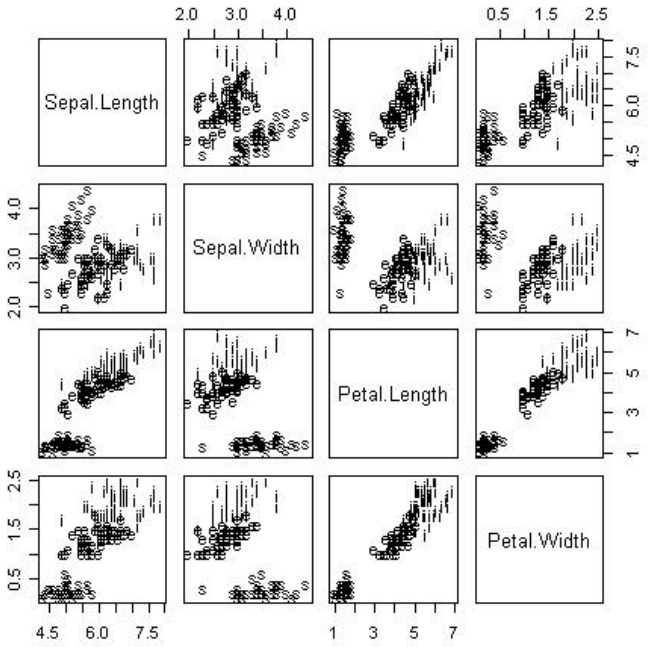

FIG. 1 – Iris de Fisher - s=setosa, e=versicolor, i=virginica

– l'appartenance à un groupe qui se traduit par une variable qualitative Y possédant m modalités y_h, $1 \leq h \leq m$.

On note G_h le groupe $\{i, Y(i) = y_h\}$. Comme dans l'exemple des iris, les techniques d'analyse discriminante visent à atteindre des objectifs de deux types : l'un à caractère descriptif et l'autre à caractère prévisionnel. Nous allons passer ici beaucoup de temps sur le premier objectif, renvoyant au cours de régression logistique pour les questions prévisionnelles.

2.2 Analyse discriminante descriptive

Ici, il s'agit avant tout d'identifier les variables, les groupes ou combinaisons de variables, qui "expliquent" au mieux l'appartenance au groupe d'un individu.

Donnons déjà quelques éléments sur les variables de notre exemple :

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Minimum	4.30	2.00	1.00	0.10
1er quartile	5.10	2.80	1.60	0.30
Médiane	5.80	3.00	4.35	1.30
Moyenne	5.84	3.06	3.76	1.20
3ème quartile	6.40	3.30	5.10	1.80
Maximum	7.90	4.40	6.90	2.50

Pour trouver les variables les plus discriminantes, une succession de boîtes à moustaches, ou boxplots, est très informative. Pour l'exemple des iris, on observe ainsi le graphique qui suit.

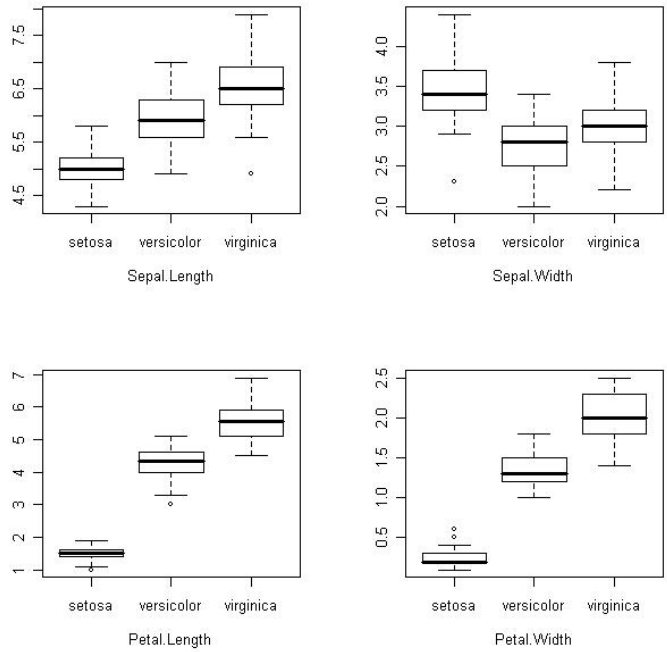

FIG. 2 – Boxplot des variables mesurées selon les espèces d'iris

Cette représentation donne les mêmes informations que la figure 1, de manière beaucoup plus lisible. On voit ainsi très simplement que les mesures concernant les pétales différentient mieux les espèces que celles qui concernent les sépales, et que toute mesure du pétale sépare très bien l'espèce setosa des deux autres. Il est par contre plus difficile de distinguer versicolor et virginica à partir de ces variables.

2.2.1 Pouvoir discriminant et estimation de densités

Partant de cette idée de représentation par espèce, on peut reproduire les densités lissées de chacune des 4 variables quantitatives, pour les 3 espèces.

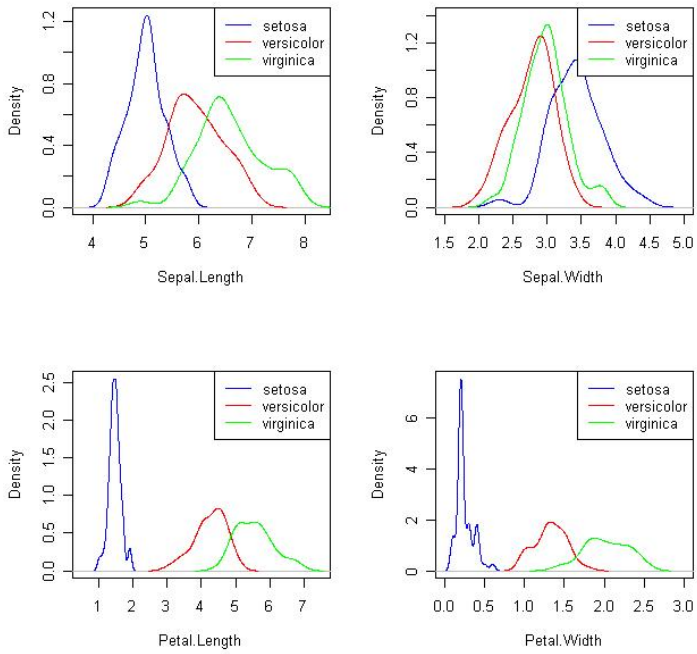

FIG. 3 – Densités estimées de la distribution de chaque variable selon les espèces

On retrouve les observations faites à partir des boxplots sous une forme plus quantitative puisque l'on peut ainsi déterminer de manière approximative le pourcentage de fleurs d'une espèce donnée, dont la longueur du pétale dépasse par exemple 4.9 cm, limite qui semble séparer le mieux possible les espèces versicolor et virginica.

On peut également comparer ces graphiques à ceux de la figure qui suit, qui donnent les distributions de chaque variable et fournissent ainsi une nouvelle façon d'appréhender le pouvoir discriminant d'une variable.

Ce graphique appelle plusieurs observations. Tout d'abord, un mélange de populations différentes ne se caractérise pas forcément par une densité multimodale (cf. les distributions des sépales), et même s'il y a plusieurs modes, leur nombre ne fait que minorer le nombre de groupes (cf. les distributions des pétales). Par ailleurs, la comparaison de ce graphique avec le précédent semble montrer que le caractère discriminant d'une variable est lié à la relation entre la dispersion totale de cette variable et celle observée dans chacun des 3 groupes. Ce qui nous mène

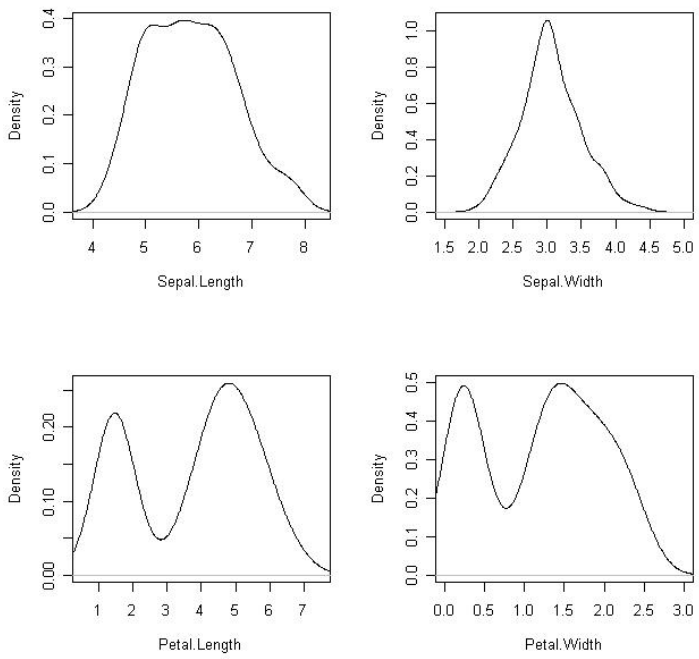

FIG. 4 – Densités estimées pour chaque variable, toutes espèces confondues

naturellement au paragraphe qui suit.

2.2.2 Pouvoir discriminant et variance

Si l'on mesure la dispersion avec une variance, ou plutôt un écart-type noté σ, nous sommes menés à construire le tableau qui suit, où par exemple σ_{set} désigne l'écart-type de la variable observée pour l'espèce setosa. La dernière colonne, intitulée *mean*, représente la moyenne des 3 rapports σ_{set}/σ, σ_{ver}/σ, σ_{vir}/σ.

variable	σ	σ_{set}	σ_{ver}	σ_{vir}	σ_{set}/σ	σ_{ver}/σ	σ_{vir}/σ	mean
Sepal.Length	0.83	0.35	0.52	0.64	0.43	0.62	0.77	0.61
Sepal.Width	0.44	0.38	0.31	0.32	0.87	0.72	0.74	0.78
Petal.Length	1.77	0.17	0.47	0.55	0.10	0.27	0.31	0.23
Petal.Width	0.76	0.11	0.20	0.27	0.14	0.26	0.36	0.25

TAB. 1 – Premiers indices évaluant le pouvoir discriminant de chaque variable

On voit clairement ici que le pouvoir discriminant d'une variable est lié à la façon dont la loi

de la variable considérée est concentrée du fait que l'observation de la variable est faite sur les seuls individus d'une espèce. Ainsi, plus la quantité *mean* est faible, plus la variable est utile pour discriminer les espèces. Et ceci est évidemment cohérent avec les observations précédentes faites sur les distributions des diverses variables.

Prenons maintenant un peu de recul et, comme nous l'avons fait sans l'expliciter pour l'instant, introduisons un modèle probabiliste. On considère que les individus indexés par i, $(1 \leq i \leq n)$ sont affectés d'un poids p_i, ce qui définit une probabilité P sur l'ensemble des individus (ici, P est uniforme sur l'ensemble des 150 iris). Sur cet espace probabilisé sont définies les variables aléatoires quantitatives X_j $(1 \leq j \leq p)$ et la variable aléatoire qualitative Y de modalités y_h, $1 \leq h \leq m$. Rappelons également la formule de la variance totale, appliquée à toute v.a. X_j et à la v.a. Y :

$$\mathrm{Var}(X_j) = E[\mathrm{Var}(X_j/Y)] + \mathrm{Var}[E(X_j/Y)] \tag{1}$$

Comme la variable Y est qualitative, les espérance et variance conditionnelles s'explicitent simplement, car :

$$P(Y = y_h)E(X_j/Y = y_h) = \sum_{Y(i)=y_h} p_i X_j(i)$$
$$P(Y = y_h)\mathrm{Var}(X_j/Y = y_h) = \sum_{Y(i)=y_h} p_i (X_j(i) - E(X_j/Y = y_h))^2$$

Prenons alors la première ligne du tableau qui précède, relatif à la variable $X =$ Sepal.Length. Les termes précédents s'interprètent facilement.

- $\sigma^2 = \mathrm{Var}(\text{Sepal.Length})$
- $\sigma^2_{set} = \mathrm{Var}(\text{Sepal.Length}/Y = \text{setosa})$
- $\sigma^2_{ver} = \mathrm{Var}(\text{Sepal.Length}/Y = \text{versicolor})$
- $\sigma^2_{vir} = \mathrm{Var}(\text{Sepal.Length}/Y = \text{virginica})$

Par suite, le premier terme du second membre de l'équation (1) n'est autre que :

$$\sigma^2_{set} P(Y = \text{setosa}) + \sigma^2_{ver} P(Y = \text{versicolor}) + \sigma^2_{vir} P(Y = \text{virginica})$$

Soit encore, du fait que les 3 espèces comportent chacune 50 individus :

$$E[\mathrm{Var}(\text{Sepal.Length}/Y)] = (\sigma^2_{set} + \sigma^2_{ver} + \sigma^2_{vir})/3$$

D'où, en divisant l'identité par Var(Sepal.Length) :

$$E[\text{Var}(\text{Sepal.Length}/Y)]/\text{Var}(\text{Sepal.Length}) = (\sigma^2_{set}/\sigma^2 + \sigma^2_{ver}/\sigma^2 + \sigma^2_{vir}/\sigma^2)/3$$

On voit ainsi que l'indicateur *mean* introduit dans le tableau qui précède n'est rien d'autre, aux carrés près, que le quotient $E[\text{Var}(X/Y)]/\text{Var}(X)$. Ce rapport est compris entre 0 et 1 et la variable X sera d'autant plus discriminante pour Y qu'il est est petit.

Cette remarque nous conduit à chercher à interpréter tous les termes de l'équation (1). Le premier, $\text{Var}(X)$, est très simple ; il s'agit de la variance – dite **variance totale** – de la v.a. X et mesure sa dispersion.

Le second est la moyenne des variances dans chaque groupe, on l'appellera plus loin **variance intra-classes**.

Le dernier représente la variance des espérances dans chaque groupe, et nous le nommerons **variance inter-classes**. Ce terme est d'autant plus fort que les espérances sont distinctes et le rapport correspondant $\text{Var}[E(X/Y)]/\text{Var}(X)$ est appelé indice de Sobol [5] Il est, comme l'indice précédent, compris entre 0 et 1, mais plus il est proche de 1, plus la variable X est discriminante (le lecteur, même fatigué, aura remarqué que la somme des deux indices vaut 1 !). Les indices de Sobol correspondants aux données des iris sont reproduites dans le tableau ci-après, et confirment les observations faites précédemment, à savoir que les deux variables concernant la taille des pétales discriminent beaucoup mieux les 3 espèces que ne le font les mesures faites sur les sépales.

variable	Indice de Sobol
Sepal.Length	0.61
Sepal.Width	0.39
Petal.Length	0.94
Petal.Width	0.93

TAB. 2 – Indices de Sobol pour les données d'iris

[5] I.M. Sobol. *Sensitivity analysis for non-linear mathematical models*, Mathematical Modelling and Computer Experiments, 1, p. 407-414 (1993).

2.2.3 Un petit tour vers l'analyse décisionnelle

Avant d'aller plus loin, on pourrait s'interroger sur la démarche qui consiste à examiner la façon dont l'espèce influence les variables mesurées pour en déduire la capacité de chaque variable à discriminer les espèces. Ce renversement des rôles est examiné en détail dans les cours de régression et de régression logistique(voir également plus bas la partie "analyse discriminante décisionnelle"). En attendant, on peut établir une règle empirique d'affectation à une espèce en fonction de l'observation d'une variable à l'aide des densités estimées plus haut. Considérons par exemple la variable Sepal.Length. On a estimé les densités de chaque espèce notées d_{set}, d_{ver}, d_{vir}. Si la valeur x pour la variable Sepal.Length a été observée, il semble assez évident que lorsque d_{set} est grande devant les deux autres densités, on aura tendance à classer la fleur considérée dans l'expèce setosa. On obtient une "probabilité" d'appartenance à l'espèce setosa en faisant simplement le rapport :

$$p_{set}(x) = d_{set}(x)/(d_{set}(x) + d_{ver}(x) + d_{vir}(x))$$

Et l'on fait de même pour les autres espèces et variables. Ce qui donne le graphique suivant.

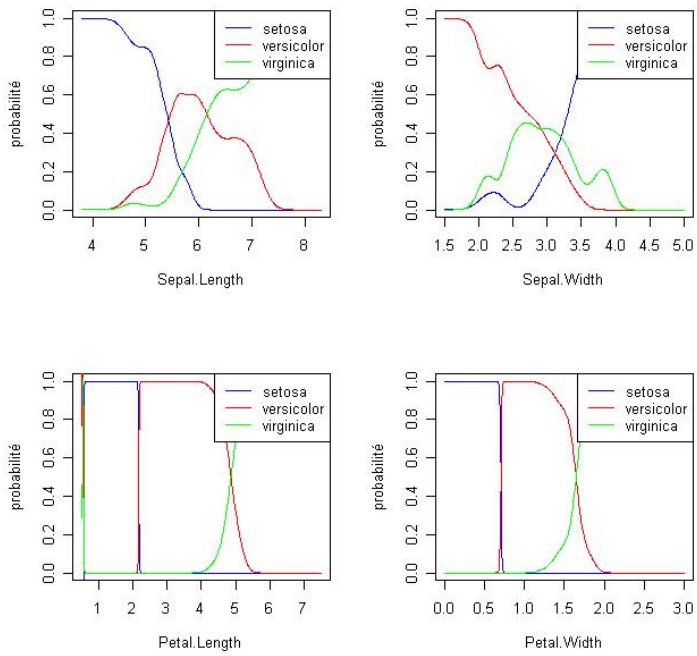

FIG. 5 – Probabilités estimées d'appartenance à une espèce

Le pouvoir discriminant des deux variables de taille des pétales sont là encore mises en évidence,

de même que la séparation nette – pour ces variables – de l'espèce setosa vis-à-vis des deux autres espèces. On peut également préciser les limites pour chaque variable qui définissent "au mieux" les 3 espèces. Par exemple, pour la variable Petal.Length, on voit que la valeur de 4.9 sépare les espèces versicolor et virginica au sens où pour cette valeur les "probabilités" d'appartenance à chaque espèce sont égales. Si nous pouvions fonder correctement les résultats que nous venons d'obtenir, et les étendre en dimension 4 en lieu et place d'une succession d'analyses mono-dimensionnelles, nous aurions répondu largement aux questions que nous nous posions, et même au-delà puisque nous serions alors en position de prédire l'espèce d'une nouvelle fleur à partir de l'observation des 4 variables, avec une probabilité d'erreur connue.

Nous allons maintenant tenter de poursuivre cette étude sur le pouvoir discriminant des variables en traitant l'une des questions qui précèdent, à savoir la réalisation d'une véritable analyse multidimensionnelle, et non – comme nous l'avons fait pour l'instant – une succession d'analyses mono-dimensionnelles. Déjà, l'analyse précédente basée sur des densités estimées ne pourra pas être utilisée en pratique pour une raison simple. Il est aisé d'estimer la densité d'une variable aléatoire à partir de 150 observations. Cela devient plus difficile si l'on veut utiliser ces données pour estimer la densité de 2 variables (la fonction à estimer devient une fonction de deux variables). Et c'est totalement illusoire d'espérer le faire avec 4 variables à la fois. Cette affirmation peut sembler étonnante. Pourtant, si l'on considère qu'il faut environ 5 points pour estimer une densité en un point (chiffre très faible, analogue aux nombre d'individus par classe dans un histogramme) et que les données relatives aux 4 variables quantitatives sont bien réparties dans l'espace, on voit que l'on peut espérer estimer la densité pour 30 valeurs différentes. Si ces valeurs sont bien réparties en dimension 4, on peut espérer $(30)^{1/4}$ points valeurs par dimension, soit un peu plus de deux points pour chaque dimension ! Autant dire que les graphiques du type de ceux de la figure 5 avec les 4 variables à la fois sont largement hors de portée en pratique ! Nous les verrons apparaître néanmoins, une fois la dimension réduite à l'aide d'autres outils.

2.2.4 L'analyse de la variance

Nous poursuivons donc l'étude à partir des indices numériques que nous avons commencé à mettre en évidence. On débute l'analyse en étendant la formule de la variance totale au cas

vectoriel. Si en effet X représente le vecteur aléatoire formé des variables X_j, $1 \leq j \leq p$ on obtient :

$$\mathrm{Var}(X) = E[\mathrm{Var}(X/Y)] + \mathrm{Var}[E(X/Y)] \tag{2}$$

formule en apparence identique à la formule (1), à la différence près que dans le cas présent, les espérances s'appliquent à des vecteurs ou des matrices et que $\mathrm{Var}(Z)$ est la matrice de covariance du vecteur aléatoire Z. Explicitons les coefficients de chacun des termes de cette formule, en utilisant la notation $g = E(X)$ et $g^h = E(X/Y = y_h)$. Comme leur nom l'indique, g représente le centre de gravité du nuage complet et g^h celui des données du groupe G_h. Comme précédemment, la variable Y était qualitative, les espérance et variance conditionnelles s'explicitent facilement. Ainsi,

$$\mathrm{Var}(X)_{j,k} = \sum_{i=1}^{n} p_i (X_j(i) - g_j)(X_k(i) - g_k)$$

$\mathrm{Var}(X)$ est la **matrice de covariance empirique totale** du nuage de points en dimension p. Elle est notée habituellement $\mathbf{T}$ (pour Totale).

$$P(Y = y_h)\mathrm{Var}(X/Y = y_h)_{j,k} = \sum_{Y(i)=y_h} p_i (X_j(i) - g_j^h)(X_k(i) - g_k^h)$$

et donc :

$$E[\mathrm{Var}(X/Y)]_{j,k} = \sum_{h=1}^{m} \sum_{Y(i)=y_h} p_i (X_j(i) - g_j^h)(X_k(i) - g_k^h)$$

soit encore en notant, pour $i \in G_h$, $p(i/Y = y_h)$ la probabilité conditionnelle $p_i/P(Y = y_h)$:

$$E[\mathrm{Var}(X/Y)]_{j,k} = \sum_{h=1}^{m} P(Y = y_h) \sum_{Y(i)=y_h} p(i/Y = y_h)(X_j(i) - g_j^h)(X_k(i) - g_k^h)$$

On voit ainsi que la quantité $E[\mathrm{Var}(X/Y)]$ s'interprète comme une moyenne, pondérée par leur probabilité, des matrices de covariances dans chaque groupe G_h. On la note $\mathbf{W}$ (pour Within) et on parle de **matrice de covariance intra-classes**. Pour le dernier terme, on rappelle que le vecteur $E[X/Y]$ prend m valeurs $g^1, \ldots, g^m$ avec comme probabilités respectives $P(Y = y_1), \ldots, P(Y = y_m)$. Son espérance vaut $g = E(X)$ et sa matrice de covariance se réduit donc à :

$$\mathrm{Var}[E(X/Y)]_{j,k} = \sum_{h=1}^{m} P(Y = y_h)(g_j^h - g_j)(g_k^h - g_k)$$

Ce terme représente la matrice de covariance du nuage des centres de gravités $g^1, \ldots, g^m$ affectés de leur probabilité. Elle est dite **matrice de covariance inter-classes** et est notée **B** pour Between. En d'autres termes, la relation (2) s'interprète selon :

$$\begin{gathered} \mathrm{T} = \mathrm{W} + \mathrm{B} \\ \text{Total} = \text{Within} + \text{Between} \\ \text{cov totale} = \text{cov intra} + \text{cov inter} \end{gathered} \tag{3}$$

2.2.5 L'analyse discriminante linéaire (LDA)

Au vu des notions exposées ci-avant, la technique d'analyse discriminante linéaire va être introduite naturellement. Nous avons vu en effet que l'indice de Sobol permettait de classer les variables quantitatives selon leur pouvoir discriminant. On peut de façon plus générale identifier parmi toutes les combinaisons linéaires de variables celle qui a l'indice de Sobol le plus important. De façon précise, soit $\beta = (\beta_1, \ldots \beta_p)'$ le vecteur des coefficients de la combinaison linéaire cherchée de sorte que celle-ci s'exprime selon :

$$\beta' X = \beta_1 X_1 + \ldots + \beta_p X_p$$

On cherche β tel que l'indice de Sobol $S(\beta) = \mathrm{Var}[E(\beta' X/Y)]/\mathrm{Var}(\beta' X)$ est maximal. Or,

$$\mathrm{Var}[E(\beta' X/Y)] = \mathrm{Var}[\beta' E(X/Y)] = \beta' \mathrm{Var}[E(X/Y)]\beta = \beta' \mathbf{B} \beta$$

$$\mathrm{Var}(\beta' X) = \beta' \mathrm{Var}(X)\beta = \beta' \mathbf{T} \beta$$

On en déduit que l'indice de Sobol vaut :

$$S(\beta) = \beta' \mathbf{B} \beta / \beta' \mathbf{T} \beta \tag{4}$$

On retrouve ici une expression bien connue en analyse en composantes principales, dite ACP. Et la recherche du vecteur β maximisant l'indice de Sobol se ramène à un calcul de plus grande valeur propre. Pour rechercher des espaces de dimension supérieure à 1 qui discriminent au mieux les groupes, nous utiliserons l'ensemble des valeurs propres, classées selon un ordre décroissant.

Propriété (réduction simultanée de deux formes quadratiques)

La matrice $\mathbf{B}$, symétrique positive de taille $n * n$, est de rang r inférieur à $m-1$. Il existe une base $(E_1, \ldots, E_n)$ orthonormale pour la forme quadratique associée à la matrice $\mathbf{T}$ et orthogonale pour la forme quadratique associée à la matrice $\mathbf{B}$. En d'autres termes, si $\mathbf{M}$ désigne la matrice de changement de base, on a :

$$\mathbf{M}'\mathbf{B}\mathbf{M} = \mathbf{\Lambda} \text{ et } \mathbf{M}'\mathbf{T}\mathbf{M} = \mathbf{Id}, \tag{5}$$

où $\mathbf{\Lambda}$ est une matrice diagonale, dont les valeurs propres sont ordonnées de manière décroissante : $1 \geq \lambda_1 \geq \lambda_2 \ldots \geq \lambda_r > 0$ et $\lambda_r + 1 = \lambda_r + 2 = \ldots = \lambda_n = 0$

La seule spécificité de cette proposition a trait aux valeurs propres. La nullité à partir du rang r+1 n'est qu'une définition du rang de la matrice $\mathbf{B}$, et que le rang r soit inférieur à $m-1$ vient du fait que la matrice $\mathbf{B}$ est la matrice de covariance des m centres de gravités des groupes. Ce nuage contenant au plus m individus est en effet inclus dans un espace affine de dimension $m-1$. Enfin, la majoration par 1 est obtenue immédiatement si on rappelle que l'indice de Sobol $S(\beta)$ est inférieur à 1, conséquence immédiate de l'identité (3).

Rappelons que les termes de valeurs et vecteurs propres sont pleinement justifiés du fait que la deuxième égalité de l'identité (5) donne $\mathbf{M}' = (\mathbf{T}\mathbf{M})^{-1} = \mathbf{M}^{-1}\mathbf{T}^{-1}$. Par suite, on a :

$$\mathbf{M}^{-1}(\mathbf{T}^{-1}\mathbf{B})\mathbf{M} = \mathbf{\Lambda}, \tag{6}$$

On en déduit que les valeurs diagonales de la matrice $\mathbf{\Lambda}$ sont les valeurs propres de la matrice $\mathbf{T}^{-1}\mathbf{B}$ et que les vecteurs propres sont ceux qui forment la matrice $\mathbf{M}$. Ces vecteurs propres $E_1, \ldots, E_n$ sont appelés **variables discriminantes** ou **variables canoniques**[6]. Si l'on reprend l'interprétation probabiliste des quantités introduites, la variable aléatoire E_1 est la variable aléatoire E, combinaison linéaire des variables aléatoires $X_1, \ldots, X_p$, dont la variabilité est réduite au minimum par la variable qualitative Y qui désigne le groupe. Elle résout donc le problème de maximisation de l'indice de Sobol donné par (4). De la même manière, le vecteur aléatoire (E_1, E_2) est le vecteur aléatoire E de dimension 2, combinaison linéaire des variables aléatoires $X_1, \ldots, X_p$, dont la variabilité est réduite au minimum par la variable qualitative Y qui désigne le groupe. Et tout ceci reste vrai quelle que soit la dimension considérée (1, 2,...ou q). La démarche correspondante est intitulée analyse discriminante linéaire ou **Linear Discriminant Analysis (LDA)**.

[6]Cette terminologie fait référence au lien entre analyse discriminante linéaire et analyse canonique.

On notera au passage la similitude entre ce qui précède et l'ACP. Celle-ci dépasse la simple analogie puisque l'on peut énoncer le résultat qui suit.

Analyse discriminante et ACP

L'algorithme de la LDA est le même que celui de l'ACP du nuage des m centres de gravité des groupes, affectés de leur poids respectif, où l'espace des individus est muni de la métrique dite de Mahalanobis $\mathbf{T}^{-1}$.

(On définit dans certains ouvrages, par exemple [Saporta], la métrique de Mahalanobis comme étant celle associée à la matrice $\mathbf{W}^{-1}$.)

Ce résultat est une simple conséquence de l'identité (6).

Pour en terminer avec ces questions générales, signalons que l'analyse discriminante linéaire est quelquefois [7] développée en remplaçant la matrice $\mathbf{T}$ par la matrice $\mathbf{W}$. C'est le cas notamment lorsque l'on effectue une analyse décisionnelle car les rapports du type de ceux donnés dans la formule (4) s'interprètent alors en termes de statistique de Fisher. Donnons très rapidement le lien, très simple, entre ces deux analyses.

On réécrit la formule (6) sous la forme :

$$\mathbf{BM} = \mathbf{TM\Lambda}$$

Puis on incorpore dans cette identité la formule de la variance totale (3) pour obtenir simplement :

$$\mathbf{M}^{-1}\mathbf{W}^{-1}\mathbf{BM}(\mathbf{Id} - \mathbf{\Lambda}) = \mathbf{\Lambda}$$

Soit encore :

$$\mathbf{M}^{-1}(\mathbf{W}^{-1}\mathbf{B})\mathbf{M} = \mathbf{\Lambda}(\mathbf{Id} - \mathbf{\Lambda})^{-1}$$

Mais la matrice $\mathbf{\Lambda}(\mathbf{Id} - \mathbf{\Lambda})^{-1}$ est diagonale, de terme général $\lambda_i/(1 - \lambda_i)$.

On déduit de tout cela que les vecteurs propres $E_1, \ldots, E_n$ de la matrice $\mathbf{T}^{-1}\mathbf{B}$ restent vecteurs propres de la matrice $\mathbf{W}^{-1}\mathbf{B}$ de valeurs propres associées données par l'identité :

$$\mu_i = \lambda_i/(1 - \lambda_i)$$

2.2.6 Retour sur l'exemple des iris

Nous appliquons donc la méthodologie LDA sur l'exemple des iris, c'est-'a-dire que nous cherchons la meilleure combinaison linéaire $\beta' X = \beta_1 X_1 + \ldots + \beta_p X_p$ au sens de la LDA. On

[7] La fonction lda du package MASS sous R effectue ce choix

projette sur un sous-espace de dimension 2 (car le nombre de groupes est de 3). Donnons à nouveau la mesure de dispersion associée à ces nouvelles variables, ainsi que les indices de Sobol correspondants :

variable	σ	σ_{set}	σ_{ver}	σ_{vir}	σ_{set}/σ	σ_{ver}/σ	σ_{vir}/σ	mean
LD1	5.72	0.85	1.04	1.10	0.15	0.18	0.19	0.17
LD2	1.13	0.91	0.87	1.18	0.81	0.78	1.05	0.88

Tab. 3 – Pouvoir discriminant de chaque variable discriminante

variable	Indice de Sobol
LD1	0.97
LD2	0.21

Tab. 4 – Indices de Sobol pour les données d'iris projetées

La matrice de passage de des variables initiales au variables discriminantes est donnée par

	LD1	LD2
Sepal.Length	0.83	0.02
Sepal.Width	1.53	2.16
Petal.Length	-2.20	-0.93
Petal.Width	-2.81	2.84

Ainsi la projection des individus iris sur le sous-espace des variables discriminantes (attention, le centre de gravité dans cet espace est 0) est représenté dans la figure 6.
Ce graphique n'est pas du tout inattendu, au vu des indices de Sobol, on voit effectivement que le premier axe a un très bon pouvoir discriminant, alors que le deuxième discrimine légèrement les espèces versicolor et virginica. En ce sens, on peut donner un autre indicateur, représentant le pouvoir discriminant relatif d'un axe, c'est-à-dire rapporté aux variables (contrairement à l'indice de Sobol qui est un indice donnant le pouvoir discriminant absolu), qui est la contribution relative des valeurs propres μ_1 et μ_2 : $(0.99, 0.01)$

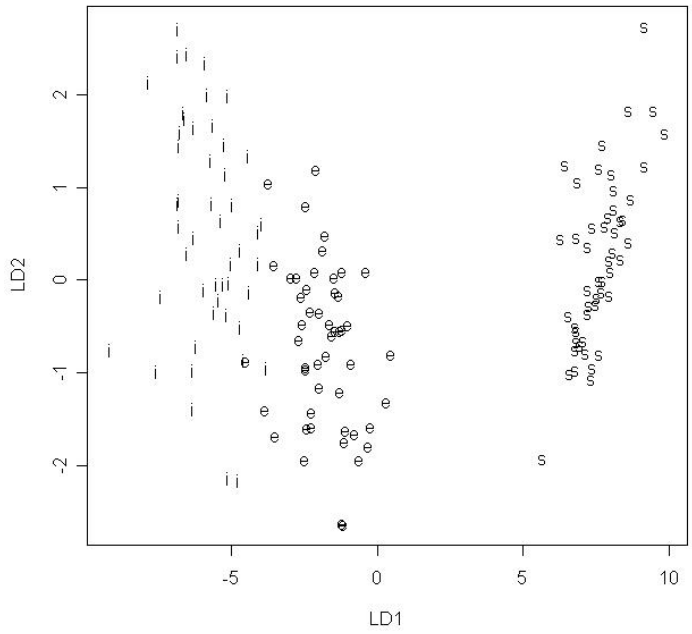

FIG. 6 – Projection des iris sur l'espace des variables discriminantes

Transformation des variables ou augmentation de l'espace des variables (1)

Comme en régression, on peut imaginer de transformer les variables initiales ou de les combiner et recommencer la procédure, i.e. identifier parmi les combinaisons linéaires de ces nouvelles variables celle dont l'indice de Sobol est le plus important. L'approche la plus commune consiste à rajouter tous les produits des variables deux à deux (distinctes ou pas).

Notons par X_{new} le vecteur

$$X_{new} = [X_1, \ldots, X_p, X_1X_2, X_2X_3, \ldots, X_{p-1}X_p, X_1^2, \ldots, X_p^2]'.$$

Ce vecteur appartient à ce que nous appellerons **espace quadratique** associé aux variables initiales. Appliquons à nouveau la méthodologie LDA : on obtient les résultats des tableaux 5 et 6.

La contribution relative des valeurs propres μ_1 et μ_2 est $(0.96, 0.04)$.

On remarque que l'indice de Sobol correspondant à la première variable discriminante augmente un peu ; cette augmentation est un phénomène général. Par contre, l'indice de Sobol pour la deuxième variable discriminante est beaucoup plus élévé qu'initialement, passant de 0.21 à 0.75.

Transformation des données (2)

variable	σ	σ_{set}	σ_{ver}	σ_{vir}	σ_{set}/σ	σ_{ver}/σ	σ_{vir}/σ	mean
LD1NEW	8.58	1.27	1.07	0.49	0.15	0.12	0.06	0.11
LD2NEW	1.99	0.58	0.78	1.43	0.29	0.39	0.72	0.47

TAB. 5 – Pouvoir discriminant de chaque variable discriminante si la LDA est faite sur l'espace quadratique

variable	Indice de Sobol
LD1NEW	0.99
LD2NEW	0.75

TAB. 6 – Indices de Sobol pour les données d'iris projetées si la LDA est faite sur l'espace quadratique

On verra par la suite (en régression logistique, car pour le moment nous n'avons pas encore les outils nécessaires) que la variable "qui compte" est le produit Petal.Length * Petal.Width. D'ailleurs, en projetant sur cet axe on trouve un indice de Sobol égal à 0.99, et la projection des iris est donnée dans la figure 8.

2.3 Analyse discriminante décisionnelle

L'analyse discriminante décisionnelle pose le problème suivant : étant donné un nouvel individu sur lequel on a observé les p variables X_j mais pas la variable qualitative Y (l'appartenance à un groupe), comment décider de la modalité y_h de Y, c'est-à-dire du groupe auquel appartient cet individu.
Pour cela nous allons définir des règles de décision et nous donner les moyens de les évaluer sur un seul individu.

2.3.1 Règle de décision issue de l'analyse discriminante descriptive

On affectera un individu x à la modalité y_h en minimisant sa distance (dans la métrique de Mahalanobis $\mathbf{W}^{-1}$) aux centres de gravité de chaque classe G_h, i.e.

$$\|x - g_h\|^2_{\mathbf{W}^{-1}} = (x - g_h)'\mathbf{W}^{-1}(x - g_h)$$

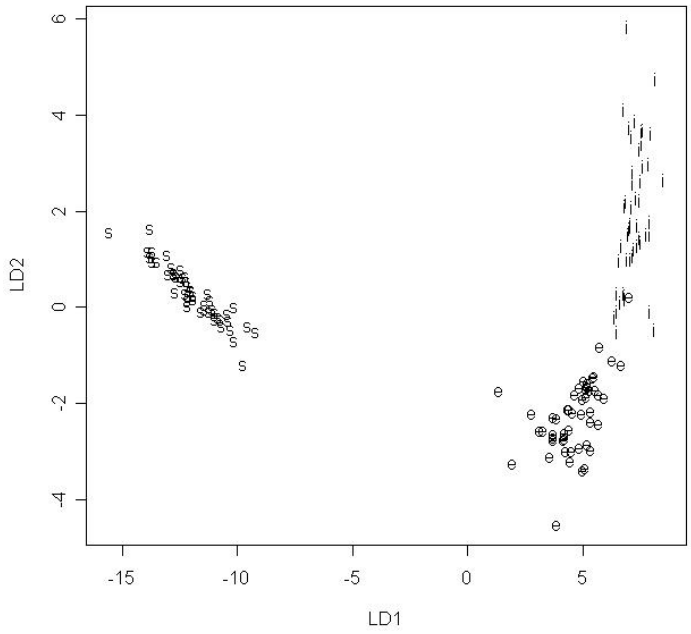

FIG. 7 – Projection des iris sur l'espace des variables discriminantes (avec polynôme de degré 2)

ce qui revient à chercher la modalité y_h qui maximise

$$l_h(x) g_h' \mathbf{W}^{-1} x - \frac{1}{2} g_h' \mathbf{W}^{-1} g_h. \tag{7}$$

Il s'agit d'une règle linéaire ! Ainsi, les groupes estimés par cette analyse sont limités par les hyperplans $l_h = l_k$ pour $h \neq k$.

Une illustration pour le cas bidimensionnel est donnée dans la figure 9. Nous avons considéré trois groupes ; dans chaque groupe les observations suivent toutes des lois normales, de même matrice de variance-covariance $\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$, et de moyennes $g_1 = (2, 2)'$, $g_2 = (-2, 2)'$, $g_3 = (0, -2)'$. Les frontières entre les groupes sont des droites.

2.3.2 Lien avec la théorie de la décision statistique

La théorie de la décision statistique nous dit qu'on a besoin de connaître les probabilités a posteriori $P(Y = y_h|X)$ pour faire une classification optimale.

Soit $d_h(x)$ la densité conditionnelle à la classe $Y = y_h$ de X et π_h la probabilité a priori de la

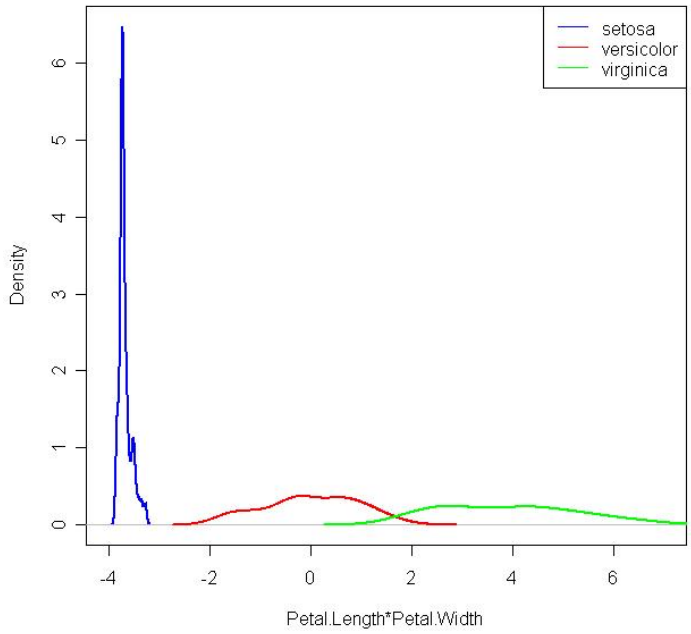

FIG. 8 – Densité estimée de la projection des iris sur Petal.Length*Petal.Width (selon l'appartenance à une espèce)

classe h [8], avec $\sum_{h=1}^{m} \pi_h = 1$. En appliquant le théorème de Bayes on a

$$P(Y = y_h | X = x) = \frac{d_h(x)\pi_h}{\sum_{l=1}^{m} d_l(x)\pi_l}$$

Ce résultat signifie que la connaissance de la densité conditionnelle $d_h(x)$ est presque équivalente à la probabilité a posteriori $P(Y = y_h|X)$. Et cela justifie les calculs que nous avons effectués en 2.2.3.

Il y a différentes techniques pour modéliser (estimer) ces densités conditionnelles, dont l'une des plus flexibles est l'estimation de densité non paramétrique, mais qui est, comme vous l'avez vu dans la section *Un petit tour vers l'analyse décisionnelle* hors de portée en pratique.

Une autre technique, qui est celle développée ici, est basée (comme vous pouvez vous y attendre) sur les densités Gaussiennes.

On suppose que le modèle pour la densité de chaque classe est Gaussien multivarié, i.e.

$$d_h(x) = \frac{1}{(2\pi)^{p/2} |\sum_h|^{1/2}} \exp\left(-\frac{1}{2}(x - g_h)' \Sigma_h^{-1} (x - g_h)\right). \quad (8)$$

Plusieurs cas peuvent alors être considérés.

[8] On rappelle que pour notre exemple des iris, $m = 3$ et $\pi_h = 1/3$

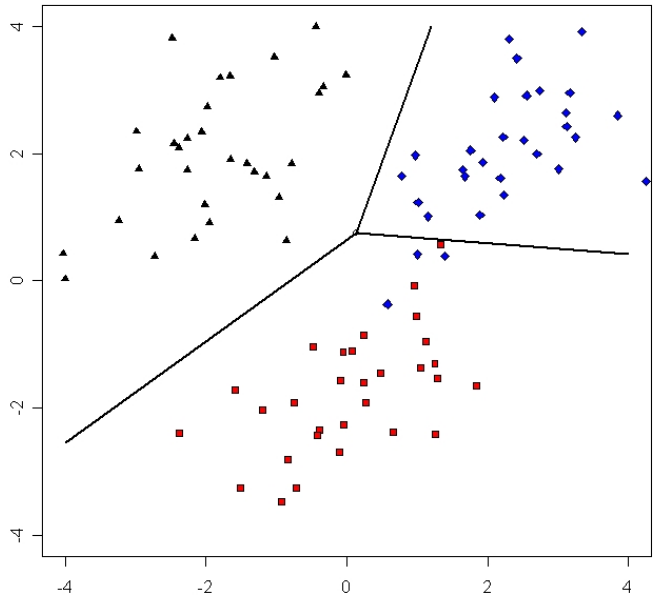

FIG. 9 – Frontières théoriques pour une LDA (trois groupes de données simulées)

2.3.3 Analyse discriminante linéaire

On suppose que l'on se trouve dans le cas de l'homoscédasticité, i.e. le cas où toutes les classes ont des matrices de covariances identiques $\Sigma_h = \Sigma$, mais leurs espérances restent bien sûr distinctes. Si on compare deux classes h et k, on obtient

$$\log \frac{P(Y = y_h|X = x)}{P(Y = y_k|X = x)} = \log \frac{d_h(x)}{d_k(x)} + \log \frac{\pi_h}{\pi_k} = \log \frac{\pi_h}{\pi_k} - \frac{1}{2}(g_h + g_k)'\Sigma^{-1}(g_h - g_k) + x'\Sigma^{-1}(g_h - g_k)$$

c'est-à-dire une équation linéaire en x. Ceci implique que la frontière de décision entre les deux classes (l'ensemble où les deux probabilités a posteriori sont égales) est linéaire en x (un hyperplan en p dimensions), et que par conséquent toutes les frontières de décision sont linéaires, comme nous avons pu le voir dans un cas particulier en 2.3.2.

De façon équivalente, on définit la fonction discriminante linéaire pour chaque classe par

$$\delta_h(x) = x'\Sigma^{-1}g_h - \frac{1}{2}g_h'\Sigma^{-1}g_h + \log \pi_h = l_h(x) + \log \pi_h. \tag{9}$$

On classe x dans la classe ayant la plus grande valeur pour sa fonction discriminante ($G(x) = \arg\max_k \delta_k(x)$).

Attention, dans la pratique nous ne connaissons pas les paramètres des différentes lois normales ; on les estime à partir des données :

- $\hat{\pi}_h = n_h/n$, n_h étant le nombre d'individus dans la classe h
- $\hat{g}_h = \sum_{y(i)=y_h} x(i)/n_h$
- $\hat{\Sigma} = \sum_{h=1}^{m} \sum_{y(i)=y_h} (x(i) - \hat{g}_h)(x(i) - \hat{g}_h)'/(n-m)$

Ainsi, pour l'exemple des iris on a

- $\hat{\pi}_h = 1/3$
- $\hat{g}_{set} = [5 \;\; 3.43 \;\; 1.46 \;\; 0.25]'$

  $\hat{g}_{ver} = [5.94 \;\; 2.78 \;\; 4.26 \;\; 1.33]'$

  $\hat{g}_{vir} = [6.59 \;\; 2.97 \;\; 5.55 \;\; 2.03]'$
- $\hat{\Sigma}$ peut être calculée, mais au vu des premiers graphiques (boxplot ou densités mono dimensionnelles estimées) l'hypothèse de homoscédasticité ne semble pas satisfaite.

2.3.4 Analyse discriminante quadratique (QDA)

Dans le cas où l'hypothèse d'homoscédasticité n'est pas vérifiée, on est dans le cas d'hétéroscédasticité, i.e. toutes les classes ont des matrices de covariances distinctes.
On obtient alors des fonctions discriminantes quadratiques pour chaque classe

$$\delta_h(x) = -\frac{1}{2}\log|\Sigma_h| - \frac{1}{2}(x - g_h)'\Sigma_h^{-1}(x - g_h) + \log \pi_h. \tag{10}$$

et la frontière de décision entre chaque paire de classes h et k est décrite par l'équation quadratique $\{x|\delta_h(x) = \delta_k(x)\}$. On est dans le cadre de l'**analyse discriminante quadratique (QDA)**.

Remarque : Dans la pratique il y a deux manières de calculer ces frontières quadratiques : soit directement par QDA, soit par LDA dans l'espace élargi formé par toutes les variables, leurs carrés et leurs interactions, que nous avons appelé espace quadratique. Les résultats obtenus par les deux méthodes sont très similaires [9], mais dans le cas de la QDA il ne faut pas oublier qu'il faut estimer les matrices de covariance pour chaque classe.

2.4 Validation

Afin de valider une démarche statistique comme celle-ci, on partage les données en deux sous-ensembles, dits **ensemble de calibration** et **ensemble de validation**, en choisissant un

[9] voir [Hastie]

échantillon dans chaque classe. Sur l'ensemble de calibration on applique la LDA (ou bien une autre méthode) et on utilise ces résultats sur l'ensemble de validation (ensemble qui n'a donc pas été utilisé dans les calculs jusqu'à maintenant). Ainsi on calcule pour l'ensemble de validation les probabilités a posteriori d'appartenance à un groupe, on affecte chaque individu au groupe pour lequel la probabilité a posteriori est la plus grande et enfin on compare cette affectation avec l'appartenance réelle de chaque individu à son groupe de départ.
Illustration sur le cas de la LDA avec les données non-transformées ; les axes ont été calculés avec les 75 données de l'ensemble de calibration (25 pour chaque classe) et l'affectation se fait sur les 75 individus restants.

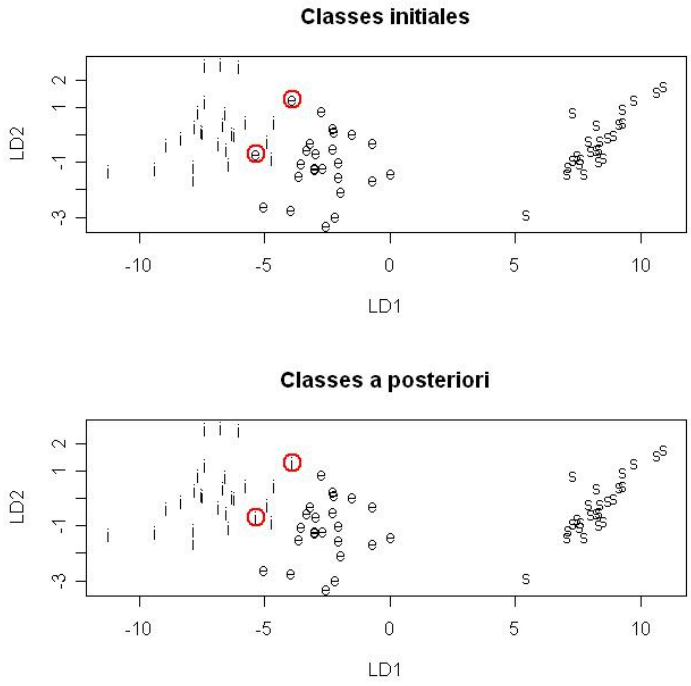

FIG. 10 – Affectation au groupe d'origine des individus de l'ensemble de validation (en haut) ; affectation selon la probabilité a posteriori des individus de l'ensemble de validation (en bas) ; sur les deux graphiques, les deux individus "mal classés a posteriori" sont mis en évidence.

3 Références

3.1 Bibliographie

[Besse] Ph. Besse, *Data mining, Modélisation Statistique et Apprentissage*, Cours Université Paul Sabatier (2005)

[Hastie] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning*, Springer (2001).

[Romeder] J.M. Romeder, *Méthodes et programmes d'analyse discriminante*, Dunod (1973).

[Saporta] G. Saporta, *Probabilités analyse de données et statistique*, Technip (1990).

[Venables] W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S*, Springer (2002).

3.2 Ressources informatiques

Packages dans l'environnement R :

- MASS : Package lié au livre de Venables et Ripley en principe associé à toute installation de R sur un poste de travail. Package qui contient en particulier les fonctionnalités élémentaires pour l'analyse discriminante, pas toujours très bien documentées. On y trouve notamment les fonctions lda, qda, mda, predict.lda, predict.qda, plot.lda, plot.qda.
- ade4 : Package très complet d'analyse de données multidimensionnelles réalisé par le laboratoire de Biométrie et Biologie Evolutive de l'Université Lyon 1. Il nécessite toutefois un investissement important (le document de description des diverses fonctions du package dépasse 200 pages !).