Aperçu sur l'Analyse Discriminante

1 Présentation de la méthode

1.1 Position du problème

On dispose de n observations sur lesquelles on a relevé :

- les valeurs d'une variable catégorielle comportant quelques modalités (2, 3, ...) : c'est le groupe ou diagnostic.
- les valeurs de p variables numériques : X1, X2, ..., Xp : ce sont les prédicteurs.

On se pose des questions telles que :

- dans quelle mesure la valeur de Y est-elle liée aux valeurs de X1, X2, ..., Xp ?
- Etant donné d'autres observations, pour lesquelles X1, X2, ..., Xp sont connues, mais Y ne l'est pas, est-il possible de prévoir Y (le groupe), et avec quel degré de certitude ?

Exemples de situations où une telle méthode peut être intéressante :

Exemple 1. On étudie les différentes espèces de poissons peuplant un lac, mais la détermination exacte de l'espèce suppose que l'on sacrifie l'animal. Peut-on se contenter de relever différents paramètres concernant les poissons prélevés, et déduire l'espèce à partir de ces paramètres avec un degré de certitude raisonnable ?

Exemple 2. Pour déterminer le type d'utilisation de parcelles agricoles, on peut évidemment faire des relevés sur le terrain. Mais pourrait-on utiliser les informations données par des images satellites ?

La méthode est également utilisée sans que l'on ait un objectif de prédiction; on souhaite seulement déterminer les prédicteurs les plus liés au groupe d'appartenance.

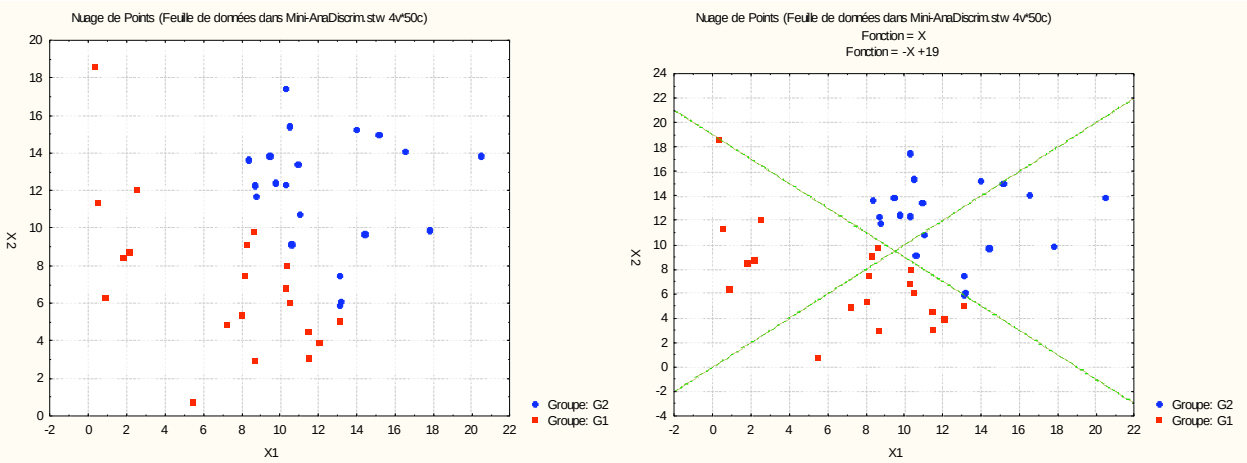
1.2 Précautions et limites de la méthode

Comme dans le cas de la régression linéaire, l'emploi de cette méthode suppose que les variables prédictrices possèdent des propriétés de régularité satisfaisantes : distribution normale (voire multinormale) des variables Xi dans les différentes populations.
Par ailleurs (comme pour la régression linéaire), l'analyse discriminante peut conduire à des résultats incorrects si les variables Xi sont trop fortement corrélées entre elles.

2 Analyse discriminante sur un mini-exemple

2.1 Présentation de l'exemple

On a relevé les valeurs de deux variables X1 et X2 sur 40 individus statistiques répartis en deux groupes. Le nuage de points représentant ces observations est le suivant :



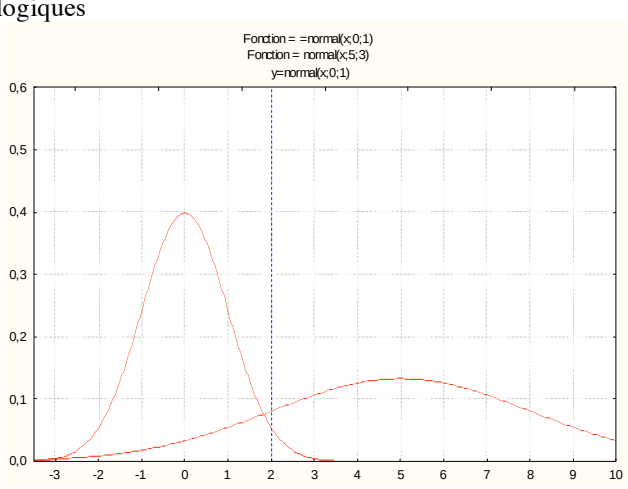
Prise isolément, aucune des deux variables X1 et X2 ne permet de différencier les deux groupes G1 et G2. Cependant, on voit bien que les deux groupes occupent des régions du plan bien spécifiques.

On voit cependant intuitivement que notre problème pourrait être résolu en considérant une variable abstraite, combinaison linéaire de X1 et X2 (approximativement X1 + X2) définie de façon que :

- la variance (dispersion) intra-groupes soit la plus petite possible
- la variance inter-groupes (variance calculée à partir des points moyens pondérés des groupes) soit la plus grande possible.

Ainsi, sur notre exemple, la droite d'équation Y= - X + 19 semble séparer correctement les deux groupes et il semblerait que c'est en projetant les points sur la droite Y=X que l'on obtiendra une dispersion minimale dans les groupes et maximale entre les groupes.

Remarque : Dans notre exemple, les deux groupes présentent à peu près la même dispersion de valeurs. Cependant, dans d'autres situations, l'un des groupes peut être nettement plus dispersé que l'autre.
Considérons la situation suivante, où l'on a représenté la distribution des valeurs issues des deux groupes sur le "facteur discriminant". On souhaite par exemple, affecter la valeur x=2 à l'un des deux groupes. Pour la distance "habituelle" (euclidienne), cette valeur est plus près du centre du premier groupe (valeur 0) que du centre du second groupe (valeur 5). Cependant, x=2 a plus de chances d'être une observation provenant du premier groupe qu'une observation provenant du second groupe.

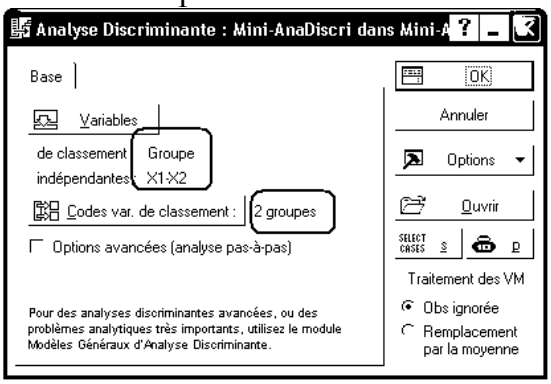


Pour résoudre ce problème, on introduit une distance particulière : la **distance de Mahalanobis** pour évaluer la distance entre un point et le centre d'un groupe.

2.2 Traitement de l'exemple précédent avec Statistica

Ouvrez le fichier Mini-AnaDiscrim.stw

Faites une analyse discriminante (menu Statistiques - Techniques exploratoires multivariées - Analyse discriminante) en indiquant les codes G2 et G1 comme codes pour la variable catégorielle "Groupe", X1 et X2 comme variables indépendantes.



L'onglet Avancé nous donne accès aux boutons suivants :

Synthèse (variables dans le modèle) :

N=40	Synthèse de l'Analyse Discriminante (Mini-AnaDiscri dans Mini-AnaDiscrim.stw) Vars dans le modèle : 2; Classmt : Groupe (2 grps) Lambda Wilk : ,38021 F approx. (2,37)=30,158 p< ,0000					
	Wilk (Lambda)	Partiel (Lambda)	F d'exc. (1,37)	niveau p	Tolér.	1-Tolér. (R?)
X1	0,676419	0,562090	28,82580	0,000004	0,838237	0,161763
X2	0,668372	0,568857	28,04266	0,000006	0,838237	0,161763

Cette feuille donne les résultats d'un test

Distances inter-groupes, qui fournit trois feuilles de résultats :

Distance entre les centroïdes des deux groupes

	Dist. de Mahalanobis au Carré (Mini-AnaDiscri dans Mini-AnaDiscrim.stw)	
Groupe	G2	G1
G2	0,000000	6,194525
G1	6,194525	0,000000

Un test statistique concernant la séparation des deux groupes

	Valeurs F ; dl 2,37 (Mini-AnaDiscri dans Mini-AnaDiscrim.stw)	
Groupe	G2	G1
G2		30,15756
G1	30,15756	

	niveau p (Mini-AnaDiscri dans Mini-AnaDiscrim.stw)	
Groupe	G2	G1
G2		1,6998E-8
G1	1,6998E-8	

Réaliser une analyse canonique, qui donne accès à un autre ensemble de résultats

Notamment, le bouton "Coefficients des variables canoniques" produit deux feuilles de résultats, dont la définition de la première variable canonique.

	Coefficients bruts des Variables Canoniques
Variable	Comp_1
X1	-0,241220
X2	-0,254916
Constte	4,776475
V.Propre	1,630138
Prop.Cum	1,000000

Ici, la première variable est C1 = -0,24122 X1 - 0,25916 X2 + 4,776475.

L'onglet Scores canoniques, puis le bouton "Scores canoniques de chaque observation" donnent la valeur de la variable canonique sur chaque observation. On voit ainsi que, sauf exception, les observations classées dans le groupe G2 ont des scores négatifs pendant que celles classées dans le groupe G1 ont des scores positifs.

En cliquant sur le bouton Annuler, on revient aux résultats de l'analyse discriminante proprement dite. L'onglet Classification donne accès aux résultats suivants :

Fonctions de classification

	Fonctions de classif. ; classement: Groupe	
Variable	G2 p=,50000	G1 p=,50000
X1	1,4380	0,83765
X2	1,5504	0,91594
Constte	-18,8231	-6,93504

La fonction discriminante linéaire du groupe G2 est :
F2 = 1,4380 X1 + 1,5504 X2 - 18,8231
Celle du groupe G1 est :
F1 = 0,83765 X1 + 0,91594 X2 - 6,93504

La méthode classe un élément dans le groupe G1 si F1 > F2 et dans G2 dans le cas contraire.

Matrice de classification

Ce tableau est encore appelé *Matrice de confusion*. Il croise la classification observée avec la classification calculée par la méthode.

	Matrice de Classification Lignes : classifications observées Colonnes : classifications prévues		
Groupe	% Correct	G2 p=,50000	G1 p=,50000
G2	90,00000	18	2
G1	95,00000	1	19
Total	92,50000	19	21

Classification d'observations

	Classification d'observations Classif. incorrectes indiquées par *		
Observation	Classif. Observée	1 p=,50000	2 p=,50000
1	G2	G2	G1
2	G1	G1	G2
3	G1	G1	G2
4	G2	G2	G1
5	G2	G2	G1
* 6	G2	G1	G2
7	G2	G2	G1

Ce tableau donne pour chaque observation, le groupe le plus probable (selon le calcul), ainsi que le second candidat. Il indique également le classement calculé des valeurs qui n'étaient pas classées a priori :

	Classification d'observations Classif. incorrectes indiquées par *		
Observation	Classif. Observée	1 p=,50000	2 p=,50000
40	G2	G2	G1
41	---	G2	G1
42	---	G1	G2
43	---	G1	G2
44	---	G2	G1
45	---	G1	G2
46	---	G1	G2

Distances de Mahalanobis au carré

Observation	Dist. Mahalanobis Carrées aux Centroïdes de Groupe Classif. incorrectes indiquées par *		
	Classif. Observée	G2 p=,50000	G1 p=,50000
1	G2	1,02516	3,21197
2	G1	9,43294	0,46701
3	G1	4,30475	0,81343
4	G2	1,71089	3,08567
5	G2	0,20940	6,53698
* 6	G2	2,95094	2,71562
7	G2	0,48679	4,17369

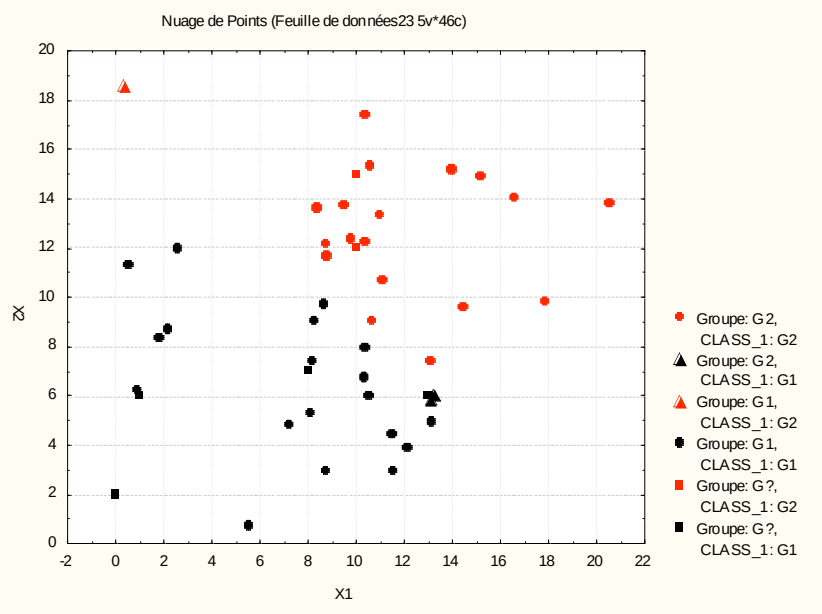
Probabilités a posteriori

Observation	Probabilités a posteriori Classif. incorrectes indiquées par *		
	Classif. Observée	G2 p=,50000	G1 p=,50000
1	G2	0,749022	0,250978
2	G1	0,011174	0,988826
3	G1	0,148595	0,851405
4	G2	0,665386	0,334614
5	G2	0,959449	0,040551
* 6	G2	0,470618	0,529382
7	G2	0,863356	0,136644

En fait pour chaque observation la méthode calcule une probabilité d'appartenance à chacun des deux groupes et affecte l'observation au groupe le plus probable.

Enregistrer les scores

Ce bouton permet de générer une feuille de données avec la classification produite par la méthode, et éventuellement, les variables et la classification initialement observées. Cette feuille de données peut être utilisée pour produire un nuage de points tel que le suivant :



Dans ce graphique, les points bien classés sont représentés par des cercles, les points mal classés par des triangles, et les points supplémentaires par des carrés. La couleur (rouge ou noir) correspond au groupe calculé.

3 Les iris de Fisher

Ouvrez le classeur Iris.stw. Il s'agit d'un exemple, initialement proposé par Fisher, et utilisé comme données de référence par la plupart des logiciels de statistiques.

On a noté, pour 150 iris, l'espèce (setosa, versicolor, virginica) et 4 variables numériques : la longueur et la largeur des sépales, la longueur et la largeur des pétales. Pour chaque espèce, on dispose de 50 observations. Les 25 premières observations de chaque espèce vont contituer l'ensemble d'apprentissage, tandis que les 25 observations restantes seront classifiées à l'aide des résultats de l'analyse discriminante. La classification ainsi obtenue pourra ainsi être comparée aux données réelles.

Procédez de même à une analyse discriminante sur ces données. Comme nous avons ici 4 variables numériques et 3 groupes, nous aurons deux facteurs discriminants, et Statistica nous permet de construire un graphique représentant les observations selon les valeurs de leurs scores canoniques :