5. Analyse discriminante

L'analyse discriminante étudie des données provenant de **groupes connus à priori**. Elle vise deux buts principaux:

i. **Description:** Parmi les groupes connus, quelles sont les principales différences que l'on peut déterminer à l'aide des variables mesurées?

ii. **Classement:** Peut- on déterminer le groupe d'appartenance d'une nouvelle observation uniquement à partir des variables mesurées?

La figure 1 (tiré de Davis, 1973, p. 444) illustre une AD pour le cas de deux groupes, dans un espace à 2 variables. La discrimination entre les deux groupes est moyenne aussi bien sur la $1^{ère}$ que sur la 2^{e} variable. On note un chevauchement non-négligeable des groupes. Par contre, si on projetait les observations sur la droite égale à $0.5x_1+0.5x_2$, on aurait une discrimination parfaite entre les deux groupes. C'est ce que cherche à faire l'AD dans le contexte plus général où l'on dispose de plusieurs variables et plusieurs groupes.

Les domaines d'application de l'analyse discriminante sont nombreux en géologie: définition d'indices de prospection géochimique, analyse d'images, caractérisation géochimique de types de roches, etc. L'analyse discriminante se rattache au champ plus vaste de la reconnaissance des formes. Par ses objectifs, elle s'apparente également aux réseaux neuronaux, sujet très à la mode en recherche informatique.

5.1 Notation et formulation du problème

Soit:

- n: nombre total d'observations.
- p: nombre de variables mesurées.
- k: nombre de groupes.
- n_k: nombre d'observations dans le groupe k.

- T: matrice de variabilité totale.
- T^*: matrice de covariances totale: T/(n-1)
- E: matrice de variabilité entre les groupes.
- E^*: matrice de covariances entre les groupes: E/(k-1)
- D: matrice de variabilité dans les groupes.
- D^*: matrice de covariances dans les groupes : D/(n-k).

Remarque: T, E et D sont des matrices p x p.
T et D sont habituellement de rang p.
E est de rang k-1.

- X: matrice n x p des observations où les observations sont placées un groupe à la suite de l'autre.
- C: matrice n x p des moyennes des p variables dans les k groupes répétées n_k fois.
- y_i: vecteur px1 des moyennes des p variables pour le groupe i.

On a donc:

$$X = \begin{bmatrix} x_{11}^1 & x_{12}^1 & \cdots & x_{1p}^1 \\ x_{21}^1 & & \cdots & x_{2p}^1 \\ \vdots & & \cdots & \vdots \\ x_{n_1 1}^1 & x_{n_1 1}^1 & \cdots & x_{n_1 p}^1 \\ \hline x_{11}^2 & x_{12}^2 & \cdots & x_{1p}^2 \\ x_{21}^2 & & \cdots & x_{2p}^2 \\ \vdots & & \cdots & \vdots \\ x_{n_2 1}^2 & x_{n_2 2}^2 & \cdots & x_{n_2 p}^2 \\ \hline \vdots & & & \\ \hline x_{11}^k & x_{12}^k & \cdots & x_{1p}^k \\ x_{21}^k & & \cdots & x_{2p}^k \\ \vdots & & \cdots & \vdots \\ x_{n_k 1}^k & x_{n_k 2}^k & \cdots & x_{n_k p}^k \end{bmatrix} \qquad C = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{11} & & \cdots & y_{1p} \\ \vdots & & \cdots & \vdots \\ y_{11} & y_{12} & \cdots & y_{1p} \\ \hline y_{21} & y_{22} & \cdots & y_{2p} \\ y_{21} & & \cdots & y_{2p} \\ \vdots & & \cdots & \vdots \\ y_{22} & y_{22} & \cdots & y_{2p} \\ \hline \vdots & & & \\ \hline y_{k1} & y_{k2} & \cdots & y_{kp} \\ y_{k1} & & \cdots & y_{kp} \\ \vdots & & \cdots & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kp} \end{bmatrix}$$

5.2 Aspect descriptif

Soit un vecteur u_1. On choisira u_1 de telle sorte que les projections des moyennes des groupes sur u_1 soient le plus espacées possible et que, simultanément, les projections des observations d'un même groupe soient le plus rapprochées possible de la projection de la moyenne du groupe. Bref, sur ce vecteur u_1 on cherche à observer des groupes compacts et distants les uns des autres.

La matrice X centrée par rapport aux moyennes calculées avec toutes les observations (sans tenir compte du groupe) est donnée par:

$$X_c = X - 11'X/n$$

De même, la matrice C centrée (i.e. la matrice contenant les moyennes de chaque groupe centrées par rapport à la moyenne globale) s'écrit:

$$C_c = C - 11'X/n$$

On pourrait également centrer chaque observation de la matrice X par rapport à la moyenne du groupe correspondant:

$$X_g = X - C$$

Bien sûr, on a:

$$X_c = C_c + X_g$$

La matrice de variabilité (totale) s'écrit alors:

$$T = X_c'X_c$$

$$T = C_c'C_c + X_g'X_g \quad \text{car } X_g'C_c = 0$$

$$T = E + D$$

Le premier membre de droite représente la matrice de variabilité entre les centres des groupes (E pour "Entre"). Le second membre représente la matrice de variabilité à l'intérieur des groupes (D pour "Dans").

Les groupes seront d'autant plus faciles à discriminer (à séparer) que E sera grand par rapport à D (où à T). En effet, si E est grand, ceci signifie que les centres des groupes sont éloignés. Si D est petit, ceci signifie que les observations d'un même groupe sont proches de leur centre. Si on a simultanément E grand et D petit alors les groupes sont éloignés les uns des autres et compacts, la situation idéale.

5.2.1 Recherche du vecteur séparant le mieux possible les groupes.

Soit un vecteur u sur lequel seront effectuées les projections des observations. Effectuons les "projections" sur u:

$$X_c u$$

La variabilité de ces projections est donnée par:

$$u'Tu$$

On a:

$$u'Tu = u'Du + u'Eu$$

Le vecteur u recherché est le vecteur qui maximise le rapport:

$$\frac{u'Eu}{u'Du} \quad ou \quad \frac{u'Eu}{u'Tu}$$

Nous choisirons le premier rapport parce qu'il est utilisé plus souvent (les deux sont admissibles et donnent des résultats identiques).

Il est équivalent de maximiser u'Eu/u'Du ou de maximiser u'Eu sujet à u'Du = 1 (en effet soit u le vecteur obtenu en solutionnant directement le rapport; si u'Du = c ≠ 1 on n'a qu'à poser u* = 1/√c u et on a le même maximum avec la contrainte respectée).

Comme déja vu en ACP, on a un problème de maximisation sous contrainte qui est résolu par la technique de Lagrange.

On trouve u est solution de

$$D^{-1}Eu = \lambda u$$
$$u'Du = 1$$

On reconnaît un problème de vecteurs propres et de valeurs propres. Le vecteur recherché est le vecteur propre associé à la plus grande valeur propre de $D^{-1}E$. Les autres vecteurs propres de cette matrice seront successivement les vecteurs, orthogonaux aux précédents (i.e. $u'_iDu_j=0$) donnant la meilleure séparation entre les groupes. On aura, au plus, k-1 valeurs propres non-nulles car le rang de la matrice E est de k-1 (k groupes centrés). Ainsi deux groupes centrés définissent une droite passant par l'origine (dimension 1), trois groupes définissent un plan (dimension 2), etc...

Remarque: Bien que D^{-1} et E soient des matrices symétriques, le produit $D^{-1}E$ ne donne pas une matrice symétrique. On pourrait donc craindre que les valeurs propres et les vecteurs propres ne soient pas réels et que les vecteurs propres ne soient pas orthogonaux. On peut démontrer que les valeurs propres et les vecteurs propres sont effectivement réels. La relation d'orthogonalité entre les vecteurs propres est légèrement modifiée; on a en effet: $u_i'Du_j=0$; $u_i'Eu_j=0$; $u_i'Tu_j = 0$ si i est différent de j. Ceci entraîne que les projections de X sur les vecteurs propres u ne sont pas corrélées.

Remarque: Si l'on a plus de groupes que de variables (et donc que la matrice E est inversible), alors les projections des observations sur les vecteurs propres représentent des distances de Mahalanobis. En effet, soit deux observations x et y (vecteurs colonnes px1). La projection de x sur l'ensemble des vecteurs propres est x'U. La projection de y est y'U. La distance (au carré) entre les deux projections est (x-y)'UU'(x-y). Or U'DU=I (condition de normalisation que l'on a imposée au départ de l'AD), ce qui implique que $UU'=D^{-1}$ (en effet prémultipliant U'DU par U et postmultipliant par U', on trouve UU'DUU'=UU'; postmultipliant par $(UU')^{-1}D^{-1}$, on trouve $UU'=D^{-1}$). La distance (au carré) entre les deux projections est : (x-y)'UU'(x-y)=(x-y)'D^{-1}(x-y). Lorsqu'il y a moins de groupes que de variables, $UU' \neq D^{-1}$ et les distances ne sont plus des distances de Mahalanobis au sens strict. En fait on montre (voir annexe) que dans ce cas $UU'=D^{-1}M$ (où $M = EU\Lambda^{-}U'$ et Λ^{-} est une matrice diagonale avec $1/\lambda_i$ sur la diagonale pour $\lambda_i > 0$ et 0 si $\lambda_i = 0$). M est une matrice idempotente (donc de projection) non symétrique, l'espace de projection étant défini par les centres des groupes (matrice E). On se trouve alors à calculer la portion de la distance de Mahalanobis contenue dans l'espace défini par les centres des groupes. C'est bien la seule distance qui importe puisque le complément est orthogonal à cet espace.

Remarque: Mesurer les distances avec la métrique D^{-1} revient à effectuer une rotation selon les axes principaux (voir ACP) de la matrice D et une normalisation pour que la dispersion intra-groupe soit 1. Par la suite on calcule la distance euclidienne habituelle. Bref, on calcule V tel que DV=VS (où S est la matrice diagonale contenant les valeurs propres de D). On

projette les observations centrées sur V (rotation) : X_cV, (note V'V=I) on normalise :$XV*S^{-0.5}$. La distance (au carré entre 2 observations transformées est alors :
$(x*-y*)(x*-y*)'=(x-y)VS^{-1}V'(x-y)'=(x-y)D^{-1}(x-y)'$. La dernière égalité provient du fait que :
DV=VS => DVV'=D=VSV' => $D^{-1}=(VSV')^{-1}=VS^{-1}V'$.
Bref, on peut interpréter l'AD comme une nouvelle façon de mesurer les distances dans l'espace original ou plutôt comme une transformation préalable à faire subir aux données avant de calculer la distance euclidienne. La transformation préalable vise à rendre les nouvelles variables non-corrélées et de dispersion (dans les groupes) unitaire.

Après avoir centré X, on calcule les coordonnées des observations sur les vecteurs propres en faisant $Co=X_cU$.

Si l'on calcule la matrice des produits croisés des coordonnées, on obtient :

$$Co'Co=U'X_c'X_cU=U' TU=U'(D+E)U=U'DU+U'EU= I+\Lambda$$

On voit que les coordonnées sur les différents vecteurs propres ne sont pas corrélées et que la dispersion sur chaque vecteur propre vaut $1+\lambda_i$ pour le $i^{ème}$ vecteur propre. λ_i est valeur propre de $D^{-1}E$.

5.2.2 Cas particulier de deux groupes

C'est un cas qui se présente très fréquemment et pour lequel la solution est particulièrement simple puisqu'on a alors un seul vecteur discriminant (vecteur propre de $D^{-1}E$). On peut montrer que le vecteur propre est donné par:

$$u = \sqrt{\frac{n_1 n_2}{n\lambda}} D^{-1}(y_1 - y_2)$$

La valeur propre associée est:

$$\lambda = (y_1\text{-}y_2)'D^{-1}(y_1\text{-}y_2)\, n_1 n_2/n = u'Eu$$

Dans le cas de deux groupes, on n'a donc aucune recherche de valeurs propres et vecteurs propres à effectuer.

5.3 Aspect classement

On a de nouvelles observations que l'on veut classer dans un des groupes connus uniquement à partir des valeurs mesurées.

Exemple: Vous prélevez un certain nombre de roches volcaniques en Abitibi pour lesquelles vous analysez les éléments majeurs. Vous formez deux groupes selon qu'il existe ou non un gisement connu situé à proximité de l'observation. Dans une nouvelle zone d'exploration, vous mesurez les mêmes variables et vous classez l'observation. Si celle-ci est classée dans le groupe "proximal", alors c'est que cette roche présente une signature géochimique plus similaire aux roches rencontrées à proximité des gisements qu'aux roches "distales". Il s'agit donc d'une zone favorable.

Exemple: Vous disposez d'images satellites dans plusieurs bandes de fréquences. Vous voulez utiliser cette information pour identifier les types de roches sur l'image. En quelques endroits (pixels), vous connaissez le type de roche pour l'avoir identifié sur le terrain. Vous formez des groupes avec ces pixels connus et vous cherchez à classer les autres pixels de l'image.

Remarque: Le classement est particulièrement indiqué lorsque les groupes sont difficiles à déterminer pour une raison ou une autre (coût, inaccessibilité,...).

Nous traiterons de deux approches différentes; une approche géométrique et une approche probabiliste (simplifiée).

5.3.1 Approche géométrique du classement

L'idée de base est très simple. Il s'agit de calculer la distance (définie par D^{-1}) entre la nouvelle observation et le centre de chacun des groupes. On classera la nouvelle observation dans le groupe pour lequel cette distance est minimale.

La distance entre une observation x (p x 1) et un groupe i s'écrit

$$d^2(x,y_i) = (x - y_i)' D^{-1} (x - y_i)$$

où y_i est le vecteur p x 1 des moyennes des p variables pour le groupe i. Développant le produit on trouve:

$$d^2(x,y_i) = x'D^{-1}x - 2x'D^{-1}y_i + y_i'D^{-1}y_i$$

Le terme $x' D^{-1} x$ ne dépend pas du groupe considéré. On veut classer dans le groupe pour lequel la distance est minimale. On peut tout aussi bien classer dans le groupe pour lequel g_i est maximal avec:

$$g_i =[x'D^{-1}y_i - 1/2\ y'_iD^{-1}y_i] * (n-k) = [x'D^{*-1}y_i - 1/2\ y'_iD^{*-1}y_i]$$

Les g_i sont ce que l'on appelle des "fonctions de classification" ou encore des "fonctions linéaires discriminantes". On en possède autant qu'il y a de groupes et on affecte la nouvelle observation au groupe pour lequel sa fonction de classification est maximale. Le facteur (n-k) est introduit pour pouvoir utiliser D* au lieu de D. En effet, D* est la matrice de covariances nécessaire pour pouvoir calculer les probabilités d'appartenance à chaque groupe (voir section 5.3.2)

5.3.1.1 Cas de deux groupes

On affecte l'observation au groupe 1 si $g_1 > g_2$ ou $g_1-g_2 > 0$

Or $g_1 - g_2$ s'écrit

$$x'D^{*-1}y_1 - 1/2y_1'D^{*-1}y_1 - x'D^{*-1}y_2 + 1/2y_2'D^{*-1}y_2 > 0$$

Ceci devient:

$$(x' - 1/2(y_1+y_2)') D^{*-1} (y_1-y_2) > 0$$

ou $$x'D^{*-1} (y_1-y_2) > 1/2(y_1+y_2)' D^{*-1} (y_1-y_2)$$

Comparant ces résultats au vecteur propre trouvé dans l'approche descriptive, on constate que le résultat du classement s'observe directement sur le premier vecteur propre. L'observation est classé dans le groupe dont le centre se projette du même côté par rapport au point milieu séparant les deux groupes.

5.3.2 Approche probabiliste (simplifiée)

L'idée est de classer une observation dans le groupe pour lequel la probabilité conditionnelle d'appartenir à ce groupe étant données les valeurs observées est maximale. En pratique on ne peut calculer ces probabilités que si les observations proviennent d'une loi multinormale. Si tel n'est pas le cas on devra au préalable transformer les données pour s'en rapprocher le plus possible. (La pratique a toutefois prouvée que l'AD était très robuste face à l'hypothèse de multinormalité).

La fonction de densité multinormale est:

$$f(x) = (2\pi)^{-p/2} \left|\Sigma\right|^{-1/2} \exp(-1/2(x - y)^1 \Sigma^{-1} (x - y))$$

Si x provient du groupe i alors sa fonction de densité est estimée par: $N(y_i, D^*_i)$.

De la définition de probabilité conditionnelle, si l'observation appartient nécessairement à un des k groupes, et si l'on suppose qu'à priori chaque groupe a une probabilité égale d'être observé, on a:

$$P(groupe\ i \mid x) = \frac{f_i(x)}{\sum_{j=1}^{k} f_j(x)}$$

Si l'on suppose de plus que les k groupes ont même matrice de covariances D alors on a:

$$P(groupe\ i|x) = \frac{\exp\left[-\frac{1}{2}(x - y_i)' D^{*-1} (x - y_i)\right]}{\sum_{j=1}^{k} \exp\left[-\frac{1}{2}(x - y_j)' D^{*-1} (x - y_j)\right]}$$

Après quelques manipulations, cette expression peut s'écrire:

$$P(groupe\ i|x) = \left[\sum_{j=1}^{k} \exp(g_j - g_i)\right]^{-1}$$

où les g_i sont les fonctions de classification décrites à la section précédente. Cette probabilité est maximale quand g_i est maximale (ou quand la distance d'un point au centre du groupe est minimale).

Conclusion: Les approches géométriques et probabilistes sont strictement équivalentes lorsque l'on a k populations multinormales avec mêmes matrices de covariances.

Remarque: Dans l'approche probabiliste, on peut inclure des probabilités à priori de rencontrer chaque groupe. Dans le cas de deux groupes, ceci revient d'un point de vue géométrique à déplacer le point milieu de façon à favoriser le groupe ayant la plus grande probabilité à priori d'être rencontrée. Également, on peut inclure des pénalités reliées au mauvais classement d'une observation. Toutefois, tous ces résultats ne sont valides que si l'hypothèse de multinormalité est respectée.

Remarque: Lorsqu'on permet que les matrices de variances-covariances D_i* varient d'un groupe à l'autre, on se trouve alors à effectuer une discrimination non-linéaire. Les zones attachées à chaque groupe ne sont plus délimitées par des plans (hyperplans) comme c'était le cas précédemment, mais plutôt par des surfaces courbes. On donne le nom de discrimination quadratique à cette approche. Elle est rarement utilisée.

Remarque: D'autres variantes existent encore pour l'AD. L'étude de celles-ci dépasse toutefois le cadre de ce cours.

5.3.3 Évaluation de la qualité de l'analyse discriminante

Il existe plusieurs façons de vérifier la qualité d'une analyse discriminante; certaines font appel à des hypothèses probabilistes, d'autres non. Les résultats présentés dans les sections suivantes le sont principalement pour référence car ces statistiques sont fréquemment utilisées dans les logiciels commerciaux.

5.3.3.1 Pourcentage de bien classés

C'est la statistique la plus utilisée et aussi la plus "parlante" tout en étant la plus simple.

L'idée est la suivante: on a une procédure de classement, alors pourquoi ne pas l'appliquer aux observations dont on connaît le véritable groupe et vérifier ainsi si l'on effectue un bon classement.

Exemple:

		Groupe AD	
		1	2
Groupe	1	50	10
véritable	2	30	110

Ici on aurait 160/200 = 80% des observations de bien classées. C'est un fort pourcentage si l'on considère qu'un classement fait entièrement de façon aléatoire donnerait en moyenne 50% de bien classés. De plus on note que les observations du groupe 1 sont bien classées dans une proportion de 83% alors que les observations du groupe 2 sont bien classées dans une proportion de 78%. Le groupe 1 est donc légèrement plus homogène que le groupe 2.

Notons que ce pourcentage de bien classés est trop optimiste, surtout lorsque le nombre d'observations est faible. En effet, si l'on forme deux groupes provenant d'une même population et que l'on applique l'analyse discriminante, on devrait trouver un pourcentage légèrement supérieur à 50% car les fonctions de classification s'ajustent aux variations échantillonnales. Une façon d'obtenir un estimé plus réaliste consiste à mettre de côté une certaine proportion des observations initiales de chaque groupe, de trouver les fonctions de classification avec les autres observations puis d'effectuer le classement des observations mises de côté (échantillon test). Une autre variante consiste à mettre de côté une observation à la fois et de répéter l'analyse et le classement n fois.

Remarque: Puisque le tableau de classement (appelé aussi matrice de confusion) est une forme de tableau de contingences, on peut tester le caractère significatif du classement à l'aide d'un test d'indépendance du Khi^2.

5.3.3.2 Lambda de Wilks

Cette statistique est définie comme étant le rapport des déterminants des matrices D et T.

$$L = |D| / |T| = |T^{-1}D|$$

$$L = \prod_{i=1}^{p} \gamma_i$$

où γ_i est une valeur propre de $T^{-1}D$.

La relation suivante relie les valeurs propres λ et γ:

$$\gamma = \frac{1}{\lambda + 1}$$

Sous hypothèse de multinormalité et d'égalité des matrices de covariances, on peut montrer que

$$-[\, n - (p+k)/2 - 1 \,] \, Ln \, L$$

où n est le nombre total d'observations.
p est le nombre de variables.
k est le nombre de groupes.

est approximativement distribuée suivant une loi Khi^2 avec p(k-1) degrés de liberté.

Lorsque l'on a plusieurs groupes (k > 2) et que l'on veut vérifier le caractère significatif des vecteurs propres qui restent après en avoir accepté q, on peut formuler le test suivant:

H_0: les vecteurs propres q+1, q+2 .... k-1 n'ajoutent rien à la discrimination des k groupes.
H_1: non H_0

alors

$$-[\, n - (p+k)/2 - 1 \,] \, Ln \, L^*$$

où L^* est donné par:

$$L^* = \prod_{i=q+1}^{k-1} \gamma_i$$

est approximativement distribué selon une loi Khi^2 avec (p-q)(k-q-1) degrés de liberté.

Un autre test similaire à ce dernier utilise le fait que (n-k) λ_q est approximativement distribué suivant une loi Khi^2 avec (p+k-2q) degrés de liberté. On vérifie successivement si la $1^{ère}$ (q=1) valeur propre est significative, puis la 2^e (q=2), et ainsi de suite.

Remarque: Ces deux derniers tests sont utiles surtout pour des fins de description. Ces résultats ne peuvent pas être incorporés dans l'étape classement.

5.3.3.3 Le "V" de Rao.

La statistique V mesure la somme des distances entre les centres des groupes et la moyenne globale. La distance est normalisée par la matrice D^{*-1} (généralisation de la distance de Mahalanobis). Elle est définie comme étant:

$$V = \sum_{i=1}^{k} n_i (y_i - y)' D^{*-1} (y_i - y)$$

où y_i: vecteur moyenne du groupe i (px1)
y : vecteur moyenne totale
D* : matrice de variance-covariance intra-groupe (i.e. D/(n-k) ou k est le nombre de groupes et D est la matrice des produits croisés intra-groupes)
n_i est le nombre d'observations dans le groupe i, n est le nombre total d'observations.

On peut démontrer que sous hypothèse de multinormalité et d'égalité des matrices de covariances, V est distribuée suivant une Khi^2 avec p(k-1) degrés de liberté.

Également si on effectue la discrimination avec p variables puis avec p+1 variables, on peut vérifier le caractère significatif de l'ajout de la variable. En effet, le changement de V (i.e. V_{fin}-V_{ini}) est alors distribué suivant une Khi^2 avec (k-1) degrés de liberté.

5.3.3.4 Corrélation canonique ou pouvoir discriminant d'un vecteur propre

Soit le rapport:

$$\alpha = \frac{u'Eu}{u'Tu}$$

Par un développement similaire à ce qui a été vu précédemment, on montre que α est valeur propre de $T^{-1}E$. Cette valeur propre est reliée aux valeurs propres λ de $D^{-1}E$ par:

$$\alpha = \frac{\lambda}{1+\lambda}$$

Ce rapport α exprime la proportion de la variabilité totale imputable aux différences entre les centres des groupes. Cette quantité est donc analogue au R^2 en régression. Pour cette raison, on définit $\alpha^{1/2}$ comme le coefficient de corrélation canonique ou pouvoir discriminant.

Remarque: Le nom corrélation canonique fait référence à une méthode appelée analyse canonique. Cette méthode étudie deux ensembles de variables mesurées sur un même ensemble d'observations. Elle cherche les combinaisons linéaires des deux ensembles de variables qui seront le plus corrélées entre elles. En AD, les deux ensembles de variables sont d'une part les p variables mesurées et d'autre part, les (k-1) variables indicatrices permettant d'identifier les groupes. La corrélation maximale que l'on peut obtenir entre ces deux ensembles de variables est précisément $\alpha^{1/2}$

5.3.4 Test d'égalité des matrices de covariances intra-groupes

Le calcul des probabilités ainsi que les différents tests présentés précédemment pour le V de Rao et le Lambda de Wilks nécessitent la multinormalité des observations et l'égalité des matrices de covariances à l'intérieur de chaque groupe. On peut tester cette dernière hypothèse par le test approximatif suivant (test de Kullback[1] (1959)) nécessitant aussi la multinormalité des observations :

$$\chi^2 = \sum_{i=1}^{k} \frac{n_i - 1}{2} \ln \frac{|D^*|}{|D_i^*|}$$

est approximativement distribué suivant une loi Khi^2 avec (k-1)*n*(n+1)/2 d.l.. On rejette l'hypothèse d'égalité des matrices de variance-covariance lorsque la statistique excède le seuil lu dans une table Khi^2.

D^* est la matrice de variance-covariance intra-groupes
D_i^* est la matrice de variance-covariance pour le groupe i
n_i est le nombre d'observations dans le groupe i
n est le nombre total d'observations
| | signifie le déterminant

5.4 Procédures de sélection des variables

On est souvent intéressé à obtenir la meilleure discrimination possible avec le minimum de variables, possiblement pour des raisons d'interprétation, de robustesse des résultats, de fiabilité, sûrement pour des raisons économiques. En effet avec des analyses géochimiques, par exemple, si on obtient une aussi bonne (et parfois meilleure) discrimination avec trois variables qu'avec huit, on vient d'économiser un coût considérable.

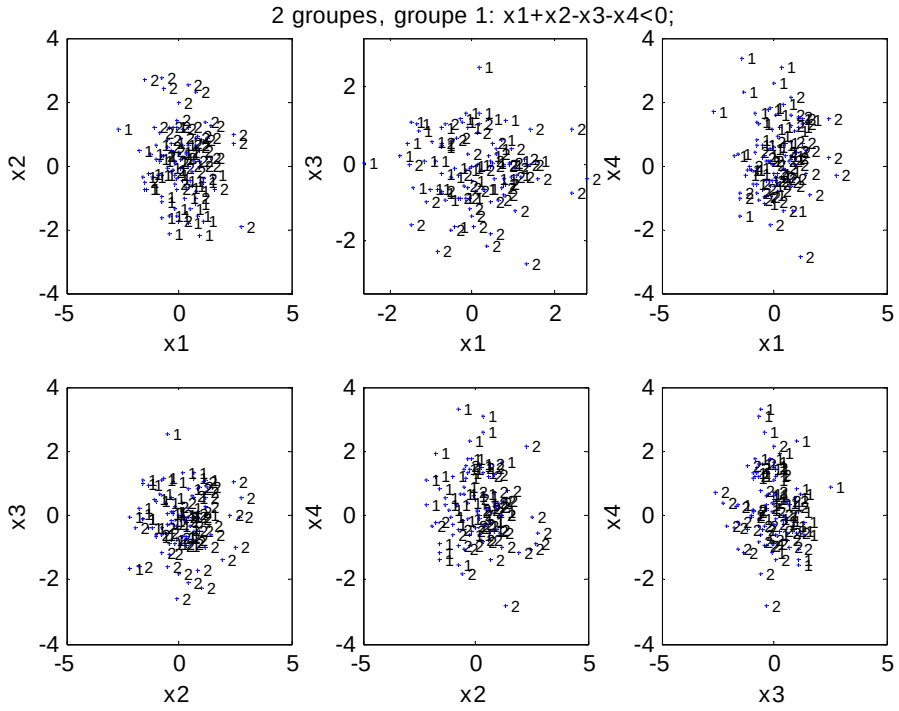
Les mêmes procédures vues en régression peuvent être utilisées ici, i.e. sélection avant, élimination arrière et "stepwise". La section 5.3.3 a été consacrée à la définition de statistiques qui peuvent toutes servir de critère d'inclusion ou d'élimination. D'autres critères sont présentés dans certains programmes d'analyse discriminante (ex. SPSS). Certains de ces critères permettent de vérifier si l'ajout d'une variable supplémentaire est significatif (ex. V de Rao) d'autres ne le permettent pas (Lambda de Wilks, pourcentage de bien classés, corrélation canonique).

[1] Kullback, 1959, Information theory and statistics, Wiley, 395p.

Malgré la diversité des méthodes, la pratique montre que le sous-ensemble de variables retenues est relativement robuste au choix du critère d'inclusion. De plus, même si deux sous-ensembles diffèrent quant aux variables retenues, très souvent l'interprétation est identique et les performances (classement) très comparables. Tout ceci est finalement rassurant pour l'utilisateur.

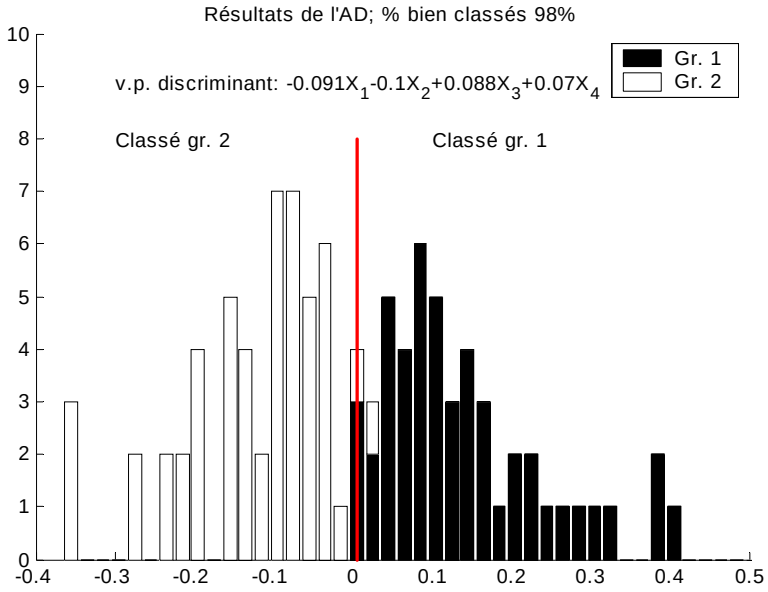
5.4.1 Exemple d'analyse discriminante (2 groupes)

Voici un exemple simulé avec 100 observations et 6 variables. Le groupe 1 est formé en prenant toutes les observations telles que x1+x2-x3-x4<0. La figure suivante montre que la discrimination n'est pas parfaite sur tous les diagrammes binaires des variables x1 à x4.



Les critères % de bien classés, lambda de Wilks et V de Rao sélectionnent les variables 1 à 4 en premier. Le vecteur propre discriminant est [–0.091 –0.1 0.088 0.07]. Les projections sur ce vecteur sont représentées sur la figure suivante. L'on constate que :

- Le vecteur discriminant forme un angle de seulement 7.3^{o} avec l'équation réelle de formation des groupes.
- Le % de bien classés est de 98%, nettement mieux que ce que l'on pouvait accomplir visuellement sur les diagrammes binaires.



5.5 Exemple d'application: indice de prospection géochimique.

En Abitibi, les gisements de cuivre et de zinc sont, la plupart, considérés comme étant d'origine volcanogène. Ce mode de formation implique la circulation d'eau marine à l'intérieur des empilements volcaniques. En circulant, cette eau altère la roche en lessivant certains éléments (tels le Na_2O et le CaO et en enrichissant d'autres éléments tels le MgO et le K_2O. Ces signatures géochimiques associées aux gisements volcanogènes constituent des cibles de prospection des plus propices puisque leur étendue est bien plus grande que celle du gisement lui même. Ces signatures peuvent toutefois varier de gisement en gisement aussi bien par les éléments les plus affectés que par l'étendue du halo d'altération.

Afin de définir un indice de prospection unique pour les gisements volcanogènes de l'Abitibi, Marcotte et David (1981) ont utilisé l'analyse discriminante. Les données, 574 analyses géochimiques, provenaient de la littérature et les variables impliquées étaient les principaux éléments majeurs. Seules des analyses provenant de roches volcaniques décrites comme roches à grain fin et ayant plus de 60% de SiO_2 ont été utilisées. Pour l'AD, deux groupes ont été formés selon un critère de distance. On fixait un seuil de distance par rapport au gisement le plus près et on formait un groupe proximal et un groupe distal. L'analyse discriminante était ensuite appliquée à ces deux groupes. En faisant varier le critère de distance de 0.25 mille à 1.5 milles par pas de 0.25 mille, on a pu déterminer, pour chaque gisement, quelle semblait être l'extension du halo d'altération. Cette extension était considérée comme étant la distance à laquelle les résultats de l'AD étaient optimaux pour le gisement considéré. On a par la suite repris l'AD en adoptant cette fois comme critère de distance l'extension estimée du halo pour chaque gisement. L'AD effectuée avec ces groupes a donné plus de 80% de bien classés, un résultat jugé intéressant. L'équation linéaire discriminante retenue indiquait un lesssivage en Na_2O, en CaO et en FeO et un enrichissement en MgO, résultats compatibles avec ce qui était décrit dans la littérature.

$$\text{Indice} = 1.91 - 0.63\ Na_2O - 0.26\ CaO - 0.18\ FeO + 0.30\ MgO + 1.44\ TiO_2$$

Appliqué au gisement de Normétal (qui n'était pas inclus dans l'analyse précédente), l'indice a su détecter la présence du gisement. Une zone favorable située plus au nord-ouest est également indiquée. Quelques anomalies sont reliées à l'intrusion granitique située au nord-ouest.