Chapitre 21

Analyse factorielle discriminante (A.F.D.)

L'analyse discriminante est une des **techniques de *scoring***. C'est une méthode de classement. Il cherche à déterminer la contribution des variables qui expliquent l'**appartenance des individus à des groupes connus *a priori***. Deux ou plusieurs groupes sont comparés sur plusieurs variables pour déterminer s'ils différent et pour comprendre la nature des différences. L'analyse discriminante permet de 1. décrire et de 2. classer.

1. Parmi les groupes connus, quelles sont les principales différences que l'on peut déterminer à l'aide des variables mesurées ?
2. Peut-on déterminer le groupe d'appartenance d'une nouvelle observation uniquement à partir des variables mesurées ?

L'**analyse factorielle discrimante** [1] (A.F.D.) poursuit deux objectifs :

1. la discrimination des classes, c'est-à-dire déterminer les fonctions linéaires discriminantes sur l'échantillon d'apprentissage, ou plus mathématiquement, déterminer une combinaison linéaire des p variables explicatives dont les valeurs séparent au mieux les k classes ;
2. l'affectation d'un nouvel individu dans une classe, c'est-à-dire déterminer la classe de nouveaux individus pour lesquels on observe les valeurs p variables explicatives. Dit autrement, c'est un problème de classement.

L'idée du principe de la discrimination repose sur le fait que la discrimination visuelle est plus aisée si :

1. les centres de gravité de chaque sous-nuage appartenant à une seule classe sont éloignées ;

1. ou analyse linéaire discriminante

2. chaque sous-nuage appartenant à une seule classe sont les plus homogènes possibles autour de ces centres de gravité.

Pour y parvenir, **il faut maximiser les variances interclasses** [2] (entre les classes) **et minimiser les variances intra-classes** [3] (à l'intérieur des classes).

S'il existe une région importante d'incertitude dans laquelle pour les mêmes valeurs des variables, on trouve des individus appartenant à plusieurs groupes. **Le but de l'analyse discriminante est de trouver un nouvel axe**, combinaison linéaire des variables, **qui permet de réduire cette zone d'incertitude et de séparer au mieux les deux groupes**.

Pour ce, on utilise l'analyse de variance multiple, mais la méthode peut comporter de grosses imprécisions. On peut dès lors considérer une nouvelle variable qui soit une combinaison linéaire des précédentes. Géométriquement, cette nouvelle variable est représentée par un axe sur lequel on projette les divers points des groupes de sujets. L'axe est appelé **axe de la fonction discriminante**. Les points projetés sur cet axe se distribuent normalement pour chacun des groupes. L'axe doit occuper une partition telle que la projection des points donne lieu au minimum de superposition des divers groupes de sujets.

L'analyse discriminante a pour objectif de déterminer cet axe optimum de la fonction discriminante, c'est-à-dire de calculer les éléments d'un vecteur k qui définissent une combinaison linéaire des variables. Des fonctions discriminantes orthogonales successives s'apparentent aux composantes principales.

Le nombre possible de ces fonctions est limité par le nombre de variables et par le nombre de groupes. Il est souhaitable de limiter à deux ou trois les fonctions discriminantes.

L'approche de l'analyse discriminante est proche de celle de la régression. On calcule un score Y à partir de X_1, X_2, ..., X_n variables tel que :

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n \tag{21.1}$$

avec α_n les **coefficients discriminants**, c'est-à-dire les poids des variables. Elle correspond à la combinaison linéaire qui différencie de manière maximale entre les groupes.

En analyse discrimante, la variable dépendante est l'appartenance à un groupe

1. On teste si, au niveau des scores discriminants, on obtient une différenciation significative entre les groupes. Pour ce, on utilise un test F aux valeurs de la statistique D^2 de Prasanta Chandra Mahalanobis datant de 1936. Il mesure la distance de chaque case à la moyenne du groupe, tout en permettant des axes corrélés et des unités de mesures différentes. Il teste l'hypothèse

Prasanta Chandra Mahalanobis (1893-1972)

2. ou externes
3. ou internes

que la distance entre les deux groupes est différente de zéro [Mahalanobis, 1936].

2. Les coefficients discriminants sont appliqués aux valeurs bruts des variables. De fait, ils sont influencés par l'échelle de mesure des variables. Il y a un risque d'effet d'échelle. Ainsi, pour comparer les contributions de chaque variable, on utilise des **coefficients standardisés**. Ils sont obtenus par la multiplication des coefficients bruts de chaque variable par l'écart type pour l'ensemble des groupes.
3. Pour clarifier les individus dans un des groupes, on doit fixer un **score critique** [4] qui joue le rôle de frontière entre les groupes. Ce score est en général la moyenne des scores des deux groupes.

   N.B. 1. Si les groupes sont de dimensions égales le score critique est égale à la moyenne des moyennes des scores des groupes.

   N.B. 2. Si les groupes ne sont pas égaux, on utilise une moyenne pondérée.

21.1 Les fonctions discriminantes

En plus d'être indépendantes, les uns des autres, les fonctions discriminantes ont un pouvoir de discrimination qui décroît d'une fonction à l'autre.

Le nombre de fonctions discriminantes est la plus petite valeur entre le nombre de variables et le nombre de groupes moins un.

Les fonctions discriminantes présentent une grande analogie avec les facteurs mis en évidence de l'analyse factorielle. Il est possible de mesurer la contribution des variables, ce qui permet d'identifier le rôle des fonctions discriminantes.

L'intérêt des fonctions additionnelles va en décroissant. Beaucoup d'analyses ne dépassent pas deux ou trois fonctions discriminantes.

L'**effet de discrimination** de la fonction discriminante i par rapport à toutes les fonctions est exprimé par la proportion :

$$\frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k} \tag{21.2}$$

Ce rapport exprime la proportion de la variance expliquée par chaque fonction discriminante, mais cette proportion ne conduit pas à une décision statistique au sens standard de l'expression. De fait, on utilise l'indicateur Λ :

$$\Lambda = \prod_{i=1}^{p} \frac{1}{1+\lambda_i} \tag{21.3}$$

4. En anglais : *cutting score*

avec p le nombre de fonction discriminante. Λ exprime la capacité de discrimination d'un ensemble de variables. Pour les fonctions au-delà de la k-ième fonction, on a :

$$\Lambda' = \prod_{j=k+1}^{p} \frac{1}{1+\lambda_i} \tag{21.4}$$

Cela correspond à une mesure de l'**inverse de la puissance discriminative expliquée** par les fonctions discriminantes à venir. La signification de la discrimination des fonctions restantes k à p à la suite de l'acceptation des k premières, peut se calculer au moyen de l'approximation de Maurice Stevenson Bartlett.

Maurice Stevenson Bartlett (1910-2002)

$$\chi^2 = \left[n - \frac{1}{2}(v+g)\right] \ln(\Lambda') \tag{21.5}$$

avec $ddl = (v-k)(g-k-1)$, v le nombre de variables et g le nombre de groupes. Si, pour les fonctions discriminantes $(k+1)\,p$, on obtient une valeur χ^2 qui ne dépasse pas le seuil critique. On considère que les k premières fonctions calculées suffisent seules à expliquer de manière significative les écarts entre les groupes.

21.2 Le choix des variables

Plusieurs méthodes peuvent être utilisées dans le choix des variables à inclure dans la construction des fonctions discriminantes.

1. On peut factoriser au préalable les variables. Cela permet de réduire le nombre de variables. On applique une analyse factorielle à l'ensemble des variables, puis on introduit les scores sur les axes factoriels dans l'analyse discriminante.
2. On peut utiliser une approche hiérarchique [5] dans laquelle les variables sont introduites une à une selon leur capacité décroissante à mettre en évidence la différence entre les groupes. Au cours des sélections successives, il est possible que des variables déjà centrées perdent leur pouvoir de discrimination à cause d'une **redondance d'information**, c'est-à-dire que le pouvoir de discrimination de cette variable est désormais inclus dans quelque combinaison de nouvelles variables retenues.
3. Divers critères mettent l'accent sur l'un ou l'autre de la dispersion des groupes. Ils permettent de sélectionner les variables :

   — Le test de Martin Bradbury Wilks vise à minimiser un rapport dans

Martin Bradbury Wilks (1922-2013)

5. En anglais : *stepwise*

lequel entrent en considération la dispersion des centroïdes et la cohésion des cas au sein des groupes.

— Plusieurs tests reliés à la notation de distance de Mahalanobis, visent à maximiser l'écart entre les deux groupes les plus rapprochés.

— La méthode de Calyampudi Radhakrishna Rao consiste à choisir la variable qui contribue le plus à une distance généralisée, évaluée sur les variables précédentes.

Pour tous les critères, une variable est sélectionnée lorsque son rapport F partiel dépasse une valeur critique, c'est-à-dire lorsque sa contribution à la dispersion additionnelle des centroïdes est statistiquement significative.

L'analyse discriminante peut être vue comme **un cas spécial d'analyse factorielle**, mais son objectif diffère. Il s'agit de **faire ressortir au maximum les différences entre des groupes mesurés dans un espace multidimensionnel** en projetant chaque cas dans l'espace unidimensionnel d'un petit nombre de fonctions linéaires orthogonales. Elle fait suite à celle de l'**analyse de variance multivariée**. En présence d'une situation dans laquelle plusieurs groupes sont mesurés sur plusieurs variables, on s'intéresse à **déterminer s'il existe une différence significative entre les groupes**. Si elle existe et est établie, il faut déterminer les variables responsables **dans un ordre décroissant d'importance des différences entre les groupes**. C'est l'**objectif de l'analyse discriminante**. Toutefois, il est possible de proposer une exploitation plus poussée des résultats **en proposant une classification des nouveaux sujets dans les divers groupes** en se donnant comme objectif une probabilité minimale d'erreurs.

Le rôle de l'analyse discriminante peut être envisagé de deux manières au niveau de l'attribution des qualitatifs d'**indépendance** et de **dépendance** aux variables mesurées sur les populations visées et aux fonctions discriminantes :

— soit les populations sont considérées comme des variables indépendantes et les fonctions discriminantes comme des variables dépendantes ;

— soit les fonctions discriminantes sont considérées comme des variables indépendantes et les populations comme des variables dépendantes.

L'analyse discriminante consiste à projeter dans un sous-espace approprié des échantillons de mesures multidimensionnelles. Elle est principalement utilisée dans le marketing.

21.3 Le principe de base

Soit une matrice $\mathbf{X}$ de scores centrés réduits de v variables et n individus. Cette matrice est partitionnée en g sous-matrices D_i de n_i cas chacune. À chaque partition correspond :

— un centroïde m_i, ligne de la matrice $\mathbf{M}$;

— une matrice de covariation $\mathbf{W}$;

— une matrice de covariance $V_i = \frac{1}{n_i - 1} W_i$

On considère que les matrices de covariance sont homogènes. La matrice de variation $\mathbf{W} = W_1 + \ldots + W_g$ et la matrice des covariances inter-groupes $\mathbf{B}$ est définie par :

$$\mathbf{B} = \frac{1}{g-1} {}^t\mathbf{M}.\mathbf{M} \tag{21.6}$$

On cherche une combinaison linéaire $\mathbf{y}$ des $\mathbf{X}$ pour obtenir la fonction discriminante y. On l'obtient avec une combinaison linéaire $\mathbf{k}$ de la matrice $\mathbf{X}$ de telle sorte qu'à une variance intra-groupe donnée corresponde un maximum de la variance inter-groupe. On peut représenter la dispersion des cas et moyennes des groupes sur l'axe $\mathbf{k}$ de la fonction discriminante.

Son expression générale est définie par :

$$\mathbf{y} = \mathbf{X}.\mathbf{k} \tag{21.7}$$

Si on considère les projections sur la fonction discriminante des centroïdes, alors ces nouveaux centroïdes auront pour scores discriminants :

$$\mathbf{S} = \mathbf{M}.\mathbf{k} \tag{21.8}$$

La variance inter-groupe de ces moyennes ${}^t\mathbf{S}.\mathbf{S}$ vaut :

$$\frac{1}{g-1} {}^t\mathbf{k}.{}^t\mathbf{M}.\mathbf{M}.\mathbf{k} = {}^t\mathbf{k}.\mathbf{B}.\mathbf{k} \tag{21.9}$$

La variance intra-groupe est donnée par :

$${}^t\mathbf{k}.\mathbf{V}.\mathbf{k} \tag{21.10}$$

Le problème consiste à maximiser la variance inter-groupe par rapport à un niveau fixé de variance intra-groupe. Si on fixe la variance intra-groupe à l'unité, on peut exprimer la fonction par le lagrangien suivant :

$$\mathbf{F} = {}^t\mathbf{k}.\mathbf{B}.\mathbf{k} - \lambda \left({}^t\mathbf{k}.\mathbf{V}.\mathbf{k} - 1 \right) \tag{21.11}$$

Il faut déterminer la valeur de $\mathbf{k}$ qui maximise $\mathbf{F}$.

$$\frac{\partial \mathbf{F}}{{}^t\mathbf{k}} = 2\mathbf{B}.\mathbf{k} - 2\lambda \mathbf{V}.\mathbf{k} = 0 \tag{21.12}$$

$$\Leftrightarrow \mathbf{B}.\mathbf{k} - \lambda \mathbf{V}.\mathbf{k} = 0 \Leftrightarrow (\mathbf{B} - \lambda \mathbf{V}).\mathbf{k} = 0 \tag{21.13}$$

On sait que cette équation comporte une solution si $(\mathbf{B} - \lambda\mathbf{V}) = 0$, car $\mathbf{k}$ ne peut être un vecteur nul.

$$(\mathbf{B} - \lambda\mathbf{V}) = 0 \tag{21.14}$$

$$\Leftrightarrow \mathbf{V}^{-1}.(\mathbf{B} - \lambda\mathbf{V}) = 0 \tag{21.15}$$

$$\Leftrightarrow \left(\mathbf{V}^{-1}.\mathbf{B} - \mathbf{V}^{-1}.\lambda.\mathbf{V}\right) = 0 \tag{21.16}$$

$$\Leftrightarrow \left(\mathbf{V}^{-1}.\mathbf{B} - \lambda.\mathbf{1}_n\right) = 0 \tag{21.17}$$

et $|\mathbf{V}^{-1}.\mathbf{B} - \lambda.\mathbf{1}_n| = 0$. La forme est plus proche de celle d'une A.C.P. La matrice est **non symétrique**. Pour la transformer une matrice symétrique, on utilise le théorème qui précise qu'une matrice non symétrique peut s'écrire comme la somme d'une matrice symétrique et d'une matrice non symétrique. Pour ce, on utilise $\mathbf{D}$ la matrice diagonale de $\mathbf{V}$.

$$\mathbf{V}^{-1}.\mathbf{B} = \mathbf{D}^{-\frac{1}{2}}.\mathbf{B}.\mathbf{D}^{-\frac{1}{2}} \tag{21.18}$$

ce qui permet de calculer les valeurs propres et les vecteurs propres, solutions de la maximisation recherchée.

21.4 L'analyse discriminante

On pose :

— n : nombre total d'observations ;

— p : nombre de variables mesurées ;

— k : nombre de groupes ;

— n_k : nombre d'observations dans le groupe k ;

— $\mathbf{T}$: matrice de variabilité totale $(p \times p)$;

— $\mathbf{T}^*$: matrice de covariance totale : $\frac{1}{n-1}\mathbf{T}$;

— $\mathbf{E}$: matrice de variabilité entre les groupes $(p \times p)$;

— $\mathbf{E}^*$: matrice de covariances entre les groupes : $\frac{1}{k-1}\mathbf{E}$;

— $\mathbf{D}$: matrice de variabilité dans les groupes $(p \times p)$;

— $\mathbf{D}^*$: matrice de covariances dans les groupes : $\frac{1}{n-k}\mathbf{D}$;

— $\mathbf{X}$: matrice des observations ;

N.B. Les observations sont placées par groupes les uns à la suite des autres.

— $\mathbf{C}$: matrice des moyennes des p variables dans les k groupes répétées n_k fois $(n \times p)$;

— $\mathbf{y_i}$: vecteur $(p \times 1)$ des moyennes des p variables pour le groupe i.

Soit la matrice $\mathbf{X}$:

$$\mathbf{X} = \left[\begin{array}{cccc} x_{11}^1 & x_{12}^1 & \dots & x_{1p}^1 \\ \dots & \dots & \dots & \dots \\ x_{n_1 1}^1 & x_{n_1 2}^1 & \dots & x_{n_1 p}^1 \\ --- & --- & --- & --- \\ x_{11}^2 & x_{12}^2 & \dots & x_{1p}^2 \\ \dots & \dots & \dots & \dots \\ x_{n_2 1}^2 & x_{n_2 2}^2 & \dots & x_{n_2 p}^2 \\ --- & --- & --- & --- \\ \dots & \dots & \dots & \dots \\ --- & --- & --- & --- \\ x_{11}^k & x_{12}^k & \dots & x_{1p}^k \\ \dots & \dots & \dots & \dots \\ x_{n_k 1}^k & x_{n_k 2}^k & \dots & x_{n_k p}^k \end{array}\right] \tag{21.19}$$

Soit la matrice $\mathbf{C}$:

$$\mathbf{C} = \left[\begin{array}{cccc} y_{11} & y_{12} & \dots & y_{1p} \\ \dots & \dots & \dots & \dots \\ y_{n_1 1} & y_{n_1 2} & \dots & y_{n_1 p} \\ --- & --- & --- & --- \\ y_{11} & y_{12} & \dots & y_{1p} \\ \dots & \dots & \dots & \dots \\ y_{n_2 1} & y_{n_2 2} & \dots & y_{n_2 p} \\ --- & --- & --- & --- \\ \dots & \dots & \dots & \dots \\ --- & --- & --- & --- \\ y_{11} & y_{12} & \dots & y_{1p} \\ \dots & \dots & \dots & \dots \\ y_{n_k 1} & y_{n_k 2} & \dots & y_{n_k p} \end{array}\right] \tag{21.20}$$

Soit un vecteur $\mathbf{u}_1$. On choisit $\mathbf{u}_1$ de telle sorte que les projections des moyennes des groupes sur $\mathbf{u}_1$ soient le plus espacées possibles, et que, simultanément, les projections des observations d'un même groupe soient le plus rapprochées possibles de la projection de la moyenne du groupe. **Sur ce vecteur $\mathbf{u}_1$, on cherche à observer des groupes compacts et distincts les uns des autres**.

La matrice $\mathbf{X}$ centrée par rapport aux moyennes calculées avec toutes les observations sans tenir compte du groupe est donnée par :

$$\mathbf{X}_c = \mathbf{X} - \frac{1}{n}\mathbf{1}_n.{}^t\mathbf{1}_n.\mathbf{X} \tag{21.21}$$

De même, la matrice $\mathbf{C}$ centrée, c'est-à-dire la matrice contenant les moyennes de chaque groupe centrées par rapport à la moyenne globale s'écrit :

$$\mathbf{C}_c = \mathbf{C} - \frac{1}{n}\mathbf{1}_n.{}^t\mathbf{1}_n.\mathbf{X} \tag{21.22}$$

On pourrait également centrer chaque observation de la matrice $\mathbf{X}$ par rapport à la moyenne du groupe correspondant :

$$\mathbf{X}_g = \mathbf{X} - \mathbf{C} \tag{21.23}$$

On a :

$$\mathbf{X}_c = \mathbf{C}_c - \mathbf{X}_g \tag{21.24}$$

La matrice de variable totale s'écrit :

$$\mathbf{T} = {}^t\mathbf{X}_c.\mathbf{X}_c \tag{21.25}$$

$$\mathbf{T} = {}^t\mathbf{C}_c.\mathbf{C}_c + {}^t\mathbf{X}_g.\mathbf{X}_g \tag{21.26}$$

car ${}^t\mathbf{X}_g.\mathbf{C}_c$.

$$\mathbf{T} = \mathbf{E} + \mathbf{D} \tag{21.27}$$

Le premier membre de droite représente la matrice de variabilité entre les centres des groupes, $\mathbf{E}$ signifiant « entre ». Ce second membre représente la matrice de variabilité à l'intérieur des groupes, $\mathbf{D}$ signifiant « dans ».

Les groupes seront d'autant plus faciles à discriminer que $\mathbf{E}$ sera grand par rapport à $\mathbf{D}$ (ou à $\mathbf{T}$).

— Si $\mathbf{E}$ est grand, alors les centres des groupes sont éloignés.

— Si $\mathbf{D}$ est petit, alors les observations d'un même groupe sont proches.

21.4.1 Comment rechercher le vecteur $\mathbf{u}$ séparant le mieux possible les groupes ?

Soit $\mathbf{u}$ un vecteur sur lequel seront effectuées les projections des observations :

$$\mathbf{X}_C.\mathbf{u} \tag{21.28}$$

La variabilité des projections est mesurée par :

$${}^t\mathbf{u}.\mathbf{T}.\mathbf{u} \tag{21.29}$$

or $\mathbf{T} = \mathbf{E} + \mathbf{D} \Rightarrow {}^t\mathbf{u}.\mathbf{T}.\mathbf{u} = {}^t\mathbf{u}.\mathbf{E}.\mathbf{u} + {}^t\mathbf{u}.\mathbf{D}.\mathbf{u}$.

Le vecteur $\mathbf{u}$ recherché maximise le rapport :

$$\frac{{}^t\mathbf{u}.\mathbf{E}.\mathbf{u}}{{}^t\mathbf{u}.\mathbf{D}.\mathbf{u}} \tag{21.30}$$

ou

$$\frac{{}^t\mathbf{u}.\mathbf{E}.\mathbf{u}}{{}^t\mathbf{u}.\mathbf{T}.\mathbf{u}} \tag{21.31}$$

Les deux rapports donnent des résultats identiques.

Si ${}^t\mathbf{u}.\mathbf{D}.\mathbf{u} = \mathbf{C} \neq 1$, on pose $\mathbf{u}^* = \frac{1}{\sqrt{c}}\mathbf{u}$ pour obtenir le maximum avec le rapport de la contrainte. Le problème de maximisation sous contrainte est résolu par la technique de Lagrange. $\mathbf{u}$ est solution de :

$$\mathbf{D}^{-1}.\mathbf{E}.\mathbf{u} = \lambda\mathbf{u} \tag{21.32}$$

et

$${}^t\mathbf{u}.\mathbf{D}.\mathbf{u} = 1 \tag{21.33}$$

Le vecteur recherché est le vecteur propre associé à la plus grande valeur propre de $\mathbf{D}^{-1}.\mathbf{E}$. Les autres vecteurs propres de cette matrice sont successivement les vecteurs orthogonaux aux précédents :

$${}^t\mathbf{u}_i.\mathbf{D}.\mathbf{u}_j = 1 \tag{21.34}$$

si $i \neq j$. Ils fournissent la meilleure séparation entre les groupes.

N.B. 1. On aura au plus $(k-1)$ valeurs propres non nulles.

N.B. 2. Les projections des observations sur les vecteurs propres représentent des distances de Mahalanobis.

N.B. 3. Mesurer les distances avec la métrique $\mathbf{D}^{-1}$ revient à effectuer une rotation selon les axes principaux de la matrice $\mathbf{D}$, et une normalisation pour que la dispersion intra-groupe vale 1.

— On calcule la distance euclidienne :

$$\mathbf{S} = \begin{bmatrix} \lambda_1 & \dots & 0 \\ 0 & \dots & 0 \\ 0 & \dots & \lambda_{k-1} \end{bmatrix} \tag{21.35}$$

avec λ_i les valeurs propres de $\mathbf{D}$.

$$\mathbf{D}.\mathbf{V} = \mathbf{V}.\mathbf{S} \tag{21.36}$$

— On projette les observations centrées sur $\mathbf{V}$ avec : $\mathbf{X}_c.\mathbf{V}$.

— On normalise avec $\mathbf{X}.\mathbf{V}^*.\mathbf{S}^{-\frac{1}{2}}$

— La distance au carré entre deux observations transformées vaut :

$$d^2 = (x^* - y^*) .^t (x^* - y^*) \tag{21.37}$$

$$d^2 = (x - y) .\mathbf{V}.\mathbf{S}^{-1}.^t\mathbf{V}.^t (x - y) \tag{21.38}$$

$$d^2 = (x - y) .\mathbf{D}^{-1}.^t (x - y) \tag{21.39}$$

— **On peut interpréter l'analyse discriminante comme une nouvelle manière de mesurer les distances dans l'espace original, ou 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 unitaires.

Après avoir centré $\mathbf{X}$, on calcule les coordonnées des observations sur les vecteurs propres en faisant :

$$\mathbf{C}_0 = \mathbf{X}_c.\mathbf{U} \tag{21.40}$$

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

$$^t\mathbf{C}_0.\mathbf{C}_0 = ^t\mathbf{U}.^t\mathbf{X}_c.\mathbf{X}_c.\mathbf{U} \tag{21.41}$$

$$\Leftrightarrow ^t\mathbf{C}_0.\mathbf{C}_0 = ^t\mathbf{U}.\mathbf{T}.\mathbf{U} \tag{21.42}$$

$$\Leftrightarrow ^t\mathbf{C}_0.\mathbf{C}_0 = ^t\mathbf{U}.(\mathbf{D} + \mathbf{E}).\mathbf{U} \tag{21.43}$$

$$\Leftrightarrow ^t\mathbf{C}_0.\mathbf{C}_0 = ^t\mathbf{U}.\mathbf{D}.\mathbf{U} + ^t\mathbf{U}.\mathbf{E}.\mathbf{U} \tag{21.44}$$

$$\Leftrightarrow ^t\mathbf{C}_0.\mathbf{C}_0 = \mathbf{1}_n + \mathbf{\Lambda} \tag{21.45}$$

avec $\mathbf{\Lambda} = \begin{bmatrix} \frac{1}{\lambda_1} & \dots & 0 \\ 0 & \dots & 0 \\ 0 & \dots & \frac{1}{\lambda_{k-1}} \end{bmatrix}$.

Les coordonnées sur les différents vecteurs propres ne sont pas corrélées.

La dispersion sur chaque vecteur propre vaut $1 + \lambda_i$ pour le i-ième vecteur propre. λ_i est la valeur propre de $\mathbf{D}^{-1}.\mathbf{E}$.

Cas particulier pour deux groupes Lorsqu'il existe deux groupes, il n'existe qu'un seul et unique vecteur discriminant, le vecteur propre de $\mathbf{D}^{-1}.\mathbf{E}$.

$$\mathbf{u} = \sqrt{\frac{n_1 n_2}{n\lambda}}.\mathbf{D}^{-1}(y_1 - y_2) \tag{21.46}$$

avec la valeur propre associée $\lambda = \frac{n_1 n_2}{n}[^t(y_1 - y_2).\mathbf{D}^{-1}.(y_1 - y_2)]$.

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

21.4.2 Comment classer ?

Si on effectue de nouvelles observations que l'on souhaite classer dans l'un des groupes existants uniquement à partir des valeurs mesurées, il existe deux approches :

1. une approche géométrique ;
2. une approche probabiliste.

L'approche géométrique consiste à calculer la distance définie par D entre la nouvelle observation et le centre de chacun des groupes. On classe la nouvelle observation dans le groupe pour lequel cette distance est **minimale**. La distance entre une observation x, un vecteur ligne avec p éléments, et un groupe i, s'écrit :

$$d^2\left(\mathbf{x}, \mathbf{y}_i\right) = {}^t\left(\mathbf{x} - \mathbf{y}_i\right).\mathbf{D}^{-1}.\left(\mathbf{x} - \mathbf{y}_i\right) \tag{21.47}$$

avec y_i le vecteur ligne avec p éléments des moyennes des p variables pour le groupe i. En développant le produit, on trouve :

$$d^2\left(\mathbf{x}, \mathbf{y}_i\right) = {}^t\mathbf{x}.\mathbf{D}^{-1}.\mathbf{x} - 2{}^t\mathbf{x}.\mathbf{D}^{-1}.\mathbf{y}_i + {}^t\mathbf{y}_i.\mathbf{D}^{-1}.\mathbf{y}_i \tag{21.48}$$

Le terme ${}^t\mathbf{x}.\mathbf{D}^{-1}.\mathbf{x}$ ne dépend pas du groupe i considéré. On veut classer dans le groupe pour lequel cette distance est minimale, mais on peut également classer $\mathbf{x}$ dans le groupe pour lequel $\mathbf{g}_i$ est maximal avec :

$$\mathbf{g}_i = (n-k)\left({}^t\mathbf{x}.\mathbf{D}^{-1}.\mathbf{y}_i - \frac{1}{2}{}^t\mathbf{y}_i.\mathbf{D}^{-1}.\mathbf{y}_i\right) \tag{21.49}$$

$$\Leftrightarrow \mathbf{g}_i = {}^t\mathbf{x}.\mathbf{D}^{*-1}.\mathbf{y}_i - \frac{1}{2}{}^t\mathbf{y}_i.\mathbf{D}^{*-1}.\mathbf{y}_i \tag{21.50}$$

Les $\mathbf{g}_i$ sont appelées **fonctions de classification** (ou fonctions linéaires discriminantes). Il en existe autant qu'il existe de groupe, et on affecte la nouvelle observation au groupe pour lequel sa fonction de classification est maximale. Le facteur $n-k$ est introduit afin de pouvoir utiliser $\mathbf{D}^*$ au lieu de $\mathbf{D}$, $\mathbf{D}^*$ étant la matrice de covariances nécessaires pour pouvoir calculer les probabilités d'appartenance à chaque groupe.

N.B. Dans le cas où il existe deux groupes, l'observation est classé dans le groupe dont le centre se projette du même côté par rapport au point milieu séparant les deux groupes.

L'approche probabiliste consiste à 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, il n'est possible de calculer ces probabilités que si les observations proviennent d'une loi **multinormale**. Si tel n'est

pas le cas, il faut transformer les données pour s'en rapprocher le plus possible. L'analyse discriminante est très robuste face à l'hypothèse de multinormalité dans la pratique. La fonction de densité s'écrit :

$$f(x) = (2\pi)^{-\frac{p}{2}} . |\mathbf{\Sigma}|^{-\frac{1}{2}} . \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{y})^1 . \mathbf{\Sigma}^{-1} . (\mathbf{x}-\mathbf{y})\right) \tag{21.51}$$

avec $\mathbf{\Sigma}$ la matrice de covariance et $|\mathbf{\Sigma}|$ le déterminant de la matrice de covariance. Si x provient du groupe i, alors sa fonction de densité est estimée par $\mathcal{N}(\mathbf{g}_i, \mathbf{D^*}_i)$. Si l'observation appartient nécessairement à un des k groupes, et si l'on suppose qu'*a priori* chaque groupe a une probabilité égale d'être observé, la probabilité conditionnelle vaut :

$$\Pr(\text{groupe } i|\mathbf{x}) = \frac{f_i(\mathbf{x})}{\sum_{j=1}^{k} f_j(\mathbf{x})} \tag{21.52}$$

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

$$\Pr(\text{groupe } i|\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}t(\mathbf{x}-\mathbf{y}_i) . \mathbf{D^*}^{-1} . (\mathbf{x}-\mathbf{y}_i)\right)}{\sum_{j=1}^{k} \exp\left(-\frac{1}{2}t(\mathbf{x}-\mathbf{y}_j) . \mathbf{D^*}^{-1} . (\mathbf{x}-\mathbf{y}_j)\right)} \tag{21.53}$$

Après quelques manipulations, cette expression peut s'écrire :

$$\Pr(\text{groupe } i|\mathbf{x}) = \left[\sum_{j=1}^{k} \exp(\mathbf{g}_j - \mathbf{g}_i)\right]^{-1} \tag{21.54}$$

avec $\mathbf{g}_i$ les fonctions de classification. Cette probabilité est maximale lorsque $\mathbf{g}_i$ est maximale, c'est-à-dire lorsque la distance d'un point au centre du groupe est minimale.

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

N.B. Il est possible de déterminer une **discrimination non linéaire** (ou quadratique).

21.4.3 Comment évaluer la qualité de l'analyse discriminante ?

La **qualité de la discrimination** est liée à la superposition des distributeurs de projections sur l'axe. On peut mesurer la qualité de la dispersion à la grandeur du rapport de la variance entre les moyennes à la variance à l'intérieur d'un groupe.

$$\frac{\text{variance inter-groupe}}{\text{variance intra-groupe}} \tag{21.55}$$

Ce rapport est analogue au F de l'analyse de variance. On suppose que la variance des scores à l'intérieur de chaque groupe répond au critère d'homogénéité de telle sorte que cette variance intra-classe est la moyenne des variances intra-classes des groupes considérés.

Il existe d'autres manières de vérifier la qualité de l'analyse discriminante :

1. le pourcentage de bien classés ;
2. la statistique λ de Wills ;
3. la statistique V de Rao ;
4. la corrélation canonique (ou le pouvoir discriminant d'un vecteur propre).

Le pourcentage de bien classés

Le pourcentage de bien classés est la **statistique la plus utilisée**. La procédure de classement étant établie, il s'agit de l'appliquer aux observations dont on connaît le véritable groupe, et de vérifier si la procédure produit le bon classement. Si un classement s'effectue entièrement de manière aléatoire, on obtiendra 50 % de bien classés. Le tableau de classement est appelée **matrice de confusion** (Tab. 21.1). Il s'agit d'une forme de tableau de contingence. On peut tester le caractère significatif du classement à l'aide d'un test d'indépendance du χ^2.

		AFFECTATION		
		Groupe 1	Groupe2	
OBSERVATION	Groupe1	a	b	N_1
	Groupe 2	c	d	N_2

TABLE 21.1 – Matrice de confusion

Les tests par ligne sont :

$$t_1 = \frac{a}{N_1} \tag{21.56}$$

et

$$t_2 = \frac{d}{N_2} \tag{21.57}$$

Le test sur l'ensemble de la matrice est :

$$t = \frac{a+d}{N_1+N_2} \tag{21.58}$$

Il est à noter que a et d sont identiques.

La statistique λ de Wilks

La statistique λ de Wilks est définie comme étant le rapport des discriminants des matrices $\mathbf{D}$ et $\mathbf{T}$.

$$L = \frac{\det \mathbf{D}}{\det (\mathbf{T})} = \det \left(\mathbf{T}^{-1}.\mathbf{D} \right) = \prod_{i=1}^{p} \gamma_i \tag{21.59}$$

avec γ_i la valeur propre de $\mathbf{T}^{-1}.\mathbf{D}$. La relation qui relie λ et γ est :

$$\gamma = \frac{1}{\lambda + 1} \tag{21.60}$$

Sous l'hypothèse de multinormalité et d'égalité des matières de covariances, on peut montrer que la quantité :

$$- \left(n - \frac{p+k}{2} - 1 \right) \ln (\mathbf{L}) \tag{21.61}$$

avec n le nombre total d'observations, p le nombre de variables, et k le nombre de groupes, est approximativement distribuée suivant une loi χ^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 statistique suivant :

$$\begin{cases} H_0 : \text{les vecteurs propres } q+1, q+2, \ldots, k-1 \text{ n'ajoutent rien à la discrimination des } k \text{ groupes.} \\ H_1 : \text{les vecteurs propres } q+1, q+2, \ldots, k-1 \text{ ajoutent quelque chose à la discrimination des } k \text{ groupe} \end{cases} \tag{21.62}$$

alors :

$$- \left(n - \frac{p+k}{2} - 1 \right) \ln (\mathbf{L}^*) \tag{21.63}$$

avec $\mathbf{L}^*$ donné par :

$$\prod_{i=q+1}^{k-1} \gamma_i \tag{21.64}$$

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

De plus, $(n-k)\,\lambda_q$ est approximativement distribué suivant une loi χ^2 avec $(p+k-2q)$ degrés de liberté. On vérifie si chaque valeur propre λ_i est significative.

La statistique V de Rao

La statistique V de Rao mesure la somme des distances entre les centres des groupes et la moyenne globale. La distance est normalisée par la matrice $\mathbf{D}^{-1}$. Elle est définie par :

$$V = \sum_{i=1}^{k} n_i{}^t (\mathbf{y}_i - \mathbf{y}) .\mathbf{D}^{*-1}. (\mathbf{y}_i - \mathbf{y}) \tag{21.65}$$

avec $\mathbf{y}_i$ le vecteur des moyennes du groupe i de dimension $p \times 1$, $\mathbf{y}$ le vecteur des moyennes, $\mathbf{D}^*$ la matrice de covariance intra-groupe $\frac{1}{n-k}\mathbf{D}$, k le nombre de groupes, $\mathbf{D}$ la matrice des produits croisés intra-groupes, n_i le nombre d'observations dans le groupe i, et n est le nombre total d'observations. Sous l'hypothèse de multinormalité et d'égalité des matrices de covariance, V est distribuée suivant une loi χ^2 avec $p-1$ degrés de liberté. Si on effectue la discrimination avec p variables, puis $p+1$ variables, on peut vérifier le caractère significatif de l'ajout de la variable. La changement de V en $V_{\text{final}} - V_{\text{initial}}$ est alors distribué suivant une loi χ^2 avec $k-1$ degrés de liberté.

La corrélation canonique

La corrélation canonique part du rapport α.

$$\alpha = \frac{{}^t\mathbf{u}.\mathbf{E}.\mathbf{u}}{{}^t\mathbf{u}.\mathbf{T}.\mathbf{u}} \tag{21.66}$$

On montre que α est valeur propre de $\mathbf{T}^{-1}.\mathbf{E}$ qui est reliée aux valeurs propres λ de $\mathbf{D}^{-1}.\mathbf{E}$ par :

$$\alpha = \frac{\lambda}{1+\lambda} \tag{21.67}$$

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

Solomon Kullback (1907-1994)

On peut également approximativement tester l'égalité des matrices de covariances intra-groupes avec le test de Solomon Kullback [Kullback, 1959]. Le test nécessite la multinormalité des observations :

$$\chi^2 = \sum_{i=1}^{k} \frac{n_i - 1}{2} \ln\left(\frac{\det(\mathbf{D}^*)}{\det(\mathbf{D}_i{}^*)}\right) \tag{21.68}$$

avec $\mathbf{D}^*$ la matrice de covariances intra-groupes, $\mathbf{D}_i{}^*$ la matrice de covariances pour le groupe i, n_i le nombre d'observations dans le groupe i, et n le nombre total

d'observations, est approximativement distribué suivant une loi χ^2 avec $\frac{1}{2}(k-1)n(n+1)$ degrés de liberté. On rejette l'hypothèse d'égalité des matrices de covariances lorsque la statistique excède le seuil lui dans une table χ^2.

21.4.4 Comment sélectionner les variables ?

Malgré la diversité des méthodes, la pratique montre que le sous-ensemble de variables retenues est relativement robuste au choix du critère d'inclusion.

Même si deux sous-ensembles diffèrent des variables retenues, très souvent l'interprétation est identique et les performances (le classement) très comparables.

Conclusion générale

L'A.F.D. s'opère souvent un complément de l'A.C.P. ou de l'A.C.M.

Bibliographie

[Kullback, 1959] KULLBACK, S. (1959). Information Theory and Statistics. John Wiley & Sons, New York.

[Mahalanobis, 1936] MAHALANOBIS, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute Science of India, 2(1):49–55.