Support de cours

ANALYSE DES DONNEES I

CHAPITRE 5 : L'ANALYSE FACTORIELLE DISCRIMINANTE

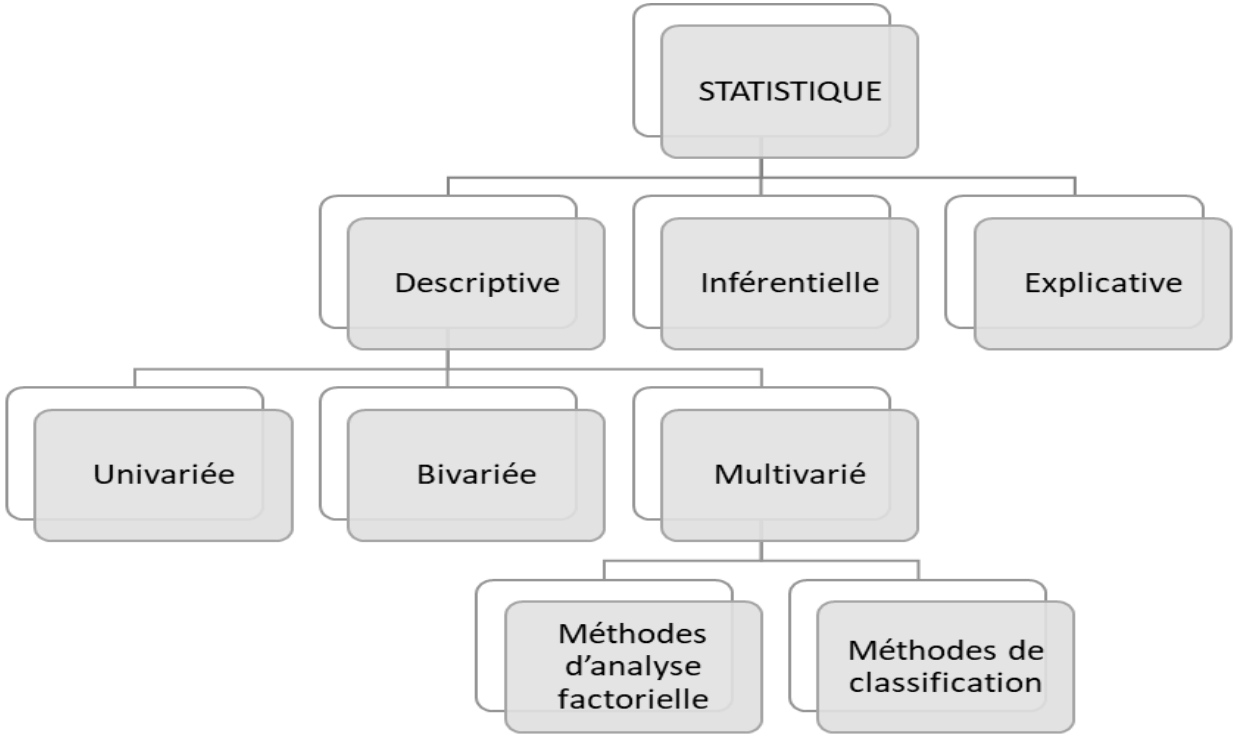
Iphygénie SARR
Ingénieure statisticienne économiste

Ecole Nationale de la Statistique et de l'Analyse Economique Pierre Ndiaye
Juin 2024

OBJECTIF DU CHAPITRE

Ce chapitre décrit l'Analyse factorielle discriminante (AFD). Il s'agit de trouver le moyen de « discriminer » au mieux plusieurs sous-ensembles d'une population décrite par une variable qualitative et plusieurs variables quantitatives.

Quelques rappels

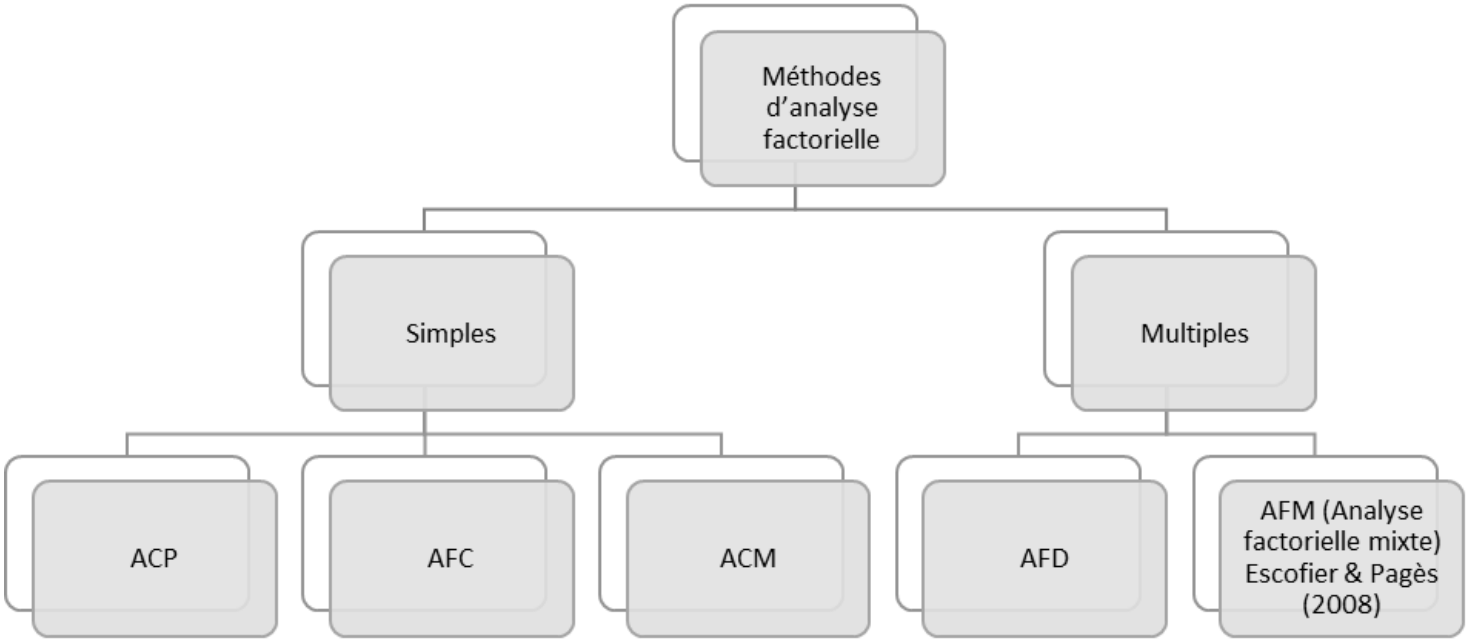


Quelques rappels

Les méthodes d'analyse factorielle répondent toutes à la double préoccupation suivante :

- résumer au mieux l'ensemble des variables à l'aide d'un petit nombre de facteurs non redondants qui matérialisent leur liaison ;
- photographier au mieux les disparités globales entre individus et cela en rapport si étroit que la photo des disparités individuelles est naturellement décrite à l'aide du résumé des variables.

Quelques rappels



Quelques rappels

- Les méthodes d'analyse factorielle simple permettent l'exploration globale d'un tableau de données décrivant un certain nombre d'individus statistiques avec diverses variables de même nature (toutes quantitatives ou toutes qualitatives).
- Les méthodes factorielles multiples permettent l'exploration simultanée de plusieurs tableaux décrivant un même ensemble d'individu, chaque tableau le faisant à l'aide d'un groupe de variables particuliers.

PLAN DU CHAPITRE

01. Approche intuitive de l'AFD

02. Données et notations

03. Analyse factorielle discriminante

04. Mesure d'efficacité d'une AFD

01.

Approche intuitive de l’AFD

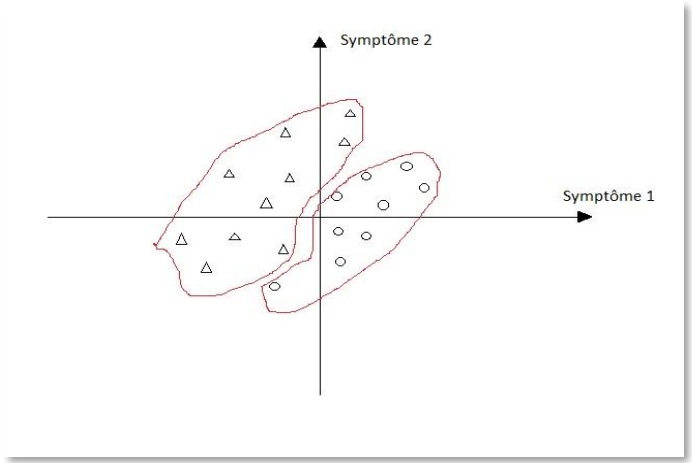
Un exemple

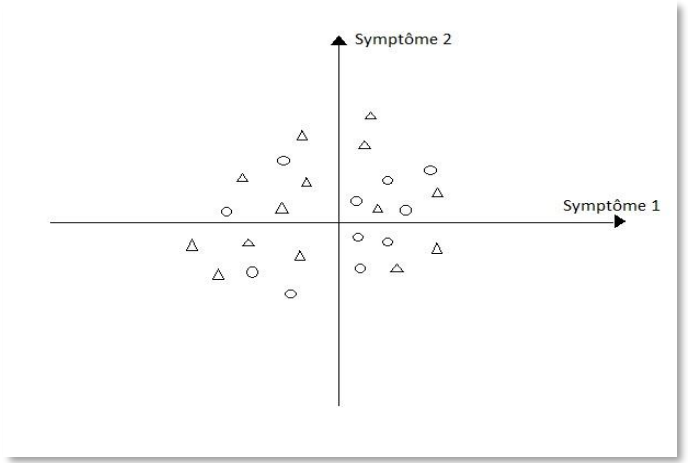
Pour introduire l'analyse discriminante, nous utiliserons un exemple. Supposons que l'on considère une population de personnes souffrant d'une maladie cardiaque.

- Dans un premier temps, on observe sur ces malades un ensemble de symptômes caractéristiques. Chaque malade peut être représenté par un point dans l'espace vectoriel défini par les symptômes.
- Il est ensuite examiné par un médecin qui pose sur son cas un diagnostic. On peut ainsi scinder le nuage des malades, repérés dans l'espace des symptômes, en plusieurs sous-nuages relatifs à un diagnostic.

Un exemple

Dans la première figure, nous avons représenté un cas où l'on observe deux symptômes x et y, et où les sous-nuages de diagnostic I et II sont bien distincts. La seconde figure décrit un cas où le diagnostic est pratiquement sans relation avec les symptômes.





Objectif

- C’est justement la relation entre diagnostic et symptômes que l’analyse discriminante se propose de mettre à jour.
- La démarche de l’analyse discriminante est ainsi tracée : quelle est la combinaison linéaire des symptômes qui permet de discriminer au mieux les diagnostics ?
- On ne cherche plus les directions qui nous permettent de mieux observer le nuage mais plutôt celles qui nous permettent d’observer une meilleure discrimination.
- L’objectif recherché est donc de trouver les directions qui permettent de séparer le mieux possible les catégories et de donner une représentation graphique qui rende au mieux compte de cette opération.

Plusieurs applications

D’autres exemples :

- cas de tumeurs décrite par diverses variables et classées en bénignes et malignes
- cas d’entreprises décrites par divers ratios d’analyse financière avec pour diagnostic le signe du résultat (positif ou négatif).
- cas de pays décrits par diverses variables macroéconomiques avec pour diagnostic le solde budgétaire (déficitaire ou excédentaire).

L’AFD devrait ainsi permettre de prévoir un risque de défaut de paiement, un risque financier, un risque médical…

Mais en plus …

Supposons maintenant que l'on considère un individu supplémentaire pour lequel on connaît les valeurs prises par les symptômes, mais non le diagnostic. *Quel diagnostic lui attribuer ?*

On observera comment l'individu se situe dans l'espace par rapport aux sous-nuages et on lui attribuera le diagnostic qui correspond au sous-nuage dont il est le plus proche.

Cette procédure permet de passer automatiquement des symptômes observés à un diagnostic, une fois cette méthode implémentée sur un nombre suffisant d'observations.

Mais attention …

- Certes l'analyse factorielle discriminante est une méthode de description et d'estimation extrêmement puissante ; le champ de ses applications possibles est vaste.
- Mais attention, mais elle est essentiellement descriptive son aspect prédictif ne se soucie pas de certaines considérations économétriques et ne saurait suffire pour faire de la prévision.

02.

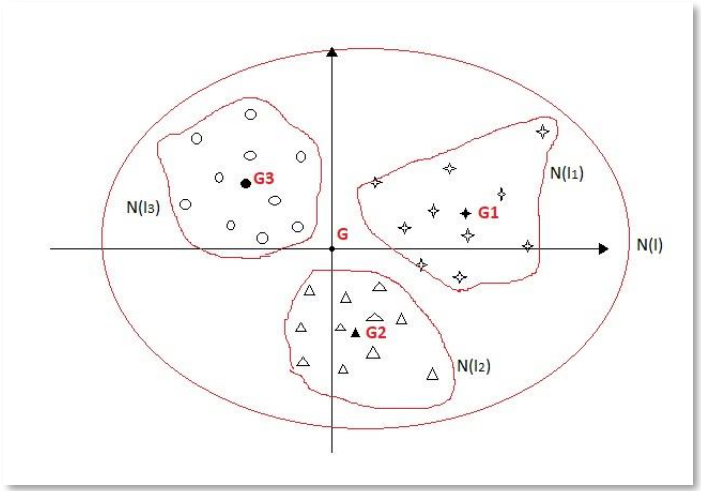
Données et notations

Données et notations

- Nous conserverons dans la suite de ce chapitre les termes de « diagnostics » et de « symptômes » pour désigner d'une part les modalités de la variable qualitative que l'on cherche à discriminer, d'autre part les variables à l'aide desquelles on opère cette discrimination.
- Considérons une population de N individus sur lesquels nous avons mesuré K variables quantitatives centrées et réduite. Individus et variables sont repérés respectivement par les indices i (i=1,…,N) et j(j=1,…,K).
- L'ensemble des individus peut évidemment être représenté par un nuage $\mathcal{N}(I)$ de N points X^i dans $\mathbb{R}^K$, en repérant chaque individu par les valeurs des variables et en associant à chaque point x_{ij} une masse unité. Le centre de gravité de ce nuage est à l'origine de $\mathbb{R}^K$.

Données et notations

Supposons maintenant que nous considérons une variable qualitative Y, telle qu'à chaque individu corresponde une modalité de Y et une seule. Nous noterons m le nombre de modalités de Y, et nous les repérerons par un indice k (k=1,...,m).
Nous noterons n_k le nombre des individus ayant la modalité k de Y.



Données et notations

Y détermine dans l'ensemble des individus I une partition. Notons $\mathcal{N}(I_k)$ le nuage des individus ayant la modalité k de Y et G_k son centre de gravité.

On a :

$$g_{jk} = \frac{1}{n_k} \sum_{i \epsilon I_k} x_{ij}$$

$$\mathcal{N}(I) = \bigcup_{k=1}^{m} \mathcal{N}(I_k)$$

03.

Analyse factorielle discriminante

Le problème

D'après le théorème de Huygens, l'inertie de $\mathcal{N}(I)$ peut s'écrire :

$$In_G(I) = In_G(G) + \sum_{k=1}^{m} In_{G_k}(I_k)$$

- $In_G(G)$ est l'inertie entre les classes : l'inertie du nuage $\mathcal{N}(G)$
- $\sum_{k=1}^{m} In_{G_k}(I_k)$ est l'inertie interne aux classes : l'inertie du nuage $\mathcal{N}'(I) = \bigcup_{k=1}^{m} \mathcal{N}(I_k)$.

Le problème

Notons T, E et D les matrices d'inertie des nuages de $\mathcal{N}(I)$, $\mathcal{N}(G)$ et $\mathcal{N}'(I)$.

Cette notation s'explique par la formule mnémotechnique suivante : inertie **T**otale = inertie **E**ntre les classes + inertie **D**ans les classes).

Soit U un vecteur de $\mathbb{R}^K$, l'inertie expliquée par la direction U s'écrit comme suit : U'TU=U'EU+U'DU

Cela implique que :

$$1 = \frac{U'EU}{U'TU} + \frac{U'DU}{U'TU}$$

Le problème

Le problème d'analyse discriminante se formule ainsi :

- Trouver une direction U dans $\mathbb{R}^K$ telle que $\frac{U'EU}{U'TU}$ soit maximal. Il s'agit de la direction dont l'inertie expliquée provient au maximum de la dispersion entre les classes. Il est clair qu'il s'agit aussi de la direction pour laquelle l'inertie expliquée rend minimale la dispersion entre les classes.
- Ainsi on cherche un angle d'observation du nuage qui discrimine au mieux les points suivant la variable Y cela en permettant simultanément une grande inertie entre les classes et une faible inertie au sein des classes.

La solution

- Rappels sur la maximisation du quotient de deux formes quadratiques.

Soient A et B deux matrices symétriques de même dimension, B étant supposée inversible. Alors, le rapport $\frac{U'AU}{U'BU}$ est maximal pour $U = U^1$, U^1 étant le vecteur propre de $B^{-1}A$ associé à la plus grande valeur propre λ^1 de cette matrice.

- On déduit du rappel ci-dessus que, pour notre problème d'optimisation, la direction U solution est le vecteur propre U^1 de $T^{-1}E$ associé à la plus grande valeur propre λ^1.

- La matrice $T^{-1}E$ est de rang min (m-1 , K-1). On pourra donc rechercher au plus m-1 directions discriminantes.

Remarque (1)

- Comme en régression et seulement lorsque les symptômes ont été préalablement réduits, plus un coefficient est fort en valeurs absolus plus le symptôme correspondant intervient dans le facteur positivement ou négativement toutes les autres étant par ailleurs considérés comme fixées.
- Et là se pose le même (gros) problème qu'en régression, l'interprétabilité des coefficients (même au seul niveau de leur signe est quasiment impossible en cas de redondances (corrélations) entre les symptômes (hypothèse d'absence de multi colinéarité est violée).

Remarque (2)

Supposons que les variables à partir desquelles nous cherchons à discriminer la variable qualitative Y soient elles-mêmes qualitatives, et non plus quantitatives comme nous l'avons supposé jusqu'ici.

On procédera comme suit : d'abord on réalise une analyse des correspondances multiples sur le tableau logique donnant le codage disjonctif complet des modalités des variables qualitatives observées (et ne comportant pas la variable Y).

Cette analyse nous permet de résumer ces variables qualitatives en un petit nombre de variables quantitatives, les facteurs $F_\alpha(i)$ de l'analyse des correspondances multiples.

Remarque (2)

A cause des deux (02) précédentes remarques, que les variables « symptômes » soient toutes quantitatives ou qualitatives, on préférera toujours réaliser une AFD sur les facteurs. Ce qui implique de faire au préalable une ACP ou une ACM.

04.

Mesure d’efficacité d’une AFD

Affectation

Supposons que nous considérons un individu supplémentaire pour lequel nous connaissons les valeurs des variables quantitatives j mais non la modalité du caractère Y. Associons à cet individus le point X de $\mathbb{R}^K$ de coordonnée courante x_j.

Dès l'instant où l'on pense que le diagnostic Y est explicable à partir des symptômes, on peut poser le problème de sa prévision. Il s'agit d'affecter le nouvel individu dont on connait le symptôme à la classe de diagnostic la plus vraisemblable empiriquement.

Affectation

Il existe plusieurs règles d'affectation. Citons parmi les plus utilisées.

1. Une règle simple et géométrique d'affectation est de choisir la classe dont le centre de gravité est le plus proche du nouvel individu. La métrique généralement utilisée dans les applications les plus courantes est celles de :

- Mahalanobis globale : $\min_{k=1 \text{ à } m} \|X - G_k\|^2_{D^{-1}}$
- Mahalanobis locale : $\min_{k=1 \text{ à } m} \|X - G_k\|^2_{D_k^{-1}}$

Affectation

2. Le scoring

La première et seule variable discriminante peut être utilisée comme score pour construire une règle de décision simple en fixant un ou deux seuils. On peut alors définir la règle de décision suivante pour un nouvel individu i :

si $F_\alpha(i) \geq c$ l'individu i est affecté au groupe 1,

si $F_\alpha(i) < c$ l'individu i est affecté au groupe 2.

Affectation

3. Le modèle bayésien d'affectation qui permet de prendre en compte les probabilités à priori des différentes classes qui peuvent être très inégales dans certaines applications.
4. Les méthodes d'estimations non paramétriques de la densité connu sous le nom de méthodes des noyaux de Rosenblatt ou de Parzen.
5. Les méthodes d'affectation non paramétrique utilisant les m plus proches voisins.

Voir Lebart, Morineau & Piron (1995) pour plus de détails sur ces trois (03) méthodes.

Qualité de l’affectation

Il existe un cadre qui permet de tester cette affectation, et par la même occasion la qualité de l’AFD réalisée.

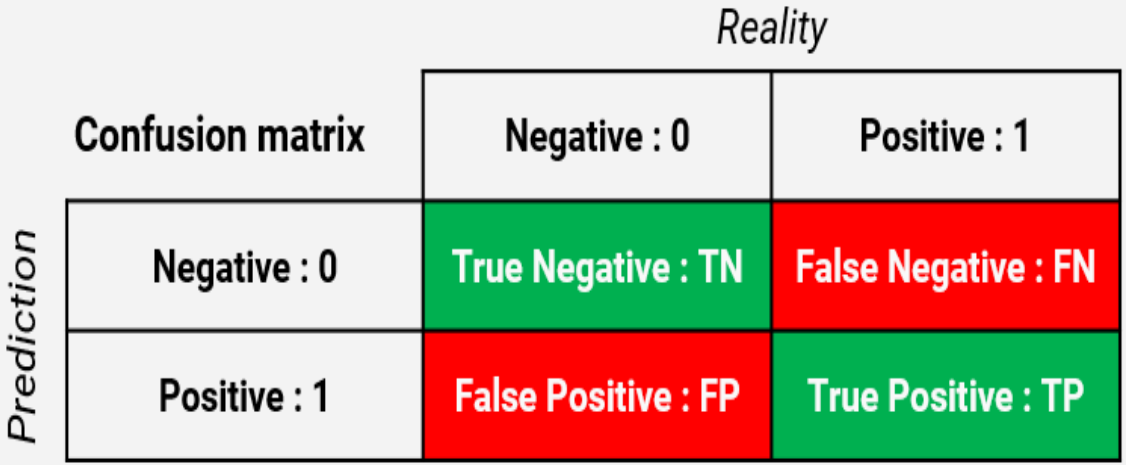
1. Tout d’abord, on effectue la discrimination sur une partie seulement de la population (l’échantillon d’apprentissage) et on teste les règles de discrimination sur la partie restante (échantillon test).

Qualité de l'affectation

2. Ensuite, on peut calculer un pourcentage de bien classé sur l'échantillon d'apprentissage ce qui donnera une idée optimiste de la qualité de discrimination. C'est-à-dire la proportion des individus de la classe k qui sont situés du bon côté des axes discriminants qui isolent G_k.
3. Enfin, on peut apprécier a priori la qualité du diagnostic ainsi posé sur les individus supplémentaires le pourcentage de « bien classés » à l'aide de la matrice de confusion de l'échantillon test.

Qualité de l'affectation

Exemple de matrice de confusion



Sélection de l’échantillon test

Il existe plusieurs manières de sélectionner un échantillon test.

1. La méthode la plus courante est celle de l’échantillon-test. Elle consiste à tirer au hasard 80% des individus de la population pour constituer l’échantillon d’apprentissage et de tester les règles de discrimination sur les 20% non utilisés.

Sélection de l’échantillon test

2. On peut aussi appliquer la méthode redistribution, c’est-à-dire faire l’affectation sur les individus qui ont permis d’obtenir les coefficients des axes factoriels (donc dans ce cas il n’y a pas d’échantillon, le test est fait sur le groupe globale).

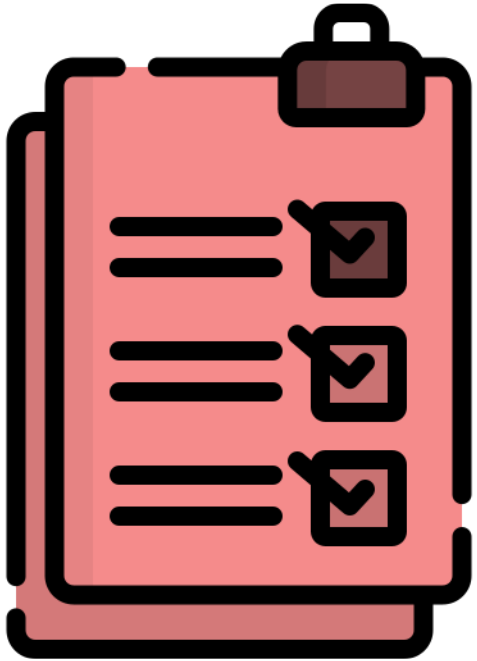
Sélection de l’échantillon test

3. Ou alors, utiliser la méthode de la validation croisée (booststrap). L’implémentation de celle-ci repose sur la division de la population totale en plusieurs échantillons, ensuite effectuer le classement autant de fois qu’il y a d’échantillons en considérant que l’échantillon test est un des échantillons et l’échantillon d’apprentissage est l’ensemble des autres échantillons. A la fin, le Taux d’Erreur de Classement avec cette méthode correspond à la moyenne arithmétique des Taux d’Erreur de Classement à chaque étape.

SYNTHESE

Quelques points essentiels à retenir :

- l'analyse discriminante cherche à résumer la liaison entre une variable qualitative et plusieurs variables quantitatives ;
- dans l'espace des symptômes, on cherche donc la direction ou le sous-espace sur lequel la projection des individus sépare le mieux possible les classes de diagnostics ;
- i.e. tel que le critère tel que l'inertie entre classe/l'inertie totale soit maximum ;
- Ce processus se fait en trois (03) étapes ;



SYNTHESE

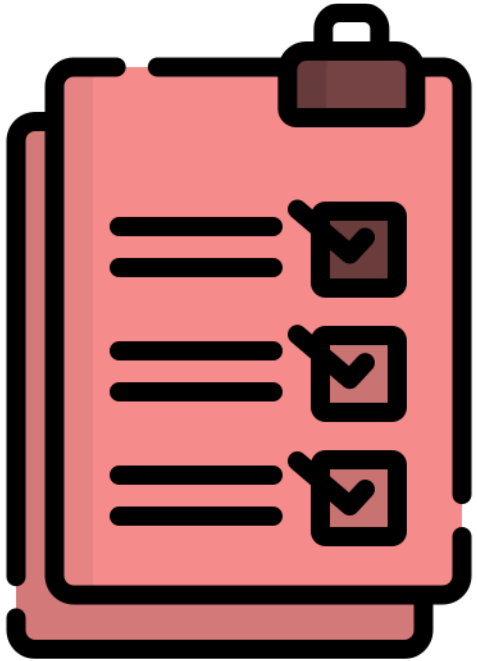
Quelques points essentiels à retenir :

1. Estimation des fonctions discriminantes : Lorsque vous effectuez une AFD, le modèle identifie des fonctions discriminantes qui maximisent la séparation entre les groupes. Ces fonctions sont des combinaisons linéaires des variables initiales. Pour chaque individu, ces fonctions sont évaluées pour obtenir les scores (facteurs).

SYNTHESE

Quelques points essentiels à retenir :

2. Calcul des scores : Les scores sont obtenus en évaluant les fonctions discriminantes pour chaque individu à partir de ses valeurs de variables originales. Les scores d'un individu correspondent des valeurs numériques qui représentent la position de cet individu dans l'espace factoriel défini par les axes discriminants (ses coordonnées sur les axes discriminants).



SYNTHESE

Quelques points essentiels à retenir :

3. Classification des individus : En fonction des scores calculés, les individus peuvent être classifiés dans les groupes ou catégories définis par l'analyse discriminante.

FIN DU CHAPITRE

MERCI POUR VOTRE ATTENTION !