**TER : Scoring**

**II- Analyse discriminante**

1. Théorie de la méthode probabiliste

*Introduction de remplacement :*

Le but premier de cette méthode est de prédire au mieux les valeurs d’une variable Y qualitative à K modalités, à partir de p variables explicatives X = (X1 , . . . , Xp ) quantitatives.

Dans cette section, nous allons définir des règles de décision bayésiennes qui vont permettre d’affecter un nouvel individu à la classe “la plus probable” et non pas au groupe « le plus proche » comme c’est le cas pour l’analyse discriminante géométrique. Pour cela, il est nécessaire de faire des hypothèses probabilistes sur les données, d’où le nom de la méthode.

* Cf annotations sur le poly de Mme Chavent

Section 5.4 à rajouter sur le lambda de wilks.

**5.4 Sélection des variables d’entrée et qualité du modèle**

5.4.1 Qualité du modèle

L’analyse factorielle discriminante probabiliste est donc utilisée pour affecter des scores à des individus. Mais donner un score à un individu sans contexte d’étude est assez absurde. C’est pourquoi, il faut évaluer le modèle choisi pour savoir si celui-ci permet une bonne prédiction des groupes d’appartenance.

Il est coutume de tester la qualité du modèle grâce aux matrices de confusions qui donnent le taux de bon et mauvais classement des individus dans chaque groupe. *(je ne m’attarde pas sur ça ici car tu en parles déjà dans la régression logistique, avec les TP et FP, peut-être parler ici de la validation croisée si ça n’a pas déjà été fait… (à moins que l’on mette la partie sur l’afd en premier)*

Nous construisons donc souvent plusieurs modèles et évaluons leurs capacités prédictives grâce à ces matrices. Or si l’on a un très grand nombre de variables explicatives, cela peut être extrêmement fastidieux et coûteux de tester tous les modèles possibles et imaginables. C’est pourquoi il est coutume d’utiliser des outils de pré-sélection de variables les plus discriminantes.

5.4.2 Sélection des variables

Le Lambda de Wilks est souvent utilisé dans les logiciels comme critère pour ne garder que les variables apportant de l’information sur l’appartenance ou non d’un individu à un groupe.

Le Lambda de Wilks est une approche paramétrique permettant de tester si plusieurs variables continues distinctes  sont liées à une variable qualitative Y à K ≥ 2 groupes, lorsqu’elles sont considérées avec leurs différentes interactions multivariées.

Les hypothèses d’utilisation de ce test sont:  suivent une loi normale et leur matrice de covariance respective sont égales (homoscédasticité).

La statistique du test du Lambda de Wilks se définie de la manière suivante :

Où SCR est la matrice de variance-covariance intragroupe et SCT la matrice de variance-covariance globale.

Cette statistique de test suit une loi de Wilks à  degrés de liberté et l’hypothèse  est : « Indépendance entre  et  ».

Une variable a un bon pouvoir discriminant si la dispersion intra-groupe est faible et si la dispersion intergroupe est forte. Donc plus le Lambda de Wilks sera faible, plus la variable considérée est discriminante.

C’est ce critère qu’utilise la commande « greedy.wilks » de Rstudio que nous avons utilisée pour trouver les variables les plus discriminantes dans notre jeu de données et ainsi se focaliser sur un nombre de modèles plus réduit.

1. ACP préliminaire (présentation du jeu de données)
2. Exemple farms

Code sur git. Peut-être organisé la rédaction en suivant la méthode de construction d’un score de Mme Chavent ? (Pages 1 et 2)