Analyse des données

Introduction

L'analyse des données est une des branches les plus vivantes de la statistique. Ses principales méthodes se séparent en deux groupes:

- Les méthodes de classification,
- Les méthodes factorielles.

Les méthodes de classification visant à réduire la taille de l'ensemble des individus en formant des groupes homogènes.

Les méthodes factorielles cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l'algèbre linéaire et donnant lieu à des représentations graphiques dans lesquelles les objets à décrire se transforment en des points sur des axes et des plans.

Les principales techniques factorielles sont :

L'analyse en composantes principales (Hotelling, 1933) qui analyse un ensemble de données (observations) faites sur un ensemble de variables quantitatives (numériques).

L'analyse des correspondances (Benzekri, 1964) qui est une technique de base pour analyser des tables de contingence qui peut être utilisé pour des variables qualitatives ou quantitatives positives de nature très divers.

L'analyse canonique.(Hotelling) qui contient à la Régression multiple et l'analyse discriminante comme des cas particulier.

Les techniques factorielles de l'analyse des données ont une partie de fondement générale commune à toutes : c'est celle qui s'appelle l' «Analyse générale», qui est basée sur les idées développées jadis par Eckart et Young (1936), qu'aujourd'hui elles sont développées encore plus théoriquement, surtout de point du vue informatique dans les dernières années et elles construisent ce qu'on appelle « Approximation d'une matrice par d'autres de rang inférieur », qui est basée sur la théorie générale de décomposition singulières d'une matrice (Singular Value Descomposition (SVD)).

Plan du cours :

- Analyse en composantes principales (ACP)
- Analyse factorielle des correspondances (AFC)
- Analyse canonique
- Analyse des correspondances
- Analyse discriminante.

Bibliographie:

- El Marhoum, A.(2005): «Analyse des données ». Toubkal.
- Labrousse, C. (1976): «Introduction à l'éconmétrie ». Dunod.
- Saporta, G. (1990): «Probabilités, Analyse des données et Statistique ». TECHNIP.

ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

Introduction :

L'analyse en composantes principales (Hotelling, 1933) est une méthode de l'analyse des données qui a pour objectif de réduire le nombre de données, souvent très élevé, d'un tableau de données représenté, algébriquement, comme une matrice et, géométriquement comme un nuage de points.

L'analyse en composantes principales consiste en l'étude des projections des points de ce nuage sur un axe (axe factoriel ou principal), un plan ou un hyperplan judicieusement déterminé. Mathématiquement, on obtiendrait le meilleur ajustement du nuage par des sous-espaces vectoriels. Algébriquement, il s'agit de chercher les valeurs propres maximales de la matrice des données et par conséquent ses vecteurs propres associés qui représenteront ces sous-espaces vectoriels (axes factoriels ou principales).

Lors de la projection, le nuage peut être déformé est donc serait différent de réel, alors les méthodes d'ajustement consistent en minimiser cette possible déformation et ce en maximisant les distances projetées.