Introduction à la Statistique exploratoire multidimensionnelle

Résumé

Cette vignette fait suite à celles, plus élémentaires, de Statistique descriptive unidimensionnelle, bidimensionnelle et multidimensionnelle pour aborder les principales méthodes factorielles de réduction de dimension et de représentation optimale ainsi que celles de classification non supervisée.

Plan du cours :

1 Historique

Les bases théoriques de ces méthodes sont anciennes et sont principalement issues de "psychomètres" américains : Spearman (1904) et Thurstone (1931, 1947) pour l'Analyse en Facteurs, Hotteling (1935) pour l'Analyse en Composantes Principales et l'Analyse Canonique, Hirschfeld (1935) et Guttman (1941, 1959) pour l'Analyse des Correspondances. Pratiquement, leur emploi ne s'est généralisé qu'avec la diffusion des moyens de calcul dans le courant des années 60. Sous l'appellation *"Multivariate Analysis"* elles poursuivent des objectifs sensiblement différents à ceux qui apparaîtront en France. Un individu ou unité statistique n'y est souvent considéré que pour l'information qu'il apporte sur la connaissance des liaisons entre variables au sein d'un échantillon statistique dont la distribution est le plus souvent soumise à des hypothèses de normalité.

En France, l'expression *"Analyse des Données"* recouvre les techniques ayant pour objectif la *description statistique des grands tableaux* (n lignes, où n varie de quelques dizaines à quelques milliers, p colonnes, où p varie de quelques unités à quelques dizaines). Ces méthodes se caractérisent par une utilisation *intensive* de l'ordinateur, leur objectif *exploratoire* et une absence quasi systématique d'hypothèses de nature *probabiliste* au profit des propriétés et résultats de géométrie euclidienne. Elles insistent sur les représentations graphiques en particulier de celles des individus qui sont considérés au même titre que les variables.

Depuis la fin des années 1970, de nombreux travaux ont permis de rapprocher ou concilier les deux points de vue en introduisant, dans des espaces multidimensionnels appropriés, les outils probabilistes et la notion de *modèle*, usuelle en statistique *inférentielle*. Les techniques se sont ainsi enrichies de notions telles que l'estimation, la convergence, la stabilité des résultats, le choix de critères...

L'objectif essentiel de ces méthodes est l'aide à la compréhension de volumes de données souvent considérables. Réduction de dimension, représentation graphique optimale, recherche de facteurs ou variables latentes... sont des formulations équivalentes.

2 Méthodes

Les méthodes de *Statistique exploratoire multidimensionnelle* se classifient selon leur objectif (réduction de dimension ou classification) et le type des données à analyser (quantitatives et/ou qualitatives) :

- Description et réduction de dimension (méthodes factorielles) :
  1. Analyse en Composantes Principales (p variables quantitatives),
  2. Analyse Factorielle Discriminante (p variables quantitatives, 1 variable qualitative),
  3. Analyse Factorielle des Correspondances simple (2 variables qualitatives) et Multiple (p variables qualitatives),
  4. Analyse Canonique (p et q variables quantitatives),

5. "Multidimensional Scaling" (M.D.S.) ou positionnement multidimensionnel ou analyse factorielle d'un tableau de distances.

Toutes ces méthodes sont basées sur des outils classiques de géométrie euclidienne qui sont développés dans les rappels et compléments d'algèbre linéaire.

- Méthodes de classification :
  1. Classification ascendante hiérarchique,
  2. Algorithmes de réallocation dynamique,
  3. Cartes de Kohonen (réseaux de neurones).

Les références introductives les plus utiles pour ce cours sont : Bouroche & Saporta (1980)[2], Jobson (1992)[3], Droesbeke, Fichet & Tassi (1992)[1], Mardia, Kent & Bibby (1979)[5], Saporta (2006)[6], Lebart, Morineau & Piron (1995)[4].

Références

[1] P.C. Besse et A. Pousse, *Extension des analyses factorielles*, Modèles pour l'Analyse des Données Multidimensionnelles (J.J. Droesbeke et al., réds.), Economica, 1992, p. 129–158.

[2] J.M. Bouroche et G. Saporta, *L'Analyse des Données*, Que Sais-je, PUF, 1980.

[3] J.D. Jobson, *Applied Multivariate Data Analysis*, t. II : Categorical and multivariate methods, Springer-Verlag, 1992.

[4] L. Lebart, A. Morineau et M. Piron, *Statistique exploratoire multidimensionnelle*, Dunod, 1995.

[5] K.V. Mardia, J.T. Kent et J.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

[6] G. Saporta, *Probabilités, Analyse des Données et Statistique*, deuxième éd., Technip, 2006.