Modélisation et extraction des liens complexes entre variables. Application à des données socio-économiques.

Martine Cadot*, Dhouha El Haj Ali**

* Université Henri Poincaré / LORIA, Nancy, France
martine.cadot@loria.fr
http://www.loria.fr/~cadot
** Université Manar I, Faculté des sciences économiques et de gestion de Tunis, Tunisie
elhajali.dhouha@yahoo.fr

Résumé. Nous nous intéressons ici à un type particulier de complexité qui est celle des liaisons entre variables. Il existe des modèles statistiques qui ont été construits pour traiter certains aspects de cette complexité. Ainsi le *modèle linéaire général* (Azaïs et Bardet 2005) permet de rendre compte d'aspects spécifiques de la complexité comme les interactions d'ordre quelconque, les liaisons négatives au même titre que les positives, et les « contrastes ». Mais ces méthodes sont mal adaptées au cas d'un grand nombre de variables et elles exigent une explicitation a priori des liaisons en jeu. Nous présentons notre méthode MIDOVA qui extrait directement des données le même type de liaisons que le modèle linéaire général, sans nécessiter d'hypothèses contraignantes, tout en étant compatible avec un grand nombre de variables, pour l'instant qualitatives. Nous l'illustrons en l'appliquant à des données issues de l'enquête PAPFEM, réalisée en 2001 par l'Office National de la Famille et de la Population en Tunisie, et nous mettons au jour le lien particulièrement complexe entre la pauvreté du ménage et la situation socio-économique des deux conjoints.

1 Introduction

Notre but est d'extraire des données ce que nous appelons les *liaisons complexes* entre variables, par opposition aux liaisons simples, c'est-à-dire entre les variables prises deux par deux . Les données que nous considérons se présentent sous la forme de tableaux individus X variables, c'est-à-dire contenant pour chaque individu sa valeur pour chaque variable. Certains modèles statistiques permettent une représentation des liaisons complexes entre variables. Nous les décrivons dans la section suivante en nous intéressant plus particulièrement au modèle statistique le plus utilisé par les chercheurs en sciences humaines, le modèle linéaire général[1], qui se base sur une décomposition des liaisons complexes en effets simples, interactions, contrastes. Nous décrivons également les conditions d'application de ces modèles statistiques qui les rendent inopérants pour ce que nous souhaitons faire : extraire automati-

[1] Pour Azaïs et Bardet (2005), *le modèle linéaire général*, ou plus *le modèle linéaire*, exprime la variable à expliquer comme combinaison linéaire des *paramètres du modèle*, et non des variables explicatives. Selon leur définition, que nous adoptons, l'équation de régression polynomiale $Y=aX^2+bX+c+\varepsilon$ fait partie du modèle linéaire car elle est linéaire en les paramètres inconnus a, b et c.