Representing interaction in multiway contingency tables: MIDOVA, CA and log-linear model

Martine Cadot, Alain Lelu

Representing interaction in multiway contingency tables: MIDOVA, CA and log-linear model.

Martine Cadot[1,2,*] , Alain Lelu[2,3,4]

1. Université Henri Poincaré, Nancy1
2. Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA, Nancy)
3. Université de Franche-Comté/LASELDI, Besançon
4. Institut des Sciences de la Communication du CNRS, Paris
*   Martine.cadot@loria.fr

Keywords: Interaction, itemsets, loglinear model, N-way contingency table, categorical data

Correspondence Analysis (CA) is particularly suited to categorical variables, as long as 2-way contingency tables are concerned. (Mourad 1983) has pointed out that its extension to 3-way contingency tables is far from trivial, due to interaction effects between the variables. (Escofier 1983) has provided a non-symmetric solution to this problem, through the example of a 3-way *qualification×profession×gender* table, disregarding interaction in this first approach. Then (Escofier & Pagès 1988) took interaction into account, still in a non-symmetric scheme, and illustrated with the same example, and in (Abdessemed & Escofier 2000) this CA approach was contrasted with the log-linear model one.

Beside CA and log-linear model, issued from the statistics domain, other research streams originating in Artificial Intelligence have coped with the same problem: we will present here the extension to categorical variables of our results on extracting and statistically validating « itemsets » in boolean datatables, results first published in (Cadot 2006) – for a survey on itemset approaches, see (Han 2001). We coined MIDOVA (Multidimensional Interaction Differential of Variation) our method for highlighting and representing complex links between qualitative variables, which includes interaction, well-suited to socio-economic data   (Haj Ali & Cadot 2010). We will compare it to the CA and log-linear model approaches, using the same 3-way example as Escofier and her colleagues. We will show that out method is effective for general N-way interactions (N may be far greater than 3), whether symmetrically or not, and results both in easy and detailed interpretability, as CA does, and in statistical significance testing, as the log-linear model does in the case of few variables.

References

Abdessemed, L. & Escofier B. (2000). Analyse de l'interaction et de la variabilité inter et intra dans un tableau de fréquence ternaire. In Moreau, J., Doudin, P.-A. & Cazes, P. (eds). *L'analyse des correspondances et les techniques connexes*, 146-164. Springer-Verlag, Berlin.

Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Ph.D. thesis, Université de Franche-Comté, France.

Escofier, B. (1983). *Généralisation de l'analyse des correspondances à la comparaison de tableaux de fréquences*. Rapport de Recherche Inria, Rennes, N°207.

Escofier, B. & Pagès, J. (1988). *Analyses factorielles simples et multiples.* Dunod Paris.

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco.

Haj Ali, D. & Cadot, M. (2010). *Estimation de l'impact de la décision du mariage sur la pauvreté des ménages tunisiens*, MASHS 2010 (Lille, France), 10–11 juin, pp. 45–56

Mourad, G. (1983). Flux de pétrole et flux de marchandises entre l'OPEP et l'OCDE de 1970 à 1979. In Benzécri J.-P. & collaborateurs (eds). *Pratique de l'analyse des données*, tome 5, économie, pp. 233–280.