Valeurs aberrantes et manquantes multivariées dans les sondages

Beat Hulliger [1] & Tobias Schoch [1]

[1] *Fachhochschule Nordwestschweiz FHNW, Hochschule für Wirtschaft Riggenbachstrasse 16, 4600 Olten,*
beat.hulliger@fhnw.ch

La détection de valeurs aberrantes multivariées est difficile pour les populations infinies mais est encore plus compliquée dans les sondages. Le plan de sondage, la non-réponse, les valeurs manquantes doivent être prises en compte et en plus les distributions rencontrées sont rarement normales. Par exemple des distribution semi-continues peuvent arriver. Le problème se pose dans des enquêtes auprès des entreprises, par exemple l'enquête sur les dépenses pour la protection de l'environnement, et dans les enquêtes auprès des ménages, par exemple l'enquête sur les revenues et les conditions de vie de l'UE (SILC).

La définition même d'une valeur aberrante devient plus complexe si plusieurs dimensions peuvent être manquantes. La notion d'une valeur aberrante aléatoirement ou complètement aléatoirement et la notion d'une contamination aléatoire ou complètement aléatoire (Béguin et Hulliger, 2008) doivent compléter les notions correspondantes pour les valeurs manquantes. Malheureusement les conditions pour récupérer complètement les distributions multivariées en présence de valeurs aberrantes se présentent rarement en pratique.

L'algorithme BACON-EEM pour la détection de valeurs aberrantes adapte l'algorithme BACON aux valeurs manquantes et au plan de sondage à l'aide d'une version de l'algorithme EM qui utilise les statistiques suffisantes estimées dans la population (Béguin et Hulliger, 2008). L'imputation qui suit la détection utilise aussi l'hypothèse d'une distribution multivariée normale et une estimation de la matrice de covariance robuste. L'algorithme épidémique se base sur une idée du plus proche voisin adaptée à la densité locale et, en principe, n'a pas d'hypothèse sur la distribution. L'imputation après détection par l'algorithme épidémique suit une épidémie en direction contraire.

En pratique il faut préparer les données avant de pouvoir utiliser ces algorithmes. La préparation a une influence importante. La détection et l'imputation dépendent de constantes d'ajustement qui sont à déterminer. L'exemple des données SILC et les simulations du projet AMELI (Alfons *et al.*, 2011) servent à illustrer les méthodes susmentionnées.

Bibliographie

Alfons, A., et Bruch, C., et Filzmoser, P., et Graf, M., et Hulliger, B., et Kolb, J.-P., et Lehtonen, R., et Lussmann, D., et Meraner, A., et Münnich, R., et Nedyalkova, D., et Schoch, T., et Templ, M., et Valaste, M., et Veijanen, A, et Zins, S. (2011). Report on the simulation results. Technical report, AMELI deliverable D7.1.

Béguin, C., et Hulliger, B. (2008). The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. Survey Methodology, 34, 91-103.