Clustering Data with Non-Ignorable Missingness using Semi-Parametric Mixture Models

Marie Du Roy De Chaumaray[1,2] and Matthieu Marbac*[2,3]

[1]Ecole Nationale de la Statistique et de lÁnalyse de lÍnformation [Bruz] – Centre de Recherche en Économie et STatistique (CREST) – France

[2]Centre de Recherche en Économie et Statistique (CREST) – Centre de Recherche en Économie et STatistique (CREST) – France

[3]Ecole Nationale de la Statistique et de lÁnalyse de lÍnformation – Ensai, Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

Abstract

This talk focus on the task of clustering of continuous data sets subject to non-ignorable missingness. We perform clustering with a specific semi-parametric mixture, under the assumption of conditional independence given the component. The mixture model is used for clustering and not for estimating the density of the full variables (observed and unobserved), thus, we do not need other assumptions concerning the component distribution or to specify the missingness mechanism. Estimation is performed by maximizing an extension of the smoothed likelihood allowing missingness. This optimization is achieved by a Majorization-Minorization algorithm. We illustrate the relevance of our approach by numerical experiments on simulated. Under mild assumptions, we show the identifiability of the model defining the distribution of the observed data and the monotonicity of the algorithm. We also propose an extension of this method to the case of mixed-type data that we illustrate on a real data set. Finally, extension of the approach relaxing the conditional independence assumption is proposed to the case of longitudinal data. The method is implemented in the R package MNARclust available on CRAN.

*Speaker