Dealing with missing data in model-based clustering through a MNAR model

Fabien Laporte$^1$

$^1$INRIA

28th August 2018

Abstract

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. Traditional ways to deal with them consist to obtain a filled data set, either by discarding missing values or by imputing them. In the first case some information is lost; in the second case the final clustering purpose is not taken into account through the imputation step. Thus both solutions risk to blur the clustering estimation result. Alternatively, we defend the need to embed the missingness mechanism directly within the clustering modeling step. There exists three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations, logistic regression is proposed as a natural and flexible candidate model. In particular, its flexibility property allows to design some meaningful parsimonious variants, as dependency on missing values or dependency on the cluster label. In this unify context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data.