Imputation multiple de données catégorielles : une approche basée sur un modèle multinormal latent

Anne De MOLINER[1] *et Philippe PÉRIÉ*[2]

La méthode d'imputation multiple basée sur l'algorithme MCMC dans le cadre multinormal (Schafer 1997) est très largement utilisée. Ses avantages sont nombreux : les algorithmes sont maintenant disponibles dans les logiciels commerciaux (SAS proc MI, SPSS version 17...), elle est simple et efficace à implémenter avec des routines de calcul largement disponibles (inversion de matrices, décompositions de Choleski, tirages dans des lois de Wishart...), on est dans un cadre théorique bien défini et, peut-être le plus important, elle accepte des dispositions arbitraires de valeurs manquantes.

Toutefois, puisque la méthode repose sur la multinormalité, elle peut achopper sur des cas de données catégorielles, en particulier pour des évènements rares (Horton, Lispsitz, Parzen 2003).

Parmi les autres approches, le modèle loglinéaire proposé par Schafer n'est plus applicable pratiquement pour des cas avec beaucoup de variables. Les séquences de régressions (Raghunathan 2001) peuvent sous-estimer les interactions dans le cas de données de grande dimensionnalité sur des évènement rares.

Notre proposition reprend la séquence de l'algorithme MCMC dans le cadre Normal et s'appuie sur la spécification d'un modèle multinormal latent, pour lequel les valeurs observées déterminent des troncatures. Les valeurs imputées sont donc tirées conditionnellement à ces troncatures.

Nous utilisons la méthode GHK (Geweke, Hajivassiliou, Keane) pour simuler des lois multinormales tronquées, et l'approche de Berens (2008) pour la détermination du modèle gaussien latent. Les jeux de données d'exemple sont tirés de données réelles d'évènements rares de grande dimension (200 variables, 40000 observations). Sur ces observées complètement, nous avons simulé plusieurs taux et arrangements de valeurs manquantes. La méthode est implémentée en SAS/IML et Matlab.

[1] Ensae

[2] TNS Sofres