Imputation de données manquantes pour des données mixtes via les méthodes factorielles grâce à missMDA

Vincent Audigier, François Husson, Julie Josse

Imputation de données manquantes pour des données mixtes via les méthodes factorielles grâce à missMDA

Vincent Audigier, Francois Husson, Julie Josse

Département de mathématiques appliquées, Agrocampus, Rennes, France
*Contact author : julie.josse@agrocampus-ouest.fr

Keywords: Imputation simple, Données mixtes, Données manquantes, Méthodes factorielles, ACP, ACM

Cette présentation a pour objet une nouvelle méthode d'imputation simple de données mixtes. L'objectif est alors de compléter des tableaux comprenant à la fois des variables quantitatives et qualitatives. L'imputation repose ici sur l'utilisation de méthodes factorielles.

Toutes les méthodes d'analyse factorielle peuvent s'écrire comme une ACP (Analyse en Composantes Principales) ou une décomposition en valeurs singulières d'un tableau de données particulier. L'ACP est donc au cœur de ces méthodes. L'approche classique pour gérer les données manquantes en ACP consiste à minimiser la fonction de coût (l'erreur de reconstitution) sur tous les éléments présents. Ceci peut être effectué à travers un algorithme d'ACP itérative (aussi appelé expectation maximisation PCA, EM-PCA) décrit dans Kiers (1997). Celui-ci consiste à attribuer une valeur initiale aux données manquantes, effectuer l'analyse (ACP) sur le jeu rendu complet, compléter les données manquantes via la formule de reconstitution pour un nombre d'axes fixé, et recommencer ces deux étapes jusqu'à convergence. Les paramètres (axes et composantes) ainsi que les données manquantes sont de cette manière simultanément estimés. Par conséquent cet algorithme peut être vu comme une méthode d'imputation simple. Il souffre cependant d'un problème de surajustement. En conséquence une version régularisée de cet algorithme doit être utilisée (Josse et al., 2009; Ilin and Raiko, 2010) afin de répondre à ce problème. De même un algorithme d'ACM régularisée permet de gérer les données manquante en ACM. Il consiste à effectuer une ACP régularisée sur une matrice judicieusement pondérée (Josse et al., 2012).

L'AFDM (Analyse Factorielle des Données Mixtes) généralise l'ACP et l'ACM, elle permet de traiter à la fois des données quantitatives propres à l'ACP et des variables qualitatives propres à l'ACM. La force de l'AFDM réside donc dans la prise en compte des relations entre individus, au même titre que toutes les autres méthodes factorielles, mais aussi, et c'est là son unicité, dans les relations entre les variables quantitatives et qualitatives équilibrées, renforçant ainsi la qualité d'imputation que l'on aurait eu en utilisant séparément une imputation par ACP et une par ACM. L'équilibre entre les différents types de variables est important au risque d'altérer l'imputation.

Le package **missMDA** (Husson and Josse, 2010) permet de gérer les données manquantes dans les méthodes d'analyse factorielle. Il s'agit d'abord d'imputer les données manquantes à l'aide des fonctions du package, puis d'effectuer l'analyse à l'aide d'un logiciel adapté comme **FactoMineR** (Lê et al., 2008; Husson et al., 2011). Ainsi, **missMDA** permet d'envisager tout type d'analyse et ceci en dépit de l'absence de données.

Bien que le problème de données manquantes sur des données mixtes soit courant, peu de méthodes d'imputation sont disponibles. Une des plus récentes (2011) et offrant de bons résultats est basée sur les forêts aléatoires et donc sur des prédicteurs par arbres (Stekhoven and Buhlmann, 2011). Les comparaisons avec cette méthode d'imputation offrent des résultats comparables et encourageants autant sur des jeux réels que simulés.

Références

Husson, F. and J. Josse (2010). *missMDA : Handling missing values with/in multivariate data analysis (principal component methods)*. R package version 1.2.

Husson, F., J. Josse, S. Le, and J. Mazet (2011). *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.16.

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research 11*, 1957–2000.

Josse, J., M. Chavent, B. Liquet, and F. Husson (2012). Handling missing values with regularized iterative multiple correspondence analysis. *Journal of classification 29*, 91–116.

Josse, J., J. Pagès, and F. Husson (2009). Gestion des données manquantes en analyse en composantes principales. *Journal de la Société Française de Statistique 150*, 28–51.

Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika 62*, 251–266.

Lê, S., J. Josse, and F. Husson (2008, 3). Factominer : An r package for multivariate analysis. *Journal of Statistical Software 25*(1), 1–18.

Stekhoven, D. and P. Buhlmann (2011). Missforest - nonparametric missing value imputation for mixed-type data. *Bioinformatics 28*, 113–118.