# Imputation multiple pour données mixtes par analyse factorielle

Vincent Audigier, François Husson, Julie Josse, Matthieu Resche-Rigon

# Imputation multiple pour données mixtes par analyse factorielle

Vincent Audigier [1] & François Husson [2] & Julie Josse [3] & Matthieu Resche-Rigon [4]

[1] *CNAM, Laboratoire Cedric MSDMA, 2 rue Conté 75003 Paris,*
*vincent.audigier[at]cnam.fr*
[2] *Laboratoire de mathématiques appliquées, Agrocampus Ouest, 65 rue de Saint Brieuc 35042 RENNES,* *husson[at]agrocampus-ouest.fr*
[3] *Centre de Mathématiques Appliquées, Ecole Polytechnique, France*
*XPOP, INRIA, France* *julie.josse[at]polytechnique.edu*
[4] *Service de Biostatistique et Information Medicale, Hopital Saint-Louis, AP-HP, Paris, France*
*Universite Paris Diderot - Paris 7, Sorbonne Paris Cite, UMR-S 1153, Paris, France*
*INSERM, UMR 1153, Equipe ECSTRA, Hopital Saint-Louis, Paris,*
*matthieu.resche-rigon[at]univ-paris-diderot.fr*

**Résumé.** La prise en compte de données toujours plus nombreuses complexifie sans cesse leur analyse. Cette complexité se traduit notamment par des variables de types différents, la présence de données manquantes, et un grand nombre de variables et/ou d'observations. L'application de méthodes statistiques dans ce contexte est généralement délicate.

L'objet de cette présentation est de proposer une nouvelle méthode d'imputation multiple basée sur l'analyse factorielle des données mixtes (AFDM). L'AFDM est une méthode d'analyse factorielle adaptée pour des jeux de données comportant des variables quantitatives et qualitatives, dont le nombre peut excéder, ou non, le nombre d'observations. En vertu de ses propriétés, le développement d'une méthode d'imputation multiple basée sur l'AFDM permet l'inférence sur des variables quantitatives et qualitatives incomplètes, en grande et petite dimension.

La méthode d'imputation multiple proposée utilise une approche bootstrap pour refléter l'incertitude sur les composantes principales et vecteurs propres de l'AFDM, utilisés ici pour prédire (imputer) les données. Chaque réplication bootstrap fournit alors une prédiction pour l'ensemble des données incomplètes du jeu de données. Ces prédictions sont ensuite bruitées pour refléter la distribution des données. On obtient ainsi autant de tableaux imputés que de réplications bootstrap.

Après avoir rappelé les principes de l'imputation multiple, nous présenterons notre méthodologie. La méthode proposée sera évaluée par simulation et comparée aux méthodes de références : imputation séquentielle par modèle linéaire généralisé, imputation par modèle de mélanges et par "general location model". La méthode proposée permet d'obtenir des estimations ponctuelles sans biais de différents paramètres d'intérêt ainsi que des intervalles de confiance au taux de recouvrement attendu. De plus, elle peut s'appliquer

sur des jeux de données de nature variée et de dimensions variées, permettant notamment de traiter les cas où le nombre d'observations est plus petit que le nombre de variables.

**Mots-clés.** Données manquantes, imputation multiple, données mixtes, analyse factorielle des données mixtes

**Abstract.** Accounting for more and more data complicates increasingly their analysis. This complexity results in variables of various types, the presence of missing data, and a large number of variables and / or observations. The application of statistical methods in this context is usually tricky.

The purpose of this presentation is to propose a new multiple imputation method based on the factorial analysis of mixed data (FAMD). FAMD is a suitable factor analysis method for datasets with quantitative and qualitative variables, the number of which may or may not exceed the number of observations. By virtue of its properties, the development of a multiple imputation method based on FAMD allows inference from quantitative and qualitative incomplete variables, in large and small dimension.

The proposed multiple imputation method uses a bootstrap approach to reflect the uncertainty on the principal components and eigenvector of FAMD, used here to predict (impute) the data. Each bootstrap replication then provides a prediction for incomplete data of the dataset. Next, these predictions are noised to reflect the distribution of the data. We thus obtain as many imputed tables as bootstrap replicates.

After recalling the principles of multiple imputation, we will present our methodology. The proposed method will be evaluated by simulation and compared to the reference multiple imputation methods : sequential imputation by generalized linear model, imputation by non-parametric Bayesian joint model, and by general location model. The proposed method provides unbiased point estimates of various parameters of interest as well as confidence intervals at the expected coverage. In addition, it can be applied to datasets of various type and of various sizes, in particular to deal with cases where the number of observations is smaller than the number of variables.

**Keywords.** Missing values, multiple imputation, mixed data, factorial analysis of mixed data

La prise en compte de la variété des données analysées est une difficulté de plus prégnante dans la pratique de la statistique d'aujourd'hui. Citons par exemple le domaine médical où l'on peut disposer pour chaque patient de réponses à un questionnaire, de comptes-rendus médicaux, de résultats d'analyse... Cette variété se traduit notamment par des variables parfois quantitatives, parfois qualitatives. Par exemple, on disposera de données qualitatives pour des questionnaires à choix multiples, ou pour des comptes-rendus, et de données quantitatives pour des mesures de marqueurs biologiques. On parle alors de *données mixtes*.

A cette première difficulté s'ajoute généralement celle des données manquantes. Les causes en sont multiples : problème de saisie, fusion de fichiers, appareil de mesure

défectueux, etc. Les données étant par ailleurs de plus en plus volumineuses, le nombre d'individus incomplets est nécessairement de plus en plus grand, rendant cette problématique de plus en plus incontournable.

La gestion des données manquantes en présence de données mixtes est délicate. Une solution consiste à remplacer les données incomplètes par des valeurs plausible, on parle alors d'*imputation simple*.

Récemment, une méthode d'imputation simple reposant sur l'analyse factorielle des données mixtes (AFDM) a été proposée (Audigier et al., 2016) avec des résultats encourageants en termes de prédiction. Toutefois, bien que prédire une valeur manquante soit intéressant en soi, le statisticien est généralement plus intéressé par la mise en oeuvre d'une analyse statistique sur ses données, par exemple l'inférence via le modèle linéaire généralisé. Or, l'imputation simple ne suffit généralement pas à une telle analyse car elle ne prend pas en compte l'incertitude liée aux données imputées. En conséquence, appliquer une méthode statistique sur un tableau imputé simplement impliquera une sous-estimation de la variabilité des estimateurs associés à cette méthode. Pour pouvoir refléter la variance de prédiction de chaque donnée manquante, on impute plusieurs fois chaque valeur manquante de façon stochastique, amenant donc à plusieurs tableaux imputés. Puis, sur chacun des différents tableaux, on applique la méthode statistique souhaitée dont on agrège ensuite les paramètres selon les règles de Rubin (Rubin, 1987). On parle d'*imputation multiple* (Rubin, 1987; Little and Rubin, 2002). On obtient ainsi une unique estimation des paramètres de la méthode ainsi qu'une estimation de la variabilité associée.

Ainsi, cette communication a pour but de présenter l'extension de l'imputation simple des données mixtes par l'AFDM à sa version imputation multiple.

Pour imputer un jeu de données mixtes à l'aide de l'AFDM on utilise un algorithme appelé AFDM itérative. Cet algorithme débute par une phase d'initialisation où les variables qualitatives sont dans un premier temps recodées en tableau disjonctif, les variables quantitatives ne sont quant à elles pas modifiées. Puis, les valeurs manquantes sont imputées par la moyenne de chaque colonne du nouveau tableau. A partir de ce tableau rendu complet, une décomposition en valeurs singulières permet d'estimer les composantes principales et les vecteurs propres qui constituent les paramètres de l'AFDM. Les données sont ensuite imputées en effectuant le produit matriciel des $S$ premières composantes principales et vecteurs propres. Ces étapes d'estimation des paramètres et d'imputation sont alors répétées jusqu'à convergence. On pourra noter que l'AFDM itérative est l'équivalent de l'ACP itérative (Kiers, 1997) qui, comme l'algorithme NIPALS (Christoffersson, 1970), permet d'estimer les paramètres d'une analyse en composantes principales avec données manquantes. A l'issue de l'algorithme, on obtient donc un tableau dans lequel les indicatrices codant pour chaque variable qualitative sont des réels sommant à 1 par variable et non uniquement des zéros et des uns comme dans un tableau disjonctif "classique". Ces valeurs réelles sont ensuite ramenées à l'intervalle $[0, 1]$ via une normalisation, ce qui

permet de les lire comme des probabilités d'appartenance. On remonte alors aux données qualitatives en effectuant un tirage selon ces probabilités pour chaque donnée incomplète, simulant ainsi la distribution originelle du jeu de données. Les valeurs imputées sur les variables quantitatives sont quant à elles bruitées par un bruit gaussien pour la même raison.

L'imputation multiple par AFDM ne peut cependant pas se limiter à une succession d'imputations simples de ce type. En effet, les paramètres du modèle d'imputation sont estimés à partir d'un même échantillon : le tableau incomplet. Il est nécessaire de prendre en compte l'incertitude vis-à-vis de cette estimation. Pour ce faire il faut se doter d'un jeu de $M$ paramètres obtenus à partir des données observées. Cela permet de refléter, à travers les données imputées, l'incertitude dans l'estimation des paramètres du modèle d'imputation. Pour se procurer un tel jeu de paramètres on propose d'adopter une approche bootstrap. Celle-ci consiste à effectuer un tirage dans les indices des individus, puis à affecter à chaque individu un poids proportionnel au nombre de fois où son indice a été tiré. Les individus dont l'indice n'a pas été tiré se voient attribuer un poids nul. On définit ainsi $M$ pondérations différentes, puis on impute de façon simple le tableau de données selon chacune de ces pondérations. La pondération intervient au niveau de la décomposition en valeurs singulières et amènera donc à une estimation particulière des paramètres. De cette façon, on obtient $M$ jeux de données imputés reflétant l'incertitude sur les paramètres du modèle d'imputation.

La vérification de la validité d'une méthode d'imputation multiple s'effectue par simulation. En particulier, la méthode d'imputation doit permettre d'obtenir des estimations ponctuelles fiables de la quantité d'intérêt associée à la méthode statistique employée, ainsi que de la variabilité associée à cette estimation. Des simulations ont été effectuées dans le cadre classique de données manquantes distribuées au hasard (MAR) pour plusieurs quantités d'intérêt.

La méthode proposée a été comparée à l'imputation multiple par le "general location model" (Schafer, 1997), méthode de référence, mais rapidement limitée en présence d'un nombre modéré de variables qualitatives ; l'imputation selon des modèles de mélanges (Murray and Reiter, 2016), méthode la plus récente gérant la complexité combinatoire engendrée par les données qualitatives ; l'imputation multiple séquentielle (van Buuren, 2018), probablement la méthode la plus plébiscitée par les utilisateurs. L'imputation multiple par AFDM fournit de bonnes estimations ponctuelles des paramètres d'intérêt tout en construisant des intervalles de confiance valides. De plus, elle peut s'appliquer sur des jeux de données de tailles quelconques et permet notamment de traiter les cas où le nombre d'individus est inférieur au nombre de variables.

# Références


Audigier, V., F. Husson, and J. Josse (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification 10*(1), 5–26.

Christoffersson, A. (1970). *The one component model with incomplete data*. Wilkinson.

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

Little, R. J. A. and D. B. Rubin (1987, 2002). *Statistical Analysis with Missing Data*. New-York : Wiley series in probability and statistics.

Murray, J. S. and J. P. Reiter (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association 111*(516), 1466–1479.

Rubin, D. B. (1987). *Multiple Imputation for Non-Response in Survey*. Wiley.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London : Chapman & Hall/CRC.

van Buuren, S. (2018). *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)* (2 ed.). Chapman and Hall/CRC.