Imputation multiple à l'aide des méthodes d'analyse factorielle

Vincent Audigier & François Husson & Julie Josse

Laboratoire de mathématiques appliquées, Agrocampus Ouest
65 rue de Saint Brieuc 35042 RENNES Cedex
audigier@agrocampus-ouest.fr ; husson@agrocampus-ouest.fr ; josse@agrocampus-ouest.fr

Résumé. Les données manquantes constituent un problème incontournable dans la pratique de la statistique. Une solution commune pour gérer ces données manquantes consiste à remplacer chacune d'entre elles par une valeur plausible. On parle d'imputation simple. Néanmoins appliquer une méthode statistique sur un tableau imputé simplement pose un problème majeur : les données imputées jouent le même rôle que les données observées alors qu'elles sont incertaines. Pour rendre compte de cette incertitude, on peut proposer plusieurs imputations pour chaque donnée manquante. On parle alors d'imputation multiple.

Cette présentation a pour objet de nouvelles méthodes d'imputation multiple pour des données quantitatives, qualitatives et mixtes dans le cadre de données manquantes au hasard. L'idée est d'étendre les méthodes d'imputation simples basées sur l'emploi de techniques de réduction de la dimension telle que l'analyse en composantes principales.

Après avoir présenté les principes de l'imputation multiple, nous détaillerons les propriétés de nos différents algorithmes. Nous proposerons ensuite des simulations pour les situer par rapports aux méthodes existantes telles que l'imputation multiple par équations enchaînées (van Buuren, 2012), ou l'imputation reposant sur l'hypothèse d'une distribution jointe à l'ensemble des données.

Mots-clés. Imputation multiple, Données mixtes, Données manquantes, Méthodes factorielles, ACP, ACM, AFDM

Abstract. Missing data are a key problem in statistical practice. A common solution to handle missing data is to use simple imputation which consists in replacing them with a plausible value. However applying a statistical method on an imputed dataset poses a major problem : the imputed values have the same status as the observed values whereas they are uncertain. To account for this uncertainty, we can propose several imputation for the same dataset. This is called multiple imputation.

This presentation proposes new methods of multiple imputation for continuous, categorical and mixed data for data missing at random. This will be done by extending the concept of simple imputation based on the use of dimensionality reduction techniques such as principal components analysis.

After presenting the principles of multiple imputation, we detail the properties of the new multiple imputation methods based on principal components methods. We then propose simulations to compare them to others existing approaches such as multiple imputation by chained equations (van Buuren, 2012) and imputation assuming a joint distribution for all data.

Keywords. Multiple imputation, Mixed data, Missing values, Principal components methods, PCA, MCA, FAMD

Les données manquantes sont incontournables dans la pratique de la statistique. La plupart des méthodes statistiques ne s'appliquant qu'à des tableaux complets, une solution usuelle consiste à supprimer les individus portant des données manquantes. Ceci n'est cependant pas judicieux, en particulier sur de petits tableaux. Dès lors une solution consiste à remplacer les données manquantes par une valeur plausible et d'appliquer ensuite la méthode statistique souhaitée. De nombreuses méthodes d'imputation sont disponibles (Little and Rubin, 2002; Schafer, 1997; Troyanskaya et al., 2001; Stekhoven and Bühlmann, 2011). Récemment, de nouvelles méthodes d'imputation reposant sur les méthodes d'analyse factorielle ont été proposées (Josse and Husson, 2012).

Toutes les méthodes d'analyse factorielle peuvent s'écrire comme une analyse en composantes principales (ACP) 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 qui peut être contourné grâce à une version régularisée de cet algorithme (Josse et al., 2009; Ilin and Raiko, 2010). De même, l'algorithme itératif d'analyse des correspondances multiple (ACM) régularisé permet de gérer les données manquante en ACM (Josse et al., 2012) et donc d'imputer des données qualitatives. L'algorithme itératif d'analyse factorielle des données mixtes (AFDM) régularisé permet quant à lui de gérer les données mixtes. Tous deux consistent à effectuer une ACP itérative régularisée sur une matrice judicieusement pondérée.

Ces méthodes d'imputation permettent de remplacer chacune des données manquantes par la valeur la plus plausible aux vues du modèle sous-jacent supposé et des données présentes. Cependant, si on applique une méthode statistique sur un tableau imputé, la variabilité de l'estimateur sera sous-estimée. Ceci pour la simple raison que le tableau

imputé ne rend pas compte de l'incertitude liée aux données manquantes. Dès lors une solution consiste à envisager ces méthodes d'imputation simple dans leur version multiple (Little and Rubin, 2002). En proposant plusieurs valeurs plausibles pour les données imputées, on va pouvoir rendre compte de l'incertitude liée aux données manquantes et ainsi estimer de manière correcte la variance des estimateurs.

L'imputation multiple par ACP a déjà été utilisée dans l'optique de vouloir estimer de façon fiable la variance des estimateurs des paramètres de l'ACP (Josse and Husson, 2011). Ceci a notamment permis de proposer des ellipses de confiance autour des coordonnées des individus. Cependant, cette méthode n'a pas été étudiée en tant que méthode d'imputation multiple à proprement parler. En approfondissant ces premiers travaux, il va être possible de proposer dans un premier temps une méthode d'imputation multiple des tableaux de données quantitatifs pour un mécanisme de données manquantes de type MAR (Missing At Random). Dans un second temps, il s'agit de s'intéresser à des tableaux de nature quelconque. Bien que constituant le cadre le plus commun, la littérature concernant l'imputation des tableaux de données mixtes est peu fournie. A fortiori l'imputation multiple pour ce type de données est peu développée. L'intérêt des méthodes d'imputation basées sur l'analyse factorielle est qu'elles s'étendent naturellement aux données mixtes. Ainsi nous présenterons comment envisager ces méthodes d'analyse factorielle dans un cadre multiple pour des données de nature quelconque et les comparerons aux méthodes existantes telles que l'imputation multiple par équations enchaînées (van Buuren, 2012), ou l'imputation reposant sur l'hypothèse d'une distribution jointe à l'ensemble des données.

Références

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res. 99*, 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. and F. Husson (2011). Multiple imputation in PCA. *Advances in data analysis and classification 5*, 231–246.

Josse, J. and F. Husson (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique 153 (2)*, 1–21.

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.

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

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

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

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman (2001). Missing value estimaion methods for dna microarrays. *Bioinformatics 17*(62001), 520–525.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. London : Chapman & Hall/CRC interdisciplinary statistics.