Imputation multiple pour variables quantitatives par analyse en composantes principales Bayésienne

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.

L'objet de cette présentation est de proposer une méthode d'imputation multiple dédiée aux variables quantitatives et basée sur le modèle d'analyse en composantes principales (ACP). L'emploi d'un traitement bayésien du modèle d'ACP va permettre de disposer d'une distribution sur les paramètres de ce modèle et ainsi de refléter l'incertitude sur les paramètres du modèle d'imputation.

Après avoir rappelé les principes de l'imputation multiple, nous présenterons notre méthodologie. La méthode proposée sera ensuite évaluée par simulation et comparée à deux méthodes existantes : l'imputation multiple par équations enchaînées, et l'imputation reposant sur l'hypothèse d'une distribution jointe à l'ensemble des données. La méthode proposée fournit de bonnes estimations ponctuelles des paramètres d'intérêt tout en construisant des intervalles de confiance valides et de tailles réduites. De plus, contrairement aux deux autres méthodes, elle permet de traiter facilement les cas où le nombre d'individus est inférieur au nombre de variables.

Mots-clés. Imputation multiple, Données manquantes, ACP, Bayésien, Data Augmentation

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 a new method of multiple imputation dedicated for continuous variables based on the principal component analysis (PCA) model. The use of a Bayesian treatment of the PCA model will allow to put a distribution on the parameters of the model and thus to reflect the uncertainty about the parameters of the imputation model.

After recalling the principles of multiple imputation, we present our methodology. The proposed method is assessed using simulations and compared to two existing methods : the multiple imputation by chained equations, and the imputation assuming a joint distribution for all data. The proposed method provides a good point estimate of the quantity of interest, an estimate of the variability of the estimator reliable while reducing the width of the confidence intervals built around the quantity of interest. Moreover, contrary to the two other methods, it easily deals with cases where the number of individuals is smaller than the number of variables.

Keywords. Multiple imputation, Missing values, PCA, Bayesian, Data Augmentation

Les données manquantes constituent un problème incontournable dans la pratique de la statistique. En effet, la plupart des méthodes statistiques ne peuvent pas être directement appliquées sur un jeu incomplet. Une solution classique pour gérer les données manquantes consiste à effectuer de l'imputation simple. Cela consiste à remplacer les données manquantes par des données plausibles. Ainsi, on obtient un jeu completé sur lequel on peut appliquer n'importe quelle méthode statistique.

Récemment, des méthodes d'imputation simple reposant sur les méthodes d'analyse factorielle ont été proposées (Josse and Husson, 2012; Audigier et al., 2013) avec des résultats encourageants : l'analyse en composantes principales (ACP) permet d'imputer des variables de nature quantitatives, l'analyse des correspondances multiple (ACM), d'imputer des variables qualitatives et l'analyse factorielle des données mixtes (AFDM), d'imputer des données mixtes.

Toutefois, l'imputation simple, aussi bonne qu'elle soit, reste une méthode limitée dans le sens où elle ne prend pas en compte l'incertitude liée aux données imputées. Ainsi, si on applique une méthode statistique sur le jeu rendu complet par l'imputation, la variabilité des estimateurs sera sous-estimée. Pour résoudre ce problème, on peut effectuer de l'imputation multiple (Rubin (1987), Little and Rubin (2002)). Dans ce cas plusieurs valeurs sont prédites pour chaque donnée manquante, ce qui amène à considérer plusieurs tableaux imputés. Ainsi, la variabilité inter-imputation reflète la variance de prédiction de chaque donnée manquante. Une fois les tableaux imputés, on applique la méthode statistique sur chacun d'entre eux, puis on agrège les résultats selon les règles de Rubin (Rubin (1987)) afin d'obtenir une unique estimation des paramètres de la méthode ainsi qu'une estimation de la variabilité associée. Cette communication a pour but de présenter des

premiers travaux d'extention des méthodes d'imputation simple par analyse factorielle à des méthodes d'imputation multiple à savoir l'imputation multiple des variables quantitatives grâce à l'ACP.

L'ACP peut s'exprimer sous la forme d'un modèle à effet fixe : $\mathbf{X}_{n\times p} = \tilde{\mathbf{X}}_{n\times p} + \mathbf{E}_{n\times p}$ avec $\mathbf{X}$ le tableau de données, $\tilde{\mathbf{X}}$ une matrice de rang inférieur S et $\mathbf{E}$ une matrice d'erreurs. L'imputation simple par ACP consiste à estimer $\tilde{\mathbf{X}}$ en dépit de la présence de données manquantes sur $\mathbf{X}$, puis à remplacer les données manquantes de $\mathbf{X}$ par les estimations correspondantes de $\tilde{\mathbf{X}}$ auxquelles une perturbation aléatoire est ajoutée afin de simuler la distribution originelle du jeu de données. L'imputation multiple par ACP ne peut cependant pas se résumer pas 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 ce procurer un tel jeu de paramètres, deux manières de procéder sont envisageables : l'approche bootstrap (Josse and Husson, 2011) et l'approche bayésienne (Audigier et al., 2014).

Le principe de l'approche bayésienne est de définir une distribution a priori sur les paramètres du modèle d'imputation et ainsi de pouvoir effectuer un tirage dans la distribution a posteriori afin d'obtenir un jeu de M paramètres. Toutefois, si l'on connaît la loi a posteriori des paramètres du modèle en combinant la loi a priori et les données dans le cas complet, sa détermination en présence de données manquantes n'est pas directe. L'algorithme de "data augmentation" permet cependant d'effectuer un tirage dans cette distribution. Il s'agit d'un algorithme itératif qui consiste à alterner les étapes d'imputation et de tirage des paramètres du modèle dans la loi a posteriori. En effet, une fois le tableau rendu complet, il devient facile de déterminer la loi a posteriori des paramètres du modèle d'ACP. Le jeu de paramètres est alors obtenu en appliquant cet algorithme M fois.

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 de qualité de la quantité d'intérêt associée à la méthode statistique employée, ainsi qu'une estimation fiable de la variabilité associée à cette estimation. Les simulations ont été effectuées dans un cadre de données manquantes distribuées complètement au hasard pour trois quantités d'intérêt différentes : une espérance, un coefficient de corrélation et un coefficient de régression. De plus l'algorithme proposé à été comparé à deux autres algorithmes de référence, particulièrement adaptés au type de données simulées : l'algorithme d'imputation multiple par équations enchaînées et celui d'imputation sous l'hypothèse d'une distribution jointe à l'ensemble des données. Ces simulations montrent que, dans des configurations variées, l'algorithme d'imputation multiple par ACP fournit de bonnes estimations ponctuelles des paramètres d'intérêt tout en produisant des intervalles

de confiance valides et de tailles réduites. De plus, la méthode proposée permet de traiter plus facilement les cas où le nombre d'individus est inférieur au nombre de variables.

Références

Audigier, V., F. Husson, and J. Josse (2013). A principal components method to impute missing values for mixed data. *ArXiv e-prints*. In revision.

Audigier, V., F. Husson, and J. Josse (2014). Multiple imputation for continuous variables using a Bayesian principal component analysis. *ArXiv e-prints*. Submitted.

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.

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

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