Imputation multiple pour variables qualitatives par analyse des correspondances multiples

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. Il est très fréquent de rencontrer des données manquantes dans la pratique de la statistique. Or la plupart des méthodes statistiques ne peuvent pas être directement appliquées sur un jeu incomplet. Pour dépasser cette difficulté on peut remplacer les données manquantes par des valeurs plausibles, on parle alors d'*imputation simple*. Cependant, l'imputation simple ne permet pas de prendre en compte l'incertitude liée aux données imputées. Pour refléter 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 qualitatives et basée sur l'analyse des correspondances multiples (ACM). L'emploi d'une approche bootstrap va permettre de se doter de M jeux de composantes principales et vecteurs propres. Ces jeux de paramètres sont ensuite utilisés pour construire M imputations du jeu de données permettant 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 aux quelques méthodes existantes : imputation multiple par modèle loglineaire, par équations enchaînées et par modèle à classes latentes. La méthode proposée fournit de bonnes estimations ponctuelles des paramètres d'intérêt et de bons intervalles de confiance. 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.

Mots-clés. Imputation multiple, Données manquantes, ACM, Bootstrap, Variables qualitatives

Abstract. Missing values are very common in statistical practice. However, most of statistical methods cannot be applied directly on incomplete datasets. To deal with missing values, substituting of missing values by plausible values can be done. This is called *single imputation*. But single imputation does not take into account the uncertainty due to imputed values. To account for this uncertainty, several imputations for the same dataset could be proposed. This is called *multiple imputation*.

This presentation proposes a new method of multiple imputation dedicated for categorical variables based on multiple correspondence analysis (MCA). The use of a bootstrap

procedure allow us to provide M sets of principal components and eigen vectors. Then, these sets are used to impute the incomplete dataset M times, reflecting in this way the uncertainty on 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 existing methods : the multiple imputation by the loglinear model, the multiple imputation by chained equations, and the multiple imputation by the latent class model. The proposed method provides a good point estimate of the quantity of interest and a reliable estimate of the variability of the estimator. Moreover, it easily deals with cases where the number of individuals is smaller than the number of variables.

Keywords. Multiple imputation, Missing values, MCA, Bootstrap, Categorical variables

Les grandes enquêtes d'opinion se font généralement via la soumission de questionnaires auprès d'un échantillon d'une population. Ceux-ci peuvent alors prendre la forme de questionnaires à choix multiples ce qui amène à collecter de nombreuses variables qualitatives. Les données qualitatives sont également très présentes dans le domaine bancaire où chaque client est décrit par un ensemble de variables telles que le département de résidence, le sexe, la situation familiale, la catégorie socio-professionnelle, etc. Des méthodes statistiques spécifiques sont par la suite appliquées sur ces données afin de répondre à des problématiques particulières. Dans le cadre des données bancaire, on peut par exemple souhaiter appliquer des méthodes de scoring afin de déterminer quel type de prêt accorder à un client.

Il est très fréquent d'observer des données manquantes dans ce type de données. Les causes en sont multiples : problème de saisie, les données sont issues d'une agrégation de plusieurs sources dans lesquelles les individus diffèrent, les questionnaires peuvent contenir des questions embarrassantes portant par exemple sur la consommation d'alcool ou sur les revenus, etc. Or la plupart des méthodes statistiques ne peuvent pas être directement appliquées sur un jeu incomplet. Pour dépasser cette difficulté on peut remplacer les données manquantes par des valeurs plausibles, on parle alors d'*imputation simple*.

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. En particulier l'analyse des correspondances multiple (ACM) permet d'imputer des variables de nature qualitative.

Toutefois l'imputation simple a ses limites : 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, l'astuce consiste à imputer plusieurs fois chaque valeurs manquante,

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 variables qualitatives par ACM à sa version imputation multiple.

Pour imputer un jeu de données qualitatif à l'aide de l'ACM on utilise un algorithme appelé ACM itérative. Cet algorithme débute par une phase d'initialisation où le tableau qualitatif incomplet est dans un premier temps recodé en tableau disjonctif (incomplet), puis imputé par la moyenne de chaque indicatrice. A partir de ce tableau disjonctif rendu complet, une décomposition en valeurs singulières permet d'estimer les composantes principales et les vecteurs propres qui constituent les paramètres de l'ACM. 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'ACM 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 donnees manquantes. A l'issue de l'algorithme, on obtient donc un tableau disjonctif complété "flou" dans le sens où les valeurs imputées de ce tableau 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.

En présence d'un petit nombre d'individus, de liaisons faibles entre variables ou d'un nombre élevé de données manquantes, cet algorithme peut souffrir de problèmes de surajustement. Pour cette raison on lui préférera sa version régularisée dont le principe est d'attribuer un poids plus important aux premières dimensions, plus stables, lors de l'étape d'imputation.

L'imputation multiple par ACM 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 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 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. Des simulations ont été effectuées dans un cadre de données manquantes distribuées complètement au hasard (MCAR) ou au hasard (MAR) pour plusieurs dizaines de quantités d'intérêt : coefficients de régression logistique ou paramètres de modèle loglinéaire.

La méthode proposée a été comparée aux quelques méthodes existantes pour imputer des variables qualitatives : le modèle loglineaire (Schafer, 1997), très performant sur des jeux de données avec peu de variables mais ne permettant pas de traiter de grands jeux de données ; l'imputation multiple par équations enchaînées (van Buuren and Groothuis-Oudshoorn, 2011), plus souple, performante, mais lente et nécessitant de définir un modèle d'imputation pour chaque variable incomplète ; l'imputation selon le modèle à classes latentes (Si and Reiter, 2013) qui repose sur une hypothèse de structure des individus en classes mais qui ne nécessite pas de paramétrage et qui peut s'appliquer sur de grands jeux de données. L'imputation multiple par ACM 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 (2013). A principal components method to impute missing values for mixed data. *ArXiv e-prints*. In revision.

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

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.

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.

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.

Si, Y. and J. Reiter (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics 38*, 499–521.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice : Multivariate imputation by chained equations in R. *Journal of Statistical Software 45*(3), 1–67.