Adv Data Anal Classif (2016) 10:5–26
DOI 10.1007/s11634-014-0195-1

REGULAR ARTICLE

A principal component method to impute missing values for mixed data

Vincent Audigier · François Husson · Julie Josse

Received: 6 February 2013 / Revised: 15 December 2014 / Accepted: 16 December 2014 /
Published online: 24 December 2014

Abstract We propose a new method to impute missing values in mixed data sets. It is based on a principal component method, the factorial analysis for mixed data, which balances the influence of all the variables that are continuous and categorical in the construction of the principal components. Because the imputation uses the principal axes and components, the prediction of the missing values is based on the similarity between individuals and on the relationships between variables. The properties of the method are illustrated via simulations and the quality of the imputation is assessed using real data sets. The method is compared to a recent method (Stekhoven and Buhlmann Bioinformatics 28:113–118, 2011) based on random forest and shows better performance especially for the imputation of categorical variables and situations with highly linear relationships between continuous variables.

Keywords Missing values · Mixed data · Imputation · Principal component method · Factorial analysis of mixed data

Mathematics Subject Classification 62H25

1 Introduction

Missing data are a key problem in statistical practice. Indeed, they are never welcome, because most statistical methods cannot be applied directly on an incomplete data set. One of the common approaches to deal with missing values consists of imputing missing values by plausible values. This leads to a complete data set that can be analyzed by any statistical method. However, results must be interpreted cautiously, since there is necessarily uncertainty associated with the prediction of values.

V. Audigier · F. Husson (✉) · J. Josse
Agrocampus Ouest, 65 rue de St-Brieuc, 35042 Rennes, France
e-mail: husson@agrocampus-ouest.fr

Several imputation methods are available for continuous data such as K nearest neighbors imputation (Troyanskaya et al. 2001), imputation based on multivariate normal model (Schafer 1997) or multivariate imputation by chained equations (van Buuren et al. 1999; van Buuren 2007). The multivariate normal model defines a joint distribution for the data which in practice can be restrictive. Imputation by chained equations consists of defining a model for each variable with missing data, which can afford a finer modeling but requires defining many models. It is also possible to impute continuous data with principal component analysis (PCA). The idea is to use an algorithm that performs PCA despite the missingness of some data (Kiers 1997; Josse et al. 2009; Ilin and Raiko 2010). Principal components and axes obtained by this algorithm are then used to reconstruct the data, which provides an imputation of the missing entries. This method has the particular advantage of simultaneously imputing any missing data by taking into account the similarities between individuals and the relationships between variables.

The imputation of categorical data can also be done with non-parametric methods such as the K nearest neighbours or with parametric methods based on different models. The most common model is the log-linear one (Schafer 1997). It has a major drawback: the number of parameters increases rapidly with the number of levels of categorical variables. Therefore, in practice, it becomes unusable in some cases. Other models have been proposed to overcome this problem, such as the latent class model (Vermunt et al. 2008), or the log-linear hierarchical model (Schafer 1997; Little and Rubin 1987, 2002).

Finally, for mixed data, i.e. both continuous and categorical, literature is less abundant. Indeed, dealing with mixed data is difficult because it requires to take into account the relationships between the variables that are of different types. One possible solution consists of coding the categorical variables using the indicator matrix of dummy variables, and using an imputation method dedicated to continuous variables on the concatenated matrix of continuous variables and the indicator matrix. However, this method is not satisfactory since the usual assumptions for continuous variables do not hold for dummy variables. Schafer (1997) proposed an imputation based on the general location model which can be seen as combination of the log-linear model and multivariate normal model; this imputation has the benefits and drawbacks of such models. Imputation by chained equations (van Buuren et al. 1999; van Buuren 2007) is one of the only approaches that can be easily extended to the mixed case by defining a specific model for each continuous and each categorical variable. However, as mentioned for the continuous case, many models have to be defined. Recently, Stekhoven and Bühlmann (2011) proposed an imputation method based on random forest Breiman (2001). The imputation is done using the following iterative algorithm: after replacing the missing data with initial values, the missing values of the variable with the fewest missing values are predicted by random forest. This operation is performed for each variable in the data set and the procedure is repeated until the predictions stabilize. For mixed data, this method was compared to the imputation by chained equations and to a version of the K nearest neighbours method adapted for mixed data. Their approach clearly outperforms the competing methods across many simulations and real data sets. It provides a good quality of imputation regardless of the number of observations and variables and the type of relationship between variables. Further, it is largely insensitive to the tuning parameters. Thus this nonparametric method based on random forest can serve as a reference among the existing methods.

We propose a new imputation method for mixed data based on a principal component method dedicated to mixed data: *the factorial analysis for mixed data* (FAMD) presented in Escofier (1979), also known as PCAMIX Kiers (1991). Because the method is based on a principal component method, imputation using FAMD allows predicting missing values using the similarities between individual and the relationships between variables simultaneously, i.e., between continuous variables, between categorical variables and between variables of different types. The specificity of FAMD lies in weighting each variable in order to balance the influence of each one, while taking into account the type of each variable in the weighting. Thus the imputation method based on this principal component method uses well the structure of the data set to impute it. We begin by presenting the imputation method (Sect. 2) and then illustrate its properties using simulations (Sect. 3). Finally, the method is assessed on real data sets (Sect. 4). The competitiveness of the method is highlighted by comparing its performances to the ones of Stekhoven and Bühlmann (2011) method.

2 Imputation for mixed type-data using the factorial analysis for mixed data

2.1 FAMD in the complete case

FAMD is a principal component method to describe, summarize and visualize a matrix with mixed data. As with any principal component method, its aim is to study the similarities between individuals, the relationships between variables (here continuous and categorical variables) and to link the study of the individuals with that of the variables. Such methods reduce the dimensionality of the data and provide the subspace that best represents the data.

The principle of FAMD is to balance the influence of the continuous and the categorical variables in the analysis. The rationale is to weight the variables in such a way that each variable of either types contributes equally to the construction of the principal components. It is the same idea as scaling for continuous variables in PCA. As mentioned by Benzécri (1973): doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize.

Let us denote I as the number of individuals, K_1 as the number of continuous variables, K_2 as the number of categorical variables and $K = K_1 + K_2$ as the total number of variables. Suppose the first K_1 variables are the continuous ones and the last K_2 variables are the categorical ones. The first step of FAMD consists of coding the categorical variables using the indicator matrix of dummy variables. We denote $\mathbf{X}_{I\times J}$ as the matrix where $\left(\mathbf{x}_j\right)_{1\leq j\leq K_1}$ are continuous variables and $\left(\mathbf{x}_j\right)_{K_1+1\leq j\leq J}$ are dummy variables. The total number of columns is $J = K_1 + \sum_{k=K_1+1}^{K} q_k$ where q_k is the number of categories of the variable k.

FAMD can be represented as the PCA of $\left((\mathbf{X}-\mathbf{M})\mathbf{D}_{\Sigma}^{-1/2}\right)$ where $\mathbf{M}_{I\times J}$ is the matrix with each row being the vector of the means of each column of $\mathbf{X}$ and $\mathbf{D}_{\Sigma}$

is the diagonal matrix $diag\left(\hat{\sigma}^2_{x_1}, \ldots, \hat{\sigma}^2_{x_{K_1}}, p_{K_1+1}, \ldots, p_j, \ldots, p_J\right)$ with $\hat{\sigma}_{x_j}$ being the standard deviation of the continuous variable $\mathbf{x}_j$ and p_j being the proportion of individuals in the category j $(j = K_1 + 1, \ldots, J)$. The matrix $\mathbf{D}_\Sigma$ is the metric used to compute distances between rows. The loss function (known as the reconstruction error) which is minimized in the PCA of matrix $\mathbf{X}$ is:

$$\|\mathbf{X} - \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\|^2 \tag{1}$$

Thus, FAMD can be defined as minimizing:

$$\|(\mathbf{X} - \mathbf{M})\mathbf{D}_\Sigma^{-1/2} - \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\|^2 \tag{2}$$

FAMD provides the best low rank $(S < (J - K_2))$ approximation of the matrix $(\mathbf{X} - \mathbf{M})\mathbf{D}_\Sigma^{-1/2}$ in the least square sense. The solution is given by the singular value decomposition (SVD) of the matrix $(\mathbf{X} - \mathbf{M})\mathbf{D}_\Sigma^{-1/2}$ with $\hat{\mathbf{U}}_{I\times S}$ being the left singular-vectors and $\hat{\mathbf{V}}_{J\times S}$ being the right-singular vectors associated with the S largest singular values gathered in the matrix $\left(\hat{\mathbf{\Lambda}}_{S\times S}\right)^{1/2} = diag\left(\sqrt{\hat{\lambda}_1}, \ldots, \sqrt{\hat{\lambda}_S}\right)$. Notice that the maximum number of non-null eigenvalues is $(J - K_2)$ because of the linear restrictions on the columns for the categorical variables (the row sum for each variable equals 1).

The specific weighting implies that the distances between two individuals i and i' in the initial space (before approximating the distances by keeping the first S dimensions obtained from the SVD) is:

$$d^2\left(i, i'\right) = \sum_{k=1}^{K_1} \frac{(x_{ik} - x_{i'k})^2}{\hat{\sigma}^2_{\mathbf{x}_k}} + \sum_{j=K_1+1}^{J} \frac{1}{p_j}\left(x_{ij} - x_{i'j}\right)^2$$

Weighting by $\frac{1}{\hat{\sigma}^2_{\mathbf{x}_k}}$ ensures that units of continuous variables do not influence the (square) distance between individuals. Furthermore, weighting by $\frac{1}{p_j}$ implies that two individuals in different categories for the same variable are more distant when one of them is in a rare category than when both of them are in frequent categories. The marginal frequencies of the categorical variables play an important role in this method. The frequencies are related to the variance in the initial space, also called inertia, of the category j Escofier (1979): Inertia$(\mathbf{x}_j) = 1 - p_j$. Categories with a small frequency have a greater inertia than the others and consequently rare categories have a greater influence on the construction of the principal components.

The specific weighting implies also that, in FAMD, the principal components maximize the associations with both continuous and categorical variables. More precisely, the first principal component $\mathbf{f}_1$ maximizes

$$\sum_{k=1}^{K_1} R^2\left(\mathbf{x}_k, \mathbf{f}_1\right) + \sum_{k=K_1+1}^{K} \eta^2\left(\mathbf{z}_k, \mathbf{f}_1\right) \tag{3}$$

with $(\mathbf{z}_k)_{k=K+1,\ldots,K}$ the categorical variables. The first principal component is the synthetic variable the most correlated with both the continuous variables in terms of the coefficient of determination (R^2), and the categorical variables in terms of the squared correlation ratio (η^2). The second principal component is the synthetic variable which maximizes the criterion among variables orthogonal to the first principal component, etc.

Regarding criterion (3), we can note that FAMD reduces to PCA when there are only continuous variables and reduces to multiple correspondence analysis Lebart et al. (1984); Greenacre and Blasius (2006) when there are only categorical variables.

2.2 The iterative FAMD algorithm

An approach commonly used to deal with missing values in exploratory data analysis methods (such as PCA) consists of ignoring the missing values by minimizing the reconstruction error over all non-missing elements. For PCA, this can be achieved by introducing a weight matrix $\mathbf{W}$ (with $w_{ij} = 0$ if x_{ij} is missing and $w_{ij} = 1$ otherwise) in the least square criterion (1):

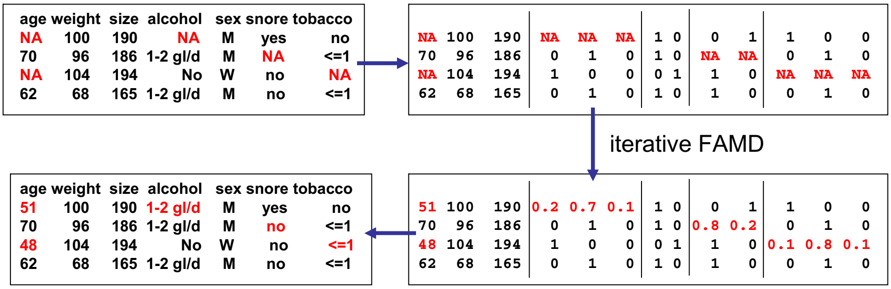
$$\left\| \mathbf{W} * \left(\mathbf{X} - \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}' \right) \right\|^2 \tag{4}$$

with $*$ being the Hadamard product. Different algorithms can be used to minimize this criterion such as an algorithm of iterative imputation of the missing entries during the estimation of the parameters Kiers (1997). The algorithm essentially sets the missing elements at initial values, performs the PCA on the completed data set, imputes the missing values with values predicted by the reconstruction formula (defined by the fitted matrix obtained with the axes and components) using a predefined number of dimensions, and repeats the procedure on the newly obtained matrix until the total change in the matrix falls below an empirically determined threshold. This type of algorithm can thus be seen as a single imputation method and consequently it is a method to impute continuous data based on PCA.

Since FAMD has been presented as the PCA of the matrix $\left((\mathbf{X} - \mathbf{M})\mathbf{D}_\Sigma^{-1/2}\right)$, criterion (2) becomes, with the addition of missing values:

$$\left\| \mathbf{W} * \left((\mathbf{X} - \mathbf{M})\mathbf{D}_\Sigma^{-1/2} - \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}' \right) \right\|^2 \tag{5}$$

The methodology used to impute data with PCA can be extended to FAMD, but the algorithm must be adapted to take into account the specificities of FAMD. More precisely, the same algorithm can be used but the matrix $\mathbf{D}_\Sigma$ as well as the mean matrix $\mathbf{M}$ must be updated during the estimation process because they depend on all the data. Indeed, after imputing data with the reconstruction formula, the variance of the continuous variables as well as the column margins of the categorical variables of the new data table change. It is not recommended to fix $\mathbf{M}$ and $\mathbf{D}_\Sigma$ from the observed data as a first step in the analysis. The estimates could be seriously biased if the mechanism generating missing values is not completely at random Rubin (1976),



age	weight	size	alcohol	sex	snore	tobacco
NA	100	190	NA	M	yes	no
70	96	186	1-2 gl/d	M	NA	<=1
NA	104	194	No	W	no	NA
62	68	165	1-2 gl/d	M	no	<=1

NA	100	190	NA	NA	NA	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	NA	NA	0	1	0
NA	104	194	1	0	0	0	1	1	0	NA	NA	NA
62	68	165	0	1	0	1	0	1	0	0	1	0

51	100	190	0.2	0.7	0.1	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	0.8	0.2	0	1	0
48	104	194	1	0	0	0	1	1	0	0.1	0.8	0.1
62	68	165	0	1	0	1	0	1	0	0	1	0

age	weight	size	alcohol	sex	snore	tobacco
51	100	190	1-2 gl/d	M	yes	no
70	96	186	1-2 gl/d	M	no	<=1
48	104	194	No	W	no	<=1
62	68	165	1-2 gl/d	M	no	<=1

Fig. 1 Diagram for the iterative FAMD algorithm: the raw mixed data, the matrix $\mathbf{X}$, the imputed data obtained by iterative FAMD and the imputed mixed data

meaning that the probability that a value is missing is unrelated to the value itself and any values in the data set, missing or observed. Since $\mathbf{M}$ and $\mathbf{D}_\Sigma$ are estimated during the algorithm, it is not guaranteed that the criterion (5) is minimized with the iterative algorithm which is then merely a heuristic.

The algorithm to impute missing values for mixed data based on FAMD begins as follow. There is initially a table of mixed data with missing values (first table in Fig. 1). This table is then transformed to obtain the matrix $\mathbf{X}$ coding categorical variables using an indicator matrix of dummy variables. A missing value on a categorical variable then leads to a row of missing values in the indicator matrix (second table in Fig. 1). Then this data table is imputed according to the following algorithm. The imputation algorithm, based on FAMD and called iterative FAMD, then proceeds as follows:

1. initialization $\ell = 0$: substitute missing values by initial values and calculate $\mathbf{M}^0$ and $\mathbf{D}_\Sigma^0$ on this completed data set.
2. step ℓ:
   (a) perform the FAMD, in other words the PCA of $\left(\mathbf{X}^{\ell-1} - \mathbf{M}^{\ell-1}\right)\left(\mathbf{D}_\Sigma^{\ell-1}\right)^{-1/2}$ to obtain $\hat{\mathbf{U}}^\ell$, $\hat{\mathbf{V}}^\ell$ and $\left(\hat{\Lambda}^\ell\right)^{1/2}$;
   (b) keep the first S dimensions and use the reconstruction formula to compute the fitted matrix:

$$\hat{\mathbf{X}}^\ell_{I\times J} = \left(\hat{\mathbf{U}}^\ell_{I\times S}\left(\hat{\Lambda}^\ell_{S\times S}\right)^{1/2}\left(\hat{\mathbf{V}}^\ell_{J\times S}\right)'\right)\left(\left(\mathbf{D}_\Sigma^{\ell-1}\right)_{I\times J}\right)^{1/2} + \mathbf{M}^{\ell-1}_{I\times J}$$

   and the new imputed data set becomes $\mathbf{X}^\ell = \mathbf{W} * \mathbf{X} + (\mathbf{1} - \mathbf{W}) * \hat{\mathbf{X}}^\ell$ with $\mathbf{1}_{I\times J}$ being a matrix with only ones. The observed values are the same but the missing ones are replaced by the fitted values;
   (c) from the new completed matrix $\mathbf{X}^\ell$, $\mathbf{D}_\Sigma^\ell$ and $\mathbf{M}^\ell$ are updated.
3. steps (2.a), (2.b) and (2.c) are repeated until the change in the imputed matrix falls below a predefined threshold $\sum_{ij}(\hat{x}_{ij}^{\ell-1} - \hat{x}_{ij}^\ell)^2 \leq \varepsilon$, with ε equals to 10^{-6} for example.

In the initialization step, missing values are replaced, for instance, by the mean of the variable for the continuous variables and the marginal proportion of the category for each category using the non-missing entries. Note that for the categorical variables, the sum of the initial entries corresponding to one individual and one categorical variable must equal one. At the end of the algorithm, imputed values for the missing entries for the categories are not equal to 0 and 1 but are real numbers (third table in Fig. 1). However, the constraint in the initialization step ensures that the sum of the entries for one individual and one categorical variable is equal to 1. This property comes from the specific weighting and is demonstrated in the framework of multiple correspondence analysis Tenenhaus and Young (1985); Josse et al. (2012). The same proof hold in the framework of FAMD. Consequently, the imputed values can be considered as degrees of membership to the associated category and it is possible to impute the categorical variable with the most plausible value (last table in Fig. 1).

Such algorithms which alternate a step of estimation of the parameters via a singular value decomposition and a step of imputation of the missing values are known to suffer from overfitting problems. These problems occur even if these methods reduce the dimensionality of the data. A fortiori, when there are missing values, the overfitting is more marked which means that imputed values are over-dispersed. Yet, the more missing values there are, the less trusted the relationships between variables are. Consequently, it would be more satisfactory to impute using the mean imputation than using a model taking into account the relationships between variables. Geometrically, mean imputation is equivalent to bring the individuals closer from the center of gravity of the point cloud, which limits the dispersion. In order to avoid these major problems of overfitting, the iterative singular value decomposition algorithm has been replaced by an iterative thresholded singular value decomposition algorithm in the framework of PCA Josse and Husson (2012); Mazumder et al. (2010). It is the same rationale as in ridge regression where the regularized version stabilizes the prediction in comparison with the ordinary least squares. We follow the approach proposed by Josse and Husson (2012) and define a regularized iterative FAMD algorithm by replacing the singular values $\left(\sqrt{\hat{\lambda}_s^{\ell}}\right)_{s=1,\ldots,S}$ of step (2.b) by $\left(\frac{\hat{\lambda}_s^{\ell}-\hat{\sigma}^2}{\sqrt{\hat{\lambda}_s^{\ell}}}\right)_{s=1,\ldots,S}$ with $\hat{\sigma}^2 = \sum_{s=S+1}^{J-K_2} \frac{\lambda_s}{J-K_2-S}$. The rationale is to remove the noise in order to avoid instabilities in the prediction. Implicitly, it is assumed that the first S dimensions contain both information and noise whereas the last ones contain only noise; hence, the variance of the noise is estimated by the mean of the last eigenvalues. The regularized method shrinks the first S singular values with a greater amount of shrinkage for the smallest ones, which is acceptable since these small singular values are more responsible for instability. When the noise is small, the regularized algorithm is quite similar to the non-regularized one. When the variance of the noise is large, $(\hat{\lambda}_s - \hat{\sigma}^2)$ tends to zero, and the regularized algorithm simply imputes continuous variables with their means and imputes categorical variables with their marginal proportions. This is acceptable because when the data are overwhelmed by noise, the structure of the data (the relationship between variables) is very weak and imputing with the mean of the variables is an effective strategy. Thus the regularized method provides imputed values less dispersed than for the non-regularized one. The regularized iterative FAMD algorithm is also a heuristic since no explicit penalized loss function is optimized.

Note that FAMD is close to another extension of the PCA for mixed data, called nonlinear PCA. When there are only categorical data, nonlinear PCA without restrictions (also known as homogeneity analysis) is equivalent to Multiple Correspondence Analysis. Nonlinear PCA allows the analysis of variables of different nature via for instance the rank-1 restrictions [see, Michailidis and de Leeuw (1998)]. There are solutions to deal with missing values in the Gifi system such as the 'missing data passive' approach Gifi (1990). The rationale for categorical variables van der Heijden and Escofier (2003) is based on the following assumption: if an individual i is missing on the variable j, one considers that the individual has not chosen any category for the variable. Consequently, in the indicator matrix, the entries in the row corresponding to individual i and variable j are marked 0. Josse et al. (2012) compared thoroughly this approach to the equivalent of iterative FAMD for categorical variables and showed that the "missing data passive" method is equivalent to adding a new category for the missing values and then to putting it as supplementary element. Consequently the rationale of both approaches is different: in the former, when there is a missing value for a categorical variable, one considers that the individual has not chosen any category for the variable whereas in the latter, one considers that a missing value represents an underlying category among the available categories. In the Gifi system, another method is available to deal with missing values named "missing multiple" which consists in adding a new category for each missing values. van der Heijden and Escofier (2003) described the issues raised by this approach. Note that in the Gifi system, no attempts have been made in order to use the approach as a possible imputation method for mixed data even if the extension could be imagined.

3 Properties of the imputation method

We discuss the main properties of the new imputation method and illustrate those properties on different toy data sets. We focus on the regularized version of the algorithm in order to avoid overfitting problem. In addition this version converges faster, which is more convenient to perform simulations. We compare the imputation results obtained with regularized iterative FAMD algorithm to the ones obtained with the method based on random forest Stekhoven and Bühlmann (2011) to highlight some important properties. In the latter method, some parameters such as the number of trees per forest, the number of variables selected for each forest as well as minimum size of terminal nodes can be tuned. However, Stekhoven and Bühlmann (2011) modify these parameters mainly to minimize computational time, finding that modifying these parameters offers little improvement of the imputations themselves. Consequently, the parameters suggested by default in their implementation of the method were used in the simulations.

Simulation process

We simulate toy data sets which differ with respect to the number of continuous variables, the number of categorical variables, the number of categories per variable, the number of individuals per category, the number of underlying dimensions and the strength of the relationship between variables through different signal to noise ratios Mazumder et al. (2010). The signal to noise ratios (SNR) is the square root of the ratio between the variance of the signal and the variance of the noise. In the particular case of continuous data with variance equal to 1, the variance of the signal corresponds to the number of variables and the variance of the noise corresponds to σ^2 times this one. Consequently the SNR is simply the inverse of σ. Thus, a high SNR implies that the variables are very correlated, whereas a low SNR implies that the data are very noisy. More precisely, the toy data sets are almost all simulated according to the following procedure:

- S' independent variables are drawn from a standard Gaussian distribution;
- each variable s' (for $s' = 1, \ldots, S'$) is replicated $K^{s'}$ times which guarantees (in expectation) S' orthogonal groups of correlated variables. S' will be called the number of underlying dimensions;
- Gaussian noise is added with different levels of variance to obtain different signal to noise ratios;
- categorical variables are obtained by splitting continuous variables in equal count categories.

Then, we insert different percentages of missing values (10, 20 and 30 %) completely at random Rubin (1976). The code to reproduce all the simulations is available on the webpage of the first author. For each set of parameters, 200 simulations are performed.

Criteria

Two criteria are used to assess the quality of the imputation, the proportion of falsely classified (PFC) entries for categorical variables and the normalized root mean squared error (NRMSE) for continuous data:

$$\text{NRMSE} = \sqrt{\frac{\sum_{k=1}^{K_1}\sum_{i=1}^{I}(1-w_{ik})\left(\frac{x_{ik}-\hat{x}_{ik}}{\hat{\sigma}_{\mathbf{x}_k}}\right)^2}{\sum_{k=1}^{K_1}\sum_{i=1}^{I}(1-w_{ik})}}$$

NRMSE allows the consideration of variables with different variances. Moreover, when NRMSE equals zero, the imputation is perfect, whereas when it is close to 1, the imputation yields results similar to those obtained using the mean imputation.

3.1 Relationships between continuous and categorical variables

The fundamental objective of FAMD is to take into account relationships between continuous and categorical variables. Taking into account both types of variables improves the imputation as is illustrated with a data set that has two underlying dimensions ($S' = 2$). Each dimension is composed of two continuous variables and two categorical variables with four categories. Missing data are then added completely at random for the three selected percentages and finally the imputation algorithm is performed according to three strategies:

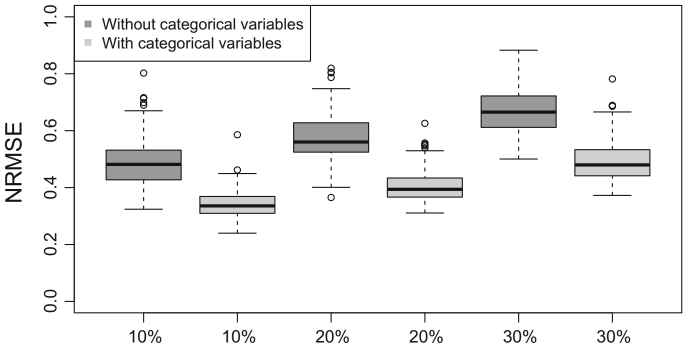


Fig. 2 Distribution of the NRMSE for different amounts of missing values (10, 20, 30 %). *Dark grey boxplots* correspond to the error of imputation for continuous variables when only continuous variables are used whereas *grey boxplots* correspond to the error when categorical and continuous variables are used

1. using only continuous variables, which leads to an imputation of continuous variables with only those variables;
2. using only categorical variables, which leads to an imputation of categorical variables with only those variables;
3. using and imputing variables of both kinds.

Figure 2 compares the distributions of the NRMSE for the three percentages of missing data according to the strategies (1) and (3). As expected, when the proportion of missing data increases, the imputation error is larger. When the SNR decreases, the quality of the imputation (not shown here) decreases which is also expected. It can be noted that even with 30 % missing data the imputation with iterative FAMD greatly outperforms the mean imputation (NRMSE less than 1). This is due to the relationships between variables, which improve the mean imputation that forms the first step of iterative FAMD. The imputation error is lower when considering both types of variables (boxplots in grey) than when considering only continuous variables (boxplots in dark grey). Taking into account categorical variables thus improves the imputation of continuous variables. This behavior is the same regardless of the proportion of missing data.

Figure 3 compares the distributions of rates of misclassification according to the strategies (2) and (3). The results of the categorical variables are similar to those obtained for continuous variables: when the rate of missing data increases, the proportion of misclassification increases, but even for 30 % missing data the imputation by FAMD yields better results than a random imputation (the latter having error 0.66). Regardless of the rate of missing data, taking into account the continuous variables for the imputation of categorical variables (light grey boxplots) reduce the proportion of misclassification.

Remark It is possible in theory to perfectly impute a categorical variable using a continuous variable. On the contrary, it is difficult to impute the continuous variables with only categorical variables. For example, using K_2 categorical variables with q categories each can produce only q^{K_2} distinct imputations. Consequently, this imputation cannot reflect all possible values that the continuous variable can take. However, in practice it can be a reasonable imputation.

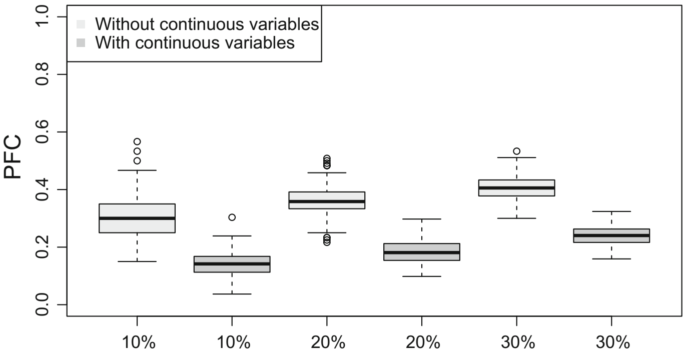


Fig. 3 Distribution of the PFC for different amounts of missing values (10, 20, 30 %). *Light grey boxplots* correspond to the error of imputation for categorical variables when only categorical variables are used, whereas *grey boxplots* correspond to the error when categorical and continuous variables are used

3.2 Influence of the relationships between variables

3.2.1 Linear and nonlinear relationships

FAMD can be seen as a variation of PCA, and PCA is based on linear relationships between variables. When there are strong linear relationships between continuous variables, the imputation of these variables with iterative FAMD will thus be accurate. To illustrate this behavior, a data set is generated according to the simulation process with $S' = 1$ initial variable from which 2 continuous variables and 3 categorical variables with 4 categories are built. The imputation by FAMD (Fig. 4, boxplots in grey) is compared to the imputation based on random forest (Fig. 4, boxplots in white). The error for continuous variables as well as the error for categorical variables with iterative FAMD is very small and is much smaller than the error of the algorithm based on random forest. Moreover, when the percentage of missing values increases, the error of the iterative FAMD algorithm increases slightly, whereas the error of the algorithm based on random forest increases more. Such results are representative of all the results obtained with different data sets.

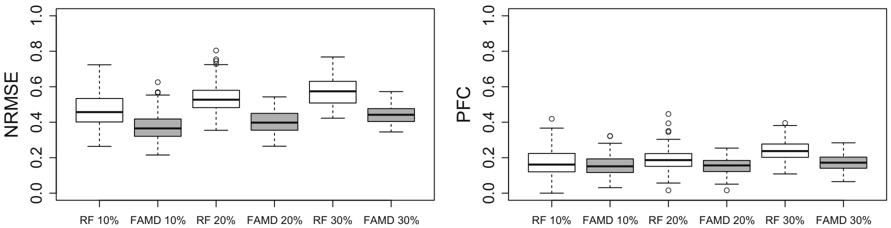


Fig. 4 Distribution of the NRMSE (*left*) and of the PFC (*right*) when the relationships between variables are linear for different amounts of missing values (10, 20, 30 %). *White boxplots* correspond to the imputation error for the algorithm based on random forest (RF) and *grey boxplots* to the imputation error for iterative FAMD

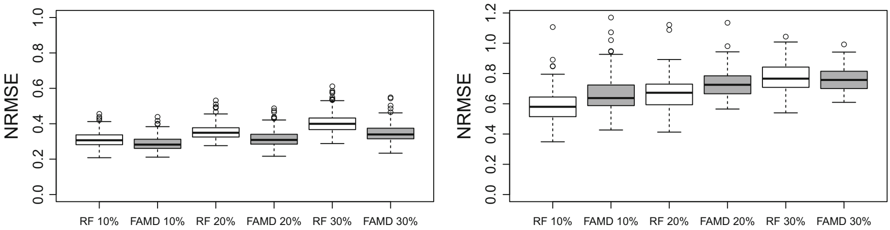


Fig. 5 Distribution of the NRMSE when the relationships between variables are linear (*left*) and nonlinear (*right*) for different amounts of missing values (10, 20, 30 %). *White boxplots* correspond to the imputation error for the algorithm based on random forest (RF), and *grey boxplots* to the imputation error for iterative FAMD

What about nonlinear relationships in FAMD? Thanks to the presence of categorical variables, FAMD may impute missing values when there are nonlinear relationships between the continuous variables. Indeed, the principal components of FAMD are linear combinations of the continuous variables and of the columns of the indicator matrix (Eq. 3). A linear combination of dummy variables may approximate a nonlinear function of a variable by a piecewise constant function. To illustrate this behavior, a data set is generated with $S' = 1$ initial variable from which 3 continuous variables and 1 categorical variable with 10 categories are built. The results of applying FAMD to this data, illustrated in the left panel of Fig. 5 are in accordance with the previous ones: the imputation with FAMD is very accurate when there are linear relationships. Then we take the same data set but the second continuous variable is squared and the cosine function is applied to the third variable. In this case, the results obtained by iterative FAMD are worse than those obtained by the algorithm based on random forest (Fig. 5, graph on the right), which is known to deal well with nonlinear relationships. However, the difference in performance is modest.

Remark In practice, if nonlinear relationships between continuous variables are suspected, a solution can be to create new categorical variables by discretizing the continuous variables into categories.

3.2.2 Taking into account interactions between categorical variables

Other relationships between variables can be challenging. FAMD is based on relationships between pairs of variables. Consequently data including complex interactions could make the imputation difficult. To illustrate this behavior, a data set with 3 variables (1 continuous and 2 categorical) illustrated in Table 1 is constructed. It consists of a fractional factorial design 3^{3-1} that is replicated several times (with replications vertically stacked). Variables are pairwise independent but there are interactions between them.

The quality of the imputation of the continuous and categorical variables is poor with iterative FAMD (Fig. 6). It is similar to the mean imputation for the continuous variable and to the imputation by the proportion of the categories for the categorical variables. The imputation based on random forest takes into account the interactions

Table 1 Data set with interaction generated with a fractional factorial design 3^{3-1}; the defining relation of the fractional design is $I = 123$

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	*a*	*a*	1
2	*b*	*a*	2
3	*c*	*a*	3
4	*a*	*b*	2
5	*b*	*b*	3
6	*c*	*b*	1
7	*a*	*c*	3
8	*b*	*c*	1
9	*c*	*c*	2
10	*a*	*a*	1
11	*b*	*a*	2
12	*c*	*a*	3
13	*a*	*b*	2
14	*b*	*b*	3
15	*c*	*b*	1
16	*a*	*c*	3
17	*b*	*c*	1
18	*c*	*c*	2
…			

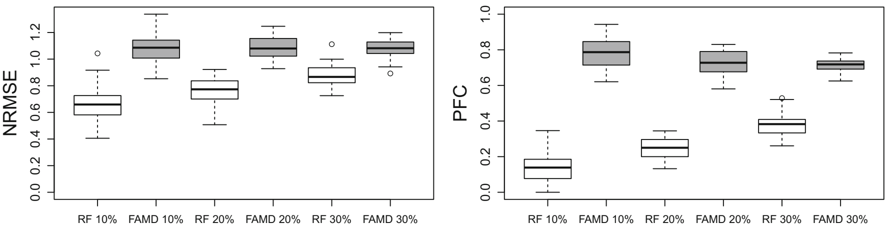


Fig. 6 Distribution of the NRMSE (*left*) and of the PFC (*right*) when there are interactions between variables. Results are given for different amounts of missing values (10, 20, 30 %). *White boxplots* correspond to the imputation error for the algorithm based on random forest (RF) and *grey boxplots* to the imputation error for iterative FAMD

between variables and provides better results. This relatively worse performance of FAMD-based imputation can be seen as a drawback. However, we can address this problem by introducing an additional variable in the data set that corresponds to the interaction, for example, by creating a variable x_4 which has 9 levels “aa”, “ba”, “ca”, etc. The imputation problem thus reduces to the case without interaction and the quality of the imputation will be very good.

Table 2 Percentage of error (PFC) over 1000 simulations when recovering a rare category for data sets with different numbers of individuals and different frequencies for the rare category (f)

Number of individuals	f (%)	FAMD	Random forest
100	10	0.060	0.096
100	4	0.082	0.173
1000	10	0.042	0.041
1000	4	0.060	0.071
1000	1	0.074	0.167
1000	0.4	0.107	0.241

Results are given for the imputation with FAMD and with the algorithm based on random forest

3.3 Imputation of rare categories

FAMD performed on a complete data set weights each category by the inverse of the number of individuals taking this category (Sect. 2.1). Thus FAMD assigns more variance to rare categories both in the cloud of categories as well as in the cloud of individuals in the initial space. Consequently, rare categories are privileged when constructing the principal components and because the algorithm uses the principal components to impute the data, rare categories may be well predicted.

In order to illustrate the ability of the method to impute rare categories, simulations have been conducted with 2 continuous variables, 3 categorical variables and 3 categories per categorical variables. The frequencies of the categories are respectively equal to $(1/3, 1/3, 1/3)$ for one categorical variable and to $(f, (1-f)/2, (1-f)/2)$ for the 2 other categorical variables, with f being a frequency that varies between 0.004 and 0.1. The rare categories (each having frequency f) of these 2 categorical variables are related in the sense that they are taken by the same individuals. Then a rare value is suppressed for one individual on one of the two categorical variables and the imputation algorithms are performed. This strategy allows us only to focus on the prediction of a rare category. Simulations are performed for different numbers of individuals and, for a given frequency f, we expect that it would be easier to recover the true category when the number of individuals is large because the category contains more individuals. The results (Table 2) show that the algorithm based on FAMD successfully recovers the rare category. The advantage of FAMD over the algorithm based on random forest is especially large when imputing a rare category.

3.4 Extensive study

Here, we assess the influence of other parameters like size, the balance between number of categorical and continuous variables and the influence of the number of categories of the categorical variables. We generate data sets using 2 underlying dimensions and vary the number of individuals (50 or 200), the number of variables (12 or 24), the proportion of continuous variables (1/3 or 2/3), the number of categories per variable (3 or 6), as well as the signal to noise ratio (2 or 4). On each data set, 10 % of missing

Table 3 Mean of NRMSE and PFC for the FAMD algorithm and for the method based on random forests

Case	I	K	$K1/K$	q	SNR	FAMD PFC	RF PFC	FAMD NRSME	RF NRMSE
1	50	12	0.667	3	2	**0.171**	**0.14**	**0.325**	**0.437**
2	50	12	0.667	3	4	**0.061**	**0.079**	**0.247**	**0.318**
3	50	12	0.667	6	2	0.498	0.451	0.408	0.425
4	50	12	0.667	6	4	0.125	0.192	0.229	0.318
5	50	12	0.333	3	2	0.107	0.168	0.402	0.471
6	50	12	0.333	3	4	0.035	0.045	0.398	0.404
7	50	12	0.333	6	2	**0.267**	**0.314**	**0.403**	**0.454**
8	50	12	0.333	6	4	0.076	0.124	0.294	0.368
9	50	24	0.667	3	2	0.149	0.16	0.285	0.373
10	50	24	0.667	3	4	0.023	0.064	0.204	0.254
11	50	24	0.667	6	2	0.258	0.296	0.364	0.367
12	50	24	0.667	6	4	0.039	0.071	0.211	0.276
13	50	24	0.333	3	2	0.093	0.136	0.371	0.394
14	50	24	0.333	3	4	0.015	0.021	0.379	0.334
15	50	24	0.333	6	2	**0.08**	**0.133**	**0.4**	**0.412**
16	50	24	0.333	6	4	0.018	0.027	0.296	0.324
17	200	12	0.667	3	2	0.165	0.167	0.26	0.309
18	200	12	0.667	3	4	0.074	0.066	0.203	0.19
19	200	12	0.667	6	2	**0.311**	**0.332**	**0.246**	**0.287**
20	200	12	0.667	6	4	0.122	0.109	0.154	0.185
21	200	12	0.333	3	2	0.094	0.126	0.363	0.371
22	200	12	0.333	3	4	0.027	0.039	0.358	0.295
23	200	12	0.333	6	2	**0.16**	**0.224**	**0.267**	**0.307**
24	200	12	0.333	6	4	0.047	0.059	0.232	0.238
25	200	24	0.667	3	2	0.088	0.092	0.22	0.259
26	200	24	0.667	3	4	0.039	0.043	0.181	0.18
27	200	24	0.667	6	2	0.16	0.195	0.212	0.247
28	200	24	0.667	6	4	0.053	0.067	0.142	0.179
29	200	24	0.333	3	2	0.07	0.075	0.335	0.327
30	200	24	0.333	3	4	0.022	0.024	**0.347**	**0.25**
31	200	24	0.333	6	2	0.085	0.103	0.266	0.299
32	200	24	0.333	6	4	0.038	0.04	**0.234**	**0.236**

Results are obtained over 200 simulations for 10 % of missing values, varying the number of individuals (I), the number of variables (K), the proportion of continuous variables ($K1/K$), the number of categories per variable (q), as well as the signal to noise ratio (SNR). The results in bold are detailed in the text and are characteristic of the general trends of the imputation methods

values are added, then the FAMD algorithm and the one based on random forests are performed. Results over 200 replications are gathered in Table 3.

As expected, when the number of data increases (i.e. when I or K increases), the imputation error decreases for the two algorithms (compare for example the mean errors on the cases 7 and 23, as well as the mean error on the cases 7 and 15). In addition, when the proportion of continuous variables increases, the NRMSE decreases for continuous variables and the PFC increases for categorical variables (see for example the cases 19 and 23). This common behaviour for the two algorithms illustrates the difficulty in taking into account the links between variables of different types: an imputation between variables of a same type is easier than between variables of different types. Then the NRMSE decreases when the number of categories increases (compare for example the cases 30 and 32). This was expected since the information carried by the categorical variable is finer when the number of categories increases, which allows to better impute continuous variables. Note that the PFC increases mechanically because the probability to make a mistake is higher. Finally, the error decreases when the SNR decreases (see the cases 1 and 2 for example). Comparing to the algorithm based on random forests, most of the time, the FAMD algorithm provides better imputations for the two types of variables.

3.5 Choice of the number of dimensions

At each iteration of the iterative FAMD algorithm, data are reconstructed using only the S first dimensions (step 2.b). If S is too small then relevant information is lost and cannot be used for the imputation. On the other hand, if S is too large then noise is considered as signal, which may lead to instability of the imputations. The number S is thus an important parameter of the algorithm, and has to be chosen *a priori*. In this section we focus on choosing this number from an incomplete mixed data set.

First of all, we can note that a categorical variable with q_k categories evolves within a space with $(q_k - 1)$ dimensions. Therefore, it is impossible to predict the values of the categories with a choice of S less than $(q_k - 1)$. This may be a clue that guiding the choice of S. However, some of these $(q_k - 1)$ dimensions may be unrelated to all the other variables, especially when they are many categories. In this case, even if S is large, it is impossible to impute this variable correctly and choosing many dimensions may lead to instability.

Many strategies are available in the literature to select a number of dimensions from a complete data set in PCA. Cross-validation Bro et al. (2008) or an approximation of cross-validation such as generalized cross-validation Josse and Husson (2011), for example, perform well. These methods have the advantage that they can be directly extended to incomplete data. Since their extension is straightforward for FAMD, in practice we use cross-validation to select the number of dimensions. However, this topic may deserve more research. Consequently, it is important to assess the impact of a poor choice for the number of dimensions on the results.

We consider a data set generated from $S' = 2$ groups of orthogonal variables ($K^{1'} = 8$ variables, 4 continuous and 4 categorical variables, and $K^{2'} = 4$ variables, 2 continuous and 2 categorical variables). Each categorical variable has 3 categories. Consequently the underlying number of dimensions of this data set is $S = 4$. Figure 7 shows the evolution of the average of the errors over 200 simulations according to S for the continuous variables (graphs on the left) and for the categorical variables (graphs on the right). Structured data sets (SNR = 3) are used in the graphs on the top whereas noisy data sets (SNR = 1) are used in the graphs on the bottom.

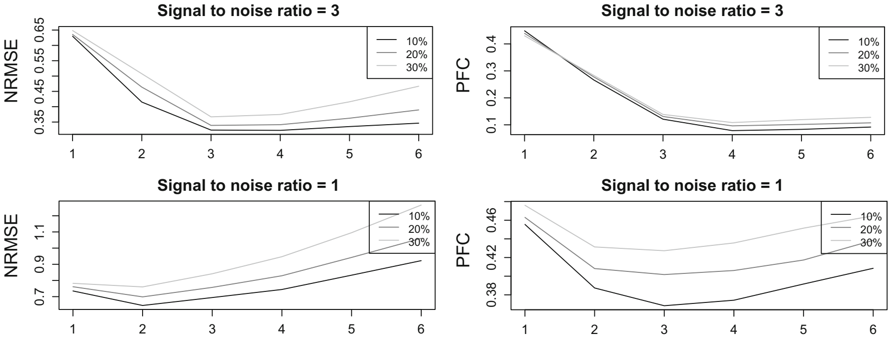


Fig. 7 Average error of imputation over 200 simulations according to the number of dimensions used in the algorithm and for 3 amounts of missing values (10, 20, 30 %): error for the continuous variables on the *left* and for the categorical variables on the *right*. The signal to noise ratio equals 3 for the simulations represented on the *top*, and 1 for the simulations represented on the *bottom*

When the signal to noise ratio is high, the error for the categorical variables decreases until it reaches the optimal value 4 and then it increases slowly, regardless of the percentage of missing values. The same comment can be made for the continuous variables even if the minimum is reached for $S = 3$, since the continuous variables are only linked to the first 3 dimensions. These results were expected. However, the behavior is very different when the signal to noise ratio is small. Indeed, even for the categorical variables, it is preferable to choose fewer than the true underlying number of dimensions. This is especially true as the percentage of missing data increases. It arises because selecting a smaller number of dimensions can be regarded as performing stronger regularization which is a good strategy when the data are very noisy.

These simulations show that it is never preferable to consider more dimensions than the true number but it may be preferable to consider fewer. A good strategy is thus to use cross-validation which, in practice, produces satisfactory results in the sense that it finds the “best” number of dimensions to impute the data set: it favors the true number of dimensions when there are strong relationships between the variables and a smaller number when the data are very noisy.

4 Comparison on real data sets

The imputation method is evaluated on real data sets that cover many situations.The regularised imputation method is evaluated on real data sets that cover many situations. They differ in terms of number of individuals, number of variables, number of categories for the categorical variables, and they represent different areas of application. Missing values are added at random to these complete sets and then the imputation is performed with iterative FAMD and with Stekhoven and Bühlmann (2011) algorithm. Each configuration is simulated 200 times for three different percentages of missing data. The number of dimensions for the reconstruction step of the iterative FAMD algorithm is determined by cross-validation. This number is held constant for the 200 simulations in order to save computational time. The evaluation is based on the following mixed data sets.

Tips This data set, from the package rggobi Lang et al. (2012) of the R software R Development Core Team (2011), concerns the tips given to a waiter in a restaurant in the US in the early 1990s. The $K = 8$ variables of the data set concern the price of the meal for $I = 244$ customers, on the tip amount and the conditions of the restaurant meal (number of guests, time of day, etc.). There are $K_1 = 3$ continuous variables and $K_2 = 5$ categorical variables with between 2 and 6 levels.

BMI This data set Lafaye de Micheaux et al. (2011) concerns body mass index of $I = 152$ French children aged 3 to 4 years. The $K = 6$ variables concern their morphology and the characteristics of their kindergarten ($K_1 = 4$, $K_2 = 2$). All the categorical variables have two levels.

Ozone This data set Cornillon et al. (2012) contains $I = 112$ daily measurements of meteorological variables (wind speed, temperature, rainfall, etc.) and ozone concentration recorded in Rennes (France) during summer 2001. There are $K_1 = 11$ continuous variables and $K_2 = 2$ categorical variables with 2 or 4 levels.

German Breast Cancer Study Group (GBSG) This data set, from the package ipred (Peters and Hothorn, 2012) of the R software, described $I = 686$ women with breast cancer using variables concerning the status of the tumours and the hormonal system of the patient ($K_1 = 7$, $K_2 = 3$). Categorical variables have between 2 or 3 levels.

Imputation results for all the data sets are presented in Fig. 8. The graphs on the left evaluate the quality of the imputation for the continuous variables (NRMSE) whereas the graphs on the right evaluate the quality of the imputation for the categorical variables. In general, the iterative FAMD provides a slightly better imputation than the one obtained by the algorithm based on random forest. However, the results obtained by the latter algorithm are often better for continuous variables. On the Ozone and BMI data sets, the difference between the errors reached 5 %. However, imputation with the iterative FAMD is more efficient on categorical variables. For the data sets on tips and BMI, the difference between the two methods is 5 %. Note that the NRMSE errors may be close to 1. This means that the imputation methods improve on the mean imputation but the gain is small.

These conclusions extend to the case of non-mixed data sets. We now consider two of them: one continuous (Parkinson) and one categorical (Credit) data set.

Parkinson This data set Stekhoven and Bühlmann (2011) contains $K = 22$ measurements on the voice of $I = 195$ patients with or without Parkinson's disease. The response categorical variable sick/healthy is excluded for these simulations.

Credit This data set Cornillon et al. (2012) concerns $I = 66$ customers profiles of subscribers to consumer credit in a bank. The $K = 11$ variables include the financial conditions under which the customer subscribes to the credit as well as some socio-demographic characteristics. The number of levels for these variables is between 2 and 5.

For imputation of the continuous data set (Parkinson), the random forest algorithm outperforms FAMD, with a difference of about 10 % in the NRMSE (Fig. 9 on the left). We thus find the same results about continuous variables as those for mixed-type data. This is accentuated by the fact that relationships between variables are here

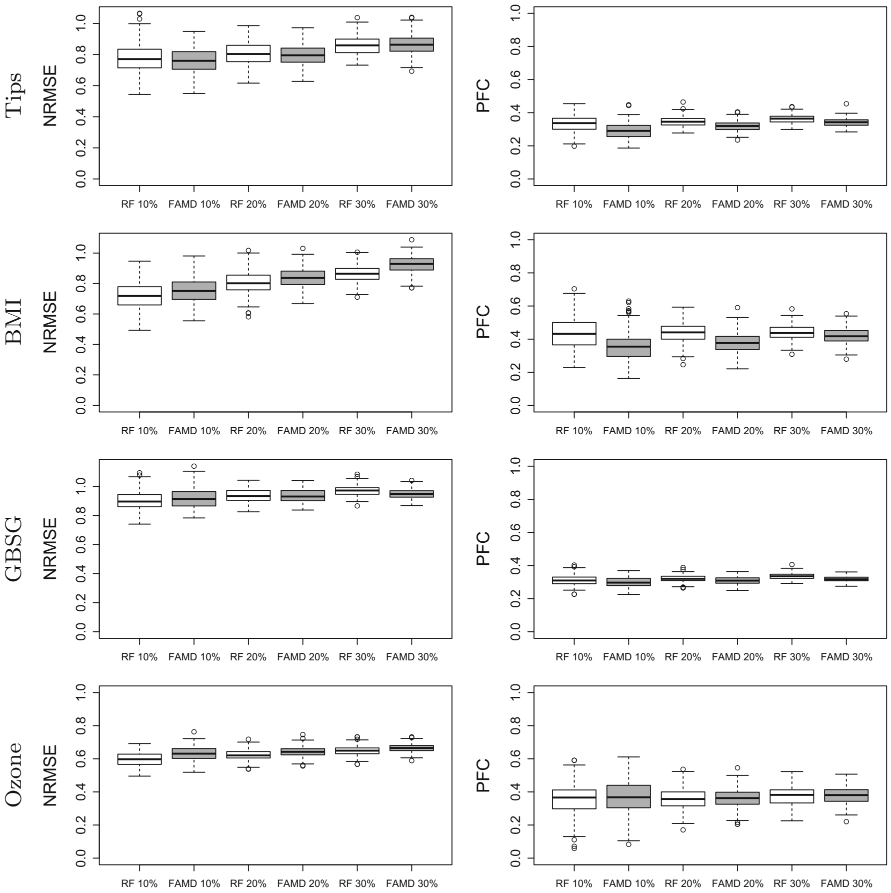


Fig. 8 Distribution of the NRMSE (*left*) and of the PFC (*right*) for different amounts of missing values (10, 20, 30 %) and for different data sets (Tips, BMI, GBSG, Ozone). *White boxplots*correspond to the imputation error of the algorithm based on random forest (RF) and *grey boxplots* correspond to the imputation error of iterative FAMD

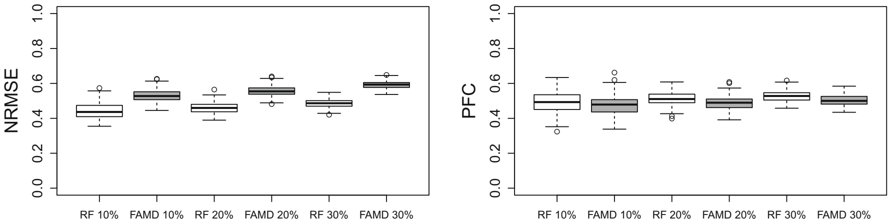


Fig. 9 Distribution of the error for 10, 20, 30 % of missing values for the data sets Parkinson (*left*) and Credit (*right*). *White boxplots* correspond to the imputation error of the algorithm based on random forest (RF) and *grey boxplots* correspond to the error of iterative FAMD

nonlinear. The lack of categorical variables implies that it is not possible to impute these variables correctly with iterative FAMD. For the categorical data set (Credit), the errors (Fig. 9 on the right) are smaller with the iterative FAMD algorithm, regardless of the percentage of missing values. These results are also similar to those observed in the mixed-type data.

As mentioned in Sect. 2.2, an approach close to FAMD is available. Consequently, to get hints of the potential of an imputation approach using the missing passive method of Gifi in the framework of nonlinear PCA, we performed the simulations on the real data sets. We used the function homals from the package Homals de Leeuw and Mair (2009) of the free R software. Note that this function performs nonlinear PCA with missing values using the missing data passive approach but does not give as an output a completed dataset. Consequently, we had to modify the code to obtain an imputed mixed data matrix. We performed this method taking first the parameters by default (rank 1 restriction) on all the simulations of the real data sets. These first attempts show very bad results compared to the two other methods especially on the continuous variables. Different combinations of the parameters (restrictions, transformation) could be considered. It can be investigated for future research, but it is out of the scope of this paper.

5 Conclusion

Imputing mixed data is very challenging and very few methods are proposed in the literature. The new imputation method proposed in this paper is based on a principal component method, the factorial analysis for mixed data, and allows imputation of missing data that simultaneously takes into account similarities between individuals and the associations between all variables, continuous variables and categorical variables. This method produces particularly good predictions for the missing entries of categorical variables and when there are linear relationships between variables. A strong point is that rare categories are well imputed. In addition, the method provides good results both in terms of quality of the imputation and computational time compared to the best method available based on random forests.

The iterative FAMD algorithm requires a tuning parameter which is the number of dimensions used to reconstruct the data. Cross-validation, while time-consuming, can in practice be used to select this parameter. Approximation of cross-validation such as generalized cross-validation could be proposed to select this number without resorting to an intensive computational method.

The imputation method based on FAMD is implemented in the package `missMDA` Husson and Josse (2012) of the `R` software. The function `imputeFAMD` takes as input the incomplete data set and the number of dimensions used to reconstruct the data at each step of the algorithm. The function returns a table with the imputed mixed data as well as the table concatenating the imputed continuous variables and the imputed indicator matrix.

As with all methods of imputation, imputation quality deteriorates with increasing percentage of missing data. However, this deterioration depends on the structure of the data set. Indeed, if the variables are uncorrelated, even a single missing value is problematic. On the other hand, if the variables are highly correlated, very little data per individual are sufficient to impute the data set. For this reason, it is of course not possible to offer a percentage of missing data below which imputation is acceptable and above which the imputation is no longer satisfactory. It would therefore be desirable to provide confidence intervals around the imputed values. We expect narrow confidence intervals when the data are highly correlated, indicating higher confidence in the imputed values.

The proposed method is a method of single imputation. Like any single imputation method, it is limited because it does not take into account the uncertainty associated with the prediction of missing values based upon observed values. Thus, if we apply a statistical method on the completed data table, the variability of the estimators will be underestimated. To avoid this problem, a solution is to perform multiple imputation Little and Rubin (1987, 2002). In this case, different values are predicted for each missing value, which leads to several imputed data sets; the variability across the imputations reflects the variance of the prediction of each missing entry. The second step of multiple imputation consists of performing the statistical analysis on each completed data set. The third step combines the results to obtain the estimators of the parameters and of their variability taking into account uncertainty due to missing data. The proposed iterative FAMD imputation algorithm could be a first step in a multiple imputation method for mixed data.

References

Benzécri JP (1973) L'analyse des données. L'analyse des correspondances. Dunod, Tome II

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Bro R, Kjeldahl K, Smilde AK, Kiers HAL (2008) Cross-validation of component model: a critical look at current methods. Anal Bioanal Chem 390:1241–1251

Cornillon PA, Guyader A, Husson F, Jégou N, Josse J, Kloareg M, Matzner-Løber E, Rouvière L (2012) R for Statistics. Chapman and Hall/CRC, Boca Raton

de Leeuw J, Mair P (2009) Gifi methods for optimal scaling in R: The package homals. J Statist Software 31(4):1–20, URL http://www.jstatsoft.org/v31/i04/

Escofier B (1979) Traitement simultané de variables quantitatives et qualitatives en analyse factorielle. Les cahiers de l'analyse des données 4(2):137–146

Gifi A (1990) Nonlinear multivariate analysis. Wiley, Chichester

Greenacre M, Blasius J (2006) Multiple correspondence analysis and related methods. Chapman and Hall/CRC.

Husson F, Josse J (2012) missMDA: Handling missing values with/in multivariate data analysis (principal component methods). URL http://www.agrocampus-ouest.fr/math/husson, r package version 1.4

Ilin A, Raiko T (2010) Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res 99:1957–2000, URL http://dl.acm.org/citation.cfm?id=1859890.1859917

Josse J, Husson F (2011) Selecting the number of components in PCA using cross-validation approximations. Comput Statist Data Anal 56(6):1869–1879

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

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

Josse J, Chavent M, Liquet B, Husson F (2012) Handling missing values with regularized iterative multiple correspondence analysis. J Classif 29:91–116

Kiers HAL (1991) Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika 56:197–212

Kiers HAL (1997) Weighted least squares fitting using ordinary least squares algorithms. Psychometrika 62:251–266

Lafaye de Micheaux P, Drouilhet R, Liquet B (2011) Le logiciel R. Springer, Paris

Lang DT, Swayne D, Wickham H, Lawrence M (2012) rggobi: Interface between R and GGobi. URL http://CRAN.R-project.org/package=rggobi, r package version 2.1.19

Lebart L, Morineau A, Werwick KM (1984) Multivariate descriptive statistical analysis. Wiley, New York

Little RJA, Rubin DB (1987, 2002) Statistical analysis with missing data. Wiley series in probability and statistics, New York

Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 11:2287–2322

Michailidis G, de Leeuw J (1998) The Gifi system of descriptive multivariate analysis. Statist Sci 13(4):307–336

Peters A, Hothorn T (2012) ipred: Improved Predictors. URL http://CRAN.R-project.org/package=ipred, R package version 0.9-1

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/, ISBN 3-900051-07-0

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman and Hall/CRC, London

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

Tenenhaus M, Young FW (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50:91–119

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17(62001):520–525

van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Statist Method Med Res 16:219–242

van Buuren S, Boshuizen H, Knook D (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Statist Med 18:681–694

van der Heijden P, Escofier B (2003) Multiple correspondence analysis with missing data. In: Analyse des correspondances, Presse universitaire de Rennes, pp 153–170

Vermunt JK, van Ginkel JR, van der Ark LA, Sijtsma K (2008) Multiple imputation of incomplete categorical data using latent class analysis. Sociol Methodol 33:369–397