

MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis

VINCENT AUDIGIER¹, FRANÇOIS HUSSON² AND JULIE JOSSE²

Applied Mathematics Department, Agrocampus Ouest, 65 rue de Saint-Brieuc,
F-35042 RENNES Cedex, France

audigier@agrocampus-ouest.fr

husson@agrocampus-ouest.fr

josse@agrocampus-ouest.fr

Abstract

We propose a multiple imputation method to deal with incomplete categorical data. This method imputes the missing entries using the principal components method dedicated to categorical data: multiple correspondence analysis (MCA). The uncertainty concerning the parameters of the imputation model is reflected using a non-parametric bootstrap. Multiple imputation using MCA (MIMCA) requires estimating a small number of parameters due to the dimensionality reduction property of MCA. It allows the user to impute a large range of data sets. In particular, a high number of categories per variable, a high number of variables or a small the number of individuals are not an issue for MIMCA. Through a simulation study based on real data sets, the method is assessed and compared to the reference methods (multiple imputation using the loglinear model, multiple imputation by logistic regressions) as well to the latest works on the topic (multiple imputation by random forests or by the Dirichlet process mixture of products of multinomial distributions model). The proposed method shows good performances in terms of bias and coverage for an analysis model such as a main effects logistic regression model. In addition, MIMCA has the great advantage that it is substantially less time consuming on data sets of high dimensions than the other multiple imputation methods.

Keywords: missing values, categorical data, multiple imputation, multiple correspondence analysis, bootstrap

1 Introduction

Data sets with categorical variables are ubiquitous in many fields such in social sciences, where surveys are conducted through multiple-choice questions. Whatever the field, missing values frequently occur and are a key problem in statistical practice since most of statistical methods cannot be applied directly on incomplete data.

To deal with missing values one solution consists in adapting the statistical method so that it can be applied on an incomplete data set. For instance, the maximum likelihood (ML) estimators can be derived from incomplete data using an Expectation-Maximization (EM) algorithm [1] and their standard error can be estimated using a Supplemented Expectation-Maximization algorithm [2]. The ML approach is suitable, but not always easy to establish [3].

Another way consists in replacing missing values by plausible values according to an *imputation model*. This is called *single imputation*. Thus, the data set is complete and any statistical method can be applied on this one. Figure 1 illustrates three simple single imputation methods. The data set used contains 1000 individuals and two variables with

¹Principal corresponding author

²Corresponding author

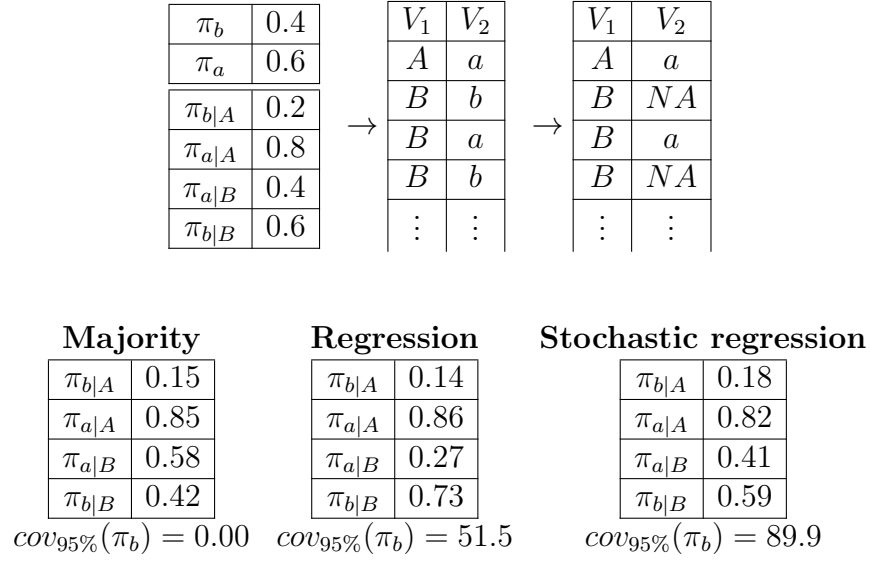


Figure 1: Illustration of three imputation methods for two categorical variables: the top part described how the data are built (marginal and conditional proportions, associated complete data, incomplete data set generated where NA denotes a missing value) and the bottom part sums up the observed conditional proportions after an imputation by several methods (majority, regression, stochastic regression). The last line indicates the coverage for the confidence interval for the proportion of b over 1000 simulations.

two categories: A and B for the first variable, a and b for the second one. The data set is built so that 40% of the individuals take the category a and 60% the category b . In addition, the variables are linked, that is to say, the probability to observe a or b on the second variable depends on the category taken on the first variable. Then, 30% of missing values are generated completely at random on the second variable. A first method could be to impute according to the most taken category of the variable. In this case, all missing values are imputed by a . Consequently, marginal proportions are modified, as well as conditional proportions (see the bottom part of Figure 1). This method is clearly not suitable. A more convenient solution consists in taking into account the relationship between the two variables, following the rationale of the imputation by regression for continuous data. To achieve this goal, the parameters of a logistic regression are estimated from the complete cases, providing fitted conditional proportions. Then, each individual is imputed according to the highest conditional proportion given the first variable. This method respects the conditional proportions better, but the relationship between variables is strengthened which is not satisfactory. In order to obtain an imputed data set with a structure as close as possible to the generated data set, a suitable single imputation method is to perform stochastic regression: instead of imputing according to the the most likely category, the imputation is performed randomly according to the fitted probabilities.

An imputation model used to perform single imputation has to be sufficiently complex compared to the statistical method desired (the *analysis model*). For instance, if the aim is to apply a logistic regression from an incomplete data set, it requires using an imputation model taking into account the relationships between variables. Thus, a suitable single imputation method, such as the stochastic regression strategy, leads to unbiased estimates of the parameters of the statistical method (see Figure 1). However, although the single imputation

method respects the structure of the data, it still has the drawback that it leads to underestimate the variability of the estimators because the uncertainty on the imputed values is not taken into account in the estimate of the variability of the estimators. However, although the single imputation method respects the structure of the data, it still has the drawback that the uncertainty on the imputed values is not taken into account in the estimate of the variability of the estimators. Thus, this variability remains underestimated. For instance, in Figure 1, the level of the confidence interval of π_b , the proportion of b , is 89.9% and does not reach the nominal rate of 95%.

Multiple imputation (MI) [4, 5] has been developed to avoid this issue. The principle of multiple imputation consists in creating M imputed data sets to reflect the uncertainty on imputed values. Then, the parameters of the statistical method, denoted ψ , are estimated from each imputed data set, leading to M sets of parameters $(\hat{\psi}_m)_{1 \leq m \leq M}$. Lastly, these sets of parameters are pooled to provide a unique estimation for ψ and for its associated variability using Rubin's rules [4].

MI is based on the *ignorability* assumption, that is to say ignoring the mechanism that generated missing values. This assumption is equivalent to: first, the parameters that govern the missing data mechanism and the parameters of the analysis model are independent; then, missing values are generated *at random*, that is to say, the probability that a missing value occurs on a cell is independent from the value of the cell itself. In practice, ignorability and value missing at random (MAR), are used interchangeably. This assumption is more plausible when the number of variables is high [6, 7], but remains difficult to verify.

Thus, under the ignorability assumption, the main challenge in multiple imputation is to reflect the uncertainty of the imputed values by reflecting *properly* [4, p. 118-128] the uncertainty on the parameters of the model used to perform imputation to get imputed data sets yielding to valid statistical inferences. To do so, two classical approaches can be considered. The first one is the Bayesian approach: a prior distribution is assumed on the parameters θ of the imputation model, it is combined with the observed entries, providing a posterior distribution from which M sets of parameters $(\tilde{\theta}_m)_{1 \leq m \leq M}$ are drawn. Then, the incomplete data set is imputed M times using each set of parameters. The second one is a bootstrap approach: M samples with replacement are drawn leading to M incomplete data sets from which the parameters of the imputation model are obtained. The M sets of parameters $(\theta_m)_{1 \leq m \leq M}$ are then used to perform M imputations of the original incomplete data set.

In this paper, we detail in Section 2 the main available MI methods to deal with categorical data. Two general modelling strategies can be distinguished for imputing multivariate data: joint modelling (JM) [6] and fully conditional specification (FCS)[8]. JM is based on the assumption that the data can be described by a multivariate distribution. Concerning FCS, the multivariate distribution is not defined explicitly, but implicitly through the conditional distributions of each variable only. Among the presented methods, three are JM methods: MI using the loglinear model, MI using the latent class model and MI using the normal distribution; the two others are FCS strategies: the FCS using logistic regressions and FCS using random forests [9]. In Section 3, a novel JM method based on a principal components method dedicated to categorical data, namely multiple correspondence analysis (MCA), is proposed. Principal components methods are commonly used to highlight the similarities between individuals and the relationships between variables, using a small number of principal components and loadings. MI based on this family of methods uses these similarities and these relationships to perform imputation, while using a restricted number of parameters. The performances of the imputation are very promising from continuous data [10, 11] which

motivates the consideration of a method for categorical data. In Section 4, a simulation study based on real data sets, evaluates the novel method and compares its performances to other main multiple imputation methods. Lastly, conclusions about MI for categorical data and possible extensions for the novel method are detailed.

2 Multiple imputation methods for categorical data

The imputation of categorical variables is rather complex. Indeed, contrary to continuous data, the variables follow a distribution on a discrete support defined by the combinations of categories observed for each individual. Because of the explosion of the number of combinations when the number of categories increases, the number of parameters defining the multivariate distribution could be extremely large. Consequently, defining an imputation model is not straightforward for categorical data. In this section we review the most popular approaches commonly used to deal with categorical data: JM using the loglinear model, JM using the latent class model, JM using the normal distribution and FCS using multinomial logistic regression or random forests.

Hereinafter, matrices and vectors will be in bold text, whereas sets of random variables or single random variables will not. Matrices will be in capital letters, whereas vectors will be in lower case letters. We denote $\mathbf{X}_{I \times K}$ a data set with I individuals and K variables. We note the observed part of \mathbf{X} by \mathbf{X}_{obs} and the missing part by \mathbf{X}_{miss} , so that $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$. Let q_k denote the number of categories for the variable \mathbf{X}_k , $J = \sum_{k=1}^K q_k$ the total number of categories. We note $\mathbb{P}(X, \theta)$ the distribution of the variables $X = (X_1, \dots, X_K)$, where θ is the corresponding set of parameters.

2.1 Multiple imputation using a loglinear model

The saturated loglinear model (or multinomial model) [12] consists in assuming a multinomial distribution $\mathcal{M}(\theta, 1)$ as joint distribution for X , where $\theta = (\theta_{x_1 \dots x_K})_{x_1 \dots x_K}$ is a vector indicating the probability to observe each event $(X_1 = x_1, \dots, X_K = x_K)$. Performing MI with the loglinear model [6] is often achieved by reflecting the variability of the imputation model's parameters with a Bayesian approach. More precisely, a Bayesian treatment of this model can be specified as follows:

$$X|\theta \sim \mathcal{M}(\theta, 1) \tag{1}$$

$$\theta \sim \mathcal{D}(\alpha) \tag{2}$$

$$\theta|X \sim \mathcal{D}(\alpha + \hat{\theta}^{ML}) \tag{3}$$

where $\mathcal{D}(\alpha)$ denotes the Dirichlet distribution with parameter α , a vector with the same dimension as θ and $\hat{\theta}^{ML}$ is the maximum likelihood for θ , corresponding to the observed proportions of each combination in the data set. A classical choice for α is $\alpha = (1/2, \dots, 1/2)$ corresponding to the non-informative Jeffreys prior [13]. Combining the prior distribution and the observed entries, a posterior distribution for the model's parameters is obtained (Equation (3)).

Because missing values occur in the data set, the posterior distribution is not tractable, therefore, drawing a set of model's parameters in it is not straightforward. Thus, a data-augmentation algorithm [14] is used. In the first step of the algorithm, missing values are imputed by random values. Then, because the data set is now completed, a draw of θ in the posterior distribution (3) can easily be obtained. Next, missing values are imputed

from the predictive distribution (1) using the previously drawn parameter and the observed values. These steps of imputation and draw from the posterior distribution are repeated until convergence. At the end, one set of parameters $\hat{\theta}_m$, drawn from the observed posterior distribution, is obtained. Repeating the procedure M times in parallel, M sets of parameters are obtained from which multiple imputation can be done. In this way, the uncertainty on the parameters of the imputation model is reflected, insuring a proper imputation.

The loglinear model is considered as the gold standard for MI of categorical data [15]. Indeed, this imputation model reflects all kind of relationships between variables, which enables applying any analysis model. However, this method is dedicated to data sets with a small number of categories because it requires a number of independent parameters equal to the number of combinations of categories minus 1. For example, it corresponds to 9 765 624 independent parameters for a data set with $K = 10$ variables with $q_k = 5$ categories for each of them. This involves two issues: the storage of θ and overfitting. To overcome these issues, the model can be simplified by adding constraints on θ . The principle is to write $\log(\theta)$ as a linear combination of a restricted set of parameters $\lambda = [\lambda_0, \lambda_{x_1}, \dots, \lambda_{x_K}, \dots, \lambda_{x_1 x_2}, \dots, \lambda_{x_1 x_K}, \dots, \lambda_{x_{K-1} x_K}]$, where each element is indexed by a category or a couple of categories. More precisely, the constraints on θ are given by the following equation:

$$\log(\theta_{x_1 \dots x_K}) = \lambda_0 + \sum_k \lambda_{x_k} + \sum_{\substack{(k,k') \\ k \neq k'}} \lambda_{x_k x_{k'}} \text{ for all } (X_1 = x_1, \dots, X_K = x_K) \quad (4)$$

where the second sum is the sum over all the couples of categories possible from the set of categories (x_1, \dots, x_K) . Thus, the imputation model reflects only the simple (two-way) associations between variables, which is generally sufficient. Equation (4) leads to 760 independent parameters for the previous example. However, although it requires a smaller number of parameters, the imputation under the loglinear model still remains difficult in this case, because the data-augmentation algorithm used [6, p.320] is based on a modification of θ at each iteration and not of λ . Thus the storage issue remains.

2.2 Multiple imputation using a latent class model

To overcome the limitation of MI using the loglinear model, another JM method based on the latent class model can be used. The latent class model [12, p.535] is a mixture model based on the assumption that each individual belongs to a latent class from which all variables can be considered as independent. More precisely, let Z denote the latent categorical variable whose values are in $\{1, \dots, L\}$. Let $\theta_Z = (\theta_\ell)_{1 \leq \ell \leq L}$ denote the proportion of the mixture and $\theta_X = (\theta_x^{(\ell)})_{1 \leq \ell \leq L}$ the parameters of the L components of the mixture. Thus, let $\theta = (\theta_Z, \theta_X)$ denote the parameters of the mixture, the joint distribution of the data is written as follows:

$$\mathbb{P}(X = (x_1, \dots, x_K); \theta) = \sum_{\ell=1}^L \left(\mathbb{P}(Z = \ell, \theta_Z) \prod_{k=1}^K \mathbb{P}(X_k = x_k | Z = \ell; \theta_x^{(\ell)}) \right) \quad (5)$$

Assuming a multinomial distribution for Z and $X|Z$, Equation (5), can be rewritten as follows:

$$\mathbb{P}(X = (x_1, \dots, x_K); \theta) = \sum_{\ell=1}^L \left(\theta_\ell \prod_{k=1}^K \theta_{x_k}^{(\ell)} \right) \quad (6)$$

The latent class model requires $L \times (J - K) + (K - 1)$ independent parameters, *i.e.* a number that linearly increases with the number of categories.

[16] reviews in detail different multiple imputation methods using a latent class model. These methods can be distinguished by the way used to reflect the uncertainty on the parameters of the imputation model and by the way that the number of components of the mixture is chosen: automatically or *a priori*. The quality of the imputation is quite similar from one method to another, the main differences remain in computation time. One of the latest contributions in this family of methods uses a non-parametric extension of the model namely the Dirichlet process mixture of products of multinomial distributions model (DPMPM) [17, 18]. This method uses a fully Bayesian approach in which the number of classes is defined automatically and is not too computationally intensive. DPMPM assumes a prior distribution on $\theta_Z = (\theta_\ell)_{1 \leq \ell \leq L}$ without fixing the number of classes which is supposed to be infinite. More precisely, the prior distribution for θ_Z is defined as follows:

$$\theta_\ell = \zeta_\ell \prod_{g < \ell} (1 - \zeta_g) \text{ for } \ell \text{ in } 1, \dots, \infty \quad (7)$$

$$\zeta_\ell \sim \mathcal{B}(1, \alpha) \quad (8)$$

$$\alpha \sim \mathcal{G}(.25, .25) \quad (9)$$

where \mathcal{G} refers to the gamma distribution, α is a positive real number, \mathcal{B} refers to the beta distribution; the prior distribution for θ_X is defined by:

$$\theta_x^{(\ell)} \sim \mathcal{D}(1, \dots, 1) \quad (10)$$

corresponding to a uniform distribution over the simplex defined by the constraint of sum to one. The posterior distribution of θ is not analytically tractable, even when no missing value occur. However, the distribution of each parameter is known if the others are given. For this reason, a Gibbs sampler is used to obtain a draw from the posterior distribution. The principle of this is to draw each parameter while fixing the others. From an incomplete data set, missing values require to be preliminarily imputed. More precisely, a draw from the posterior distribution is obtained as follows: first, the parameters and missing values are initialized; then, given the current parameters, particularly θ_Z and θ_X , each individual is randomly affected to one class according to its categories; next, each parameter (θ_Z , θ_X , α) is drawn conditionally to the others; finally, missing values are imputed according to the mixture model. These steps are then repeated until convergence (for more details, see [18]).

Despite the infinite number of classes, the prior on θ_ℓ typically implies that the posterior distribution for θ_ℓ is non negligible for a finite number of classes only. Moreover, for computational reasons, the number of classes has to be bounded. Thus, [18] recommends to fix the maximum number of latent classes to twenty. Consequently, the simulated values of θ are some realisations of an approximated posterior distribution only.

Multiple imputation using the latent class model has the advantages and drawbacks of this model: because the latent class model approximates quite well any kind of relationships between variables, MI using this model enables the use of complex analysis models such as logistic regression with some interaction terms and provides good estimates of the parameters of the analysis model. However, the imputation model implies that given a class, each individual is imputed in the same way, whatever the categories taken. If the class is very homogeneous, all the individuals have the same observed values, and this behaviour makes sense. However, when the number of missing values is high and when the number of variables is high, it is not straightforward to obtain homogeneous classes. It can explain why [16] observed that the multiple imputation using the latent class model can lead to biased estimates for the analysis model in such cases.

2.3 Multiple imputation using a multivariate normal distribution

Another popular strategy to perform MI for categorical data is to adapt the methods developed for continuous data. Because multiple imputation using the normal multivariate distribution is a robust method for imputing continuous non-normal data [6], imputation using the multivariate normal model is an attractive method for this. The principle consists in recoding the categorical variables as dummy variables and applying the multiple imputation under the normal multivariate distribution on the recoded data. The imputed dummy variables are seen as a set of latent continuous variables from which categories can be independently derived. More precisely, let $\mathbf{Z}_{I \times J}$ denote the disjunctive table coding for $\mathbf{X}_{I \times K}$, *i.e.*, the set of dummy variables corresponding to the incomplete matrix. Note that one missing value on \mathbf{x}_k implies q_k missing values for \mathbf{z}_k . The following procedure implemented in [19, 20] enables the multiple imputation of a categorical data set using the normal distribution:

- perform a non-parametric bootstrap on \mathbf{Z} : sample the rows of \mathbf{Z} with replacement M times. M incomplete disjunctive tables $(\mathbf{Z}_m^{boot})_{1 \leq m \leq M}$ are obtained;
- estimate the parameters of the normal distribution on each bootstrap replicate: calculate the ML estimators of (μ_m, Σ_m) , the mean and the variance of the normal distribution for the m^{th} bootstrap incomplete replicate, using an EM algorithm. Note that the set of M parameters reflects the uncertainty required for a proper multiple imputation method;
- create M imputed disjunctive tables: impute \mathbf{Z} from the normal distribution using $(\mu_m, \Sigma_m)_{1 \leq m \leq M}$ and the observed values of \mathbf{Z} . M imputed disjunctive tables $(\mathbf{Z}_m)_{1 \leq m \leq M}$ are obtained. In \mathbf{Z}_m , the observed values are still zeros and ones, whereas the missing values have been replaced by real numbers;
- create M imputed categorical data sets: from the latent continuous variables contained in $(\mathbf{Z}_m)_{1 \leq m \leq M}$, derive categories for each incomplete individual.

Several ways have been proposed to get the imputed categories from the imputed continuous values. For example [21] recommends to attribute the category corresponding to the highest imputed value, while [22–24] propose some rounding strategies. However, “*A single best rounding rule for categorical data has yet to be identified.*” [25, p. 107]. A common one proposed by [22] is called *Coin flipping*. Coin flipping consists in considering the set of imputed values of the q_k dummy variables \mathbf{z}_k as an expectation given the observed values $\theta_k = \mathbb{E} \left[(z_1, \dots, z_{q_k}) | Z_{obs}; \hat{\mu}, \hat{\Sigma} \right]$. Thus, randomly drawing one category according to a multinomial distribution $\mathcal{M}(\theta_k, 1)$, suitably modified so that θ_k remains between 0 and 1, imputes plausible values. The values lower than 0 are replaced by 0 and the imputed values higher than 1 are replaced by 1. In this case, the imputed values are scaled to respect the constraint of sum to one.

Because imputation under the normal multivariate distribution is based on the estimate of a covariance matrix, the imputation under the normal distribution can detect only two-way associations between categorical variables. In addition, this method assumes independence between categories conditionally to the latent continuous variables. This implies that if two variables are linked, and if an individual has missing values on these ones, then the categories derived from the imputed disjunctive table will be drawn independently. Consequently, the two-way associations can not be perfectly reflected in the imputed data set. Note that, contrary to the MI using the latent class, the parameter of the multinomial distribution θ_k is specific to each individual, because the imputation of the disjunctive table is performed

given the observed values. This behaviour makes sense if the variables on which missing values occur are linked with the others. The main drawback of the MI using the normal distribution is the number of independent parameters estimated. This number is equal to $\frac{(J-K) \times (J-K+1)}{2} + (J-K)$, representing 860 parameters for a data set with 10 variables with 5 categories. It increases rapidly when the total number of categories (J) increases, leading quickly to overfitting. Moreover, the covariance matrix is not invertible when the number of individuals is lower than $(J-K)$. To overcome these issues, it is possible to add a ridge term on its diagonal to improve the conditioning of the regression problem.

2.4 Fully conditional specification

Categorical data can be imputed using a FCS approach instead of a JM approach: for each variable with missing values, an imputation model is defined, (*i.e.* a conditional distribution), and each incomplete variable is sequentially imputed according to this, while reflecting the uncertainty on the model's parameters. Typically, the models used for each incomplete variable are some multinomial logistic regressions and the variability of the models parameter is reflected using a Bayesian point of view. More precisely, we denote by $\theta_k = (\theta_{k\ell})_{1 \leq \ell \leq q_k}$ the set of parameters for the multinomial distribution of the variable to impute X_k (the set of the other variables is denoted X_{-k}). We also denote by $\beta_k = (\beta_{k1}, \dots, \beta_{kL})$ the set of regression parameters that defines θ_k , such as $\beta_{k\ell}$ is the regression parameter vector associated with the category ℓ of the response variable X_k and \mathbf{Z}_k is the design matrix associated. Note that identifiability constraints are required on β_k , that is why β_{kL} is fixed to the null vector. Thus, the imputation is built on the following assumptions:

$$X_k | \theta_k \sim \mathcal{M}(\theta_k, 1) \quad (11)$$

$$\theta_{k\ell} = \mathbb{P}(X_k = \ell | X_{-k}, \beta) = \frac{\exp(\mathbf{Z}_k \beta_{k\ell})}{1 + \sum_{\ell=1}^{L-1} \exp(\mathbf{Z}_k \beta_{k\ell})} \quad (12)$$

$$\beta | X \sim \mathcal{N}(\hat{\beta}, \hat{V}) \quad (13)$$

where $\hat{\beta}, \hat{V}$ are the estimators of β and of its associated variance. For simplicity, suppose that the data set contains 2 binary variables \mathbf{x}_1 and \mathbf{x}_2 , with \mathbf{x}_2 as incomplete and \mathbf{x}_1 as complete. To impute \mathbf{x}_2 given \mathbf{x}_1 the first step is to estimate β and its associated variance using complete cases by iteratively reweighted least squares. Then, a new parameter $\hat{\beta}_k$ is drawn from a normal distribution centred in the previous estimate with the covariance matrix previously obtained. Lastly, the fitted probability θ_k are obtained from the logistic regression model with parameter $\hat{\beta}_k$ and \mathbf{x}_2 is imputed according to a multinomial distribution with parameters θ_k [25, p.76]. Note that β is drawn in an approximated posterior distribution. Indeed, as explained by [4, p.169-170], the posterior distribution has not a neat form for reasonable prior distributions. However, on a large sample, assuming a weak prior on β , the posterior distribution can be approximated by a normal distribution. Thus, draw β in a normal distribution with $\hat{\beta}$ and \hat{V} as parameters makes sense.

In the general case, where the data set contains K variables with missing values, each variable is imputed according to a multinomial logistic regression given all the others. More precisely, the incomplete data set is firstly randomly imputed. Then, the missing values of the variable \mathbf{x}_k are imputed as explained previously: a value of β_k is drawn from the approximated posterior distribution and an imputation according to $\mathbb{P}(X_k | X_{-k}; \theta_k)$ is performed. The next incomplete variable is imputed in the same way given the other variables, and particularly from the new imputed values of \mathbf{x}_k . We proceed in this way for all variables and repeat it

until convergence, this provides one imputed data set. The procedure is performed M times in parallel to provide M imputed data sets.

Implicitly, the choices of the conditional distributions $\mathbb{P}(X_k|X_{-k};\theta_k)$ determine a joint distribution $\mathbb{P}(X_k;\theta)$, in so far as a joint distribution is compatible with these choices [26]. The convergence to the joint distribution is often obtained for a low number of iterations (5 can be sufficient), but [25, p.113] underlines that this number can be higher in some cases. In addition, FCS is more computationally intensive than JM [15, 25]. This is not a practical issue when the data set is small, but it becomes so on a data set of high dimensions. In particular, checking the convergence becomes very difficult.

The imputation using logistic regressions on each variable performs quite well, that is why this method is often used as a benchmark to perform comparative studies [7, 18, 27, 28]. However, the lack of multinomial regression can affect the multiple imputation procedure using this model. Indeed, when separability problems occur [29], or when the number of individuals is smaller than the number of categories [12, p.195], it is not possible to get the estimates of the parameters. In addition, the number of parameters is very large when the number of categories per variable is high, implying overfitting when the number of individuals is small. When the number of categories becomes too large, [25, 30] advise to use a method dedicated to continuous data: the predictive mean matching (PMM). PMM treats each variable as continuous variables, predicts them using linear regression, and draws one individual among those the nearest to the predicted value. However, PMM often yields to biased estimates [7].

Typically, the default models selected for each logistic regression are main effects models. Thus, the imputation model captures the two-way associations between variables well, which is generally sufficient for the analysis model. However, models taking into account interactions can be used but the choice of these models requires a certain effort by the user. To overcome this effort, in particular when the variables are numerous, conditional imputations using random forests instead of logistic regression have been proposed [27, 28]. According to [27], an imputation of one variable X_k given the others is obtained as follows:

- draw 10 bootstrap samples from the individuals without missing value on X_k ;
- fit one tree on each sample: for a given bootstrap sample, draw randomly a subset of $\sqrt{K-1}$ variables among the $K-1$ explanatory variables. Build one tree from this bootstrap sample and this subset of explanatory variables. A random forest of 10 trees is obtained. Note that the uncertainty due to missing values is reflected by the use of one random forest instead of a unique tree;
- impute missing values on X_k according to the forest: for an individual i with a missing value on X_k , gather all the donors from the 10 predictive leaves from each tree and draw randomly one donor from it.

Then, the procedure is performed for each incomplete variable and repeated until convergence. Using random forests as conditional models allows capturing complex relationships between variables. In addition, the method is very robust to the number of trees used, as well as to the number of explanatory variables retained. Thus, the default choices for these parameters (10 trees, $\sqrt{K-1}$ explanatory variables) are very suitable in most of the cases. However, the method is more computationally intensive than the one based on logistic regressions.

3 Multiple Imputation using multiple correspondence analysis

This section deals with a novel MI method for categorical data based on multiple correspondence analysis (MCA) [31, 32], *i.e.* the principal components method dedicated for categorical data. Like the imputation using the normal distribution, it is a JM method based on the imputation of the disjunctive table. We first introduce MCA as a specific singular value decomposition on specific matrices. Then, we present how to perform this SVD with missing values and how it is used to perform single imputation. We explain how to introduce uncertainty to obtain a proper MI method. Finally, the properties of the method are discussed and the differences with MI using the normal distribution highlighted.

3.1 MCA for complete data

MCA is a principal components method to describe, summarise and visualise multidimensional matrices with categorical data. This powerful method allows us to understand the two-way associations between variables as well as the similarities between individuals. Like any principal components method, MCA is a method of dimensionality reduction consisting in searching for a subspace of dimension S providing the best representation of the data in the sense that it maximises the variability of the projected points (*i.e.* the individuals or the variables according to the space considered). The subspace can be obtained by performing a specific singular value decomposition (SVD) on the disjunctive table.

More precisely, let $\mathbf{Z}_{I \times J}$ denote the disjunctive table corresponding to $\mathbf{X}_{I \times K}$. We define a metric between individuals through the diagonal matrix $\frac{1}{K}\mathbf{D}_\Sigma^{-1}$ where $\mathbf{D}_\Sigma = \text{diag}(\mathbf{p}_1^{\mathbf{x}_1}, \dots, \mathbf{p}_{q_1}^{\mathbf{x}_1}, \dots, \mathbf{p}_1^{\mathbf{x}_K}, \dots, \mathbf{p}_{q_K}^{\mathbf{x}_K})$ is a diagonal matrix with dimensions $J \times J$, $p_\ell^{\mathbf{x}_k}$ is the proportion of observations taking the category ℓ on the variable \mathbf{x}_k . In this way, two individuals taking different categories for the same variable are more distant from the others when one of them takes a rare category than when both of them take frequent categories. We also define a uniform weighting for the individuals through the diagonal matrix $\frac{1}{I}\mathbb{1}_I$ with $\mathbb{1}_I$ the identity matrix of dimensions I . By duality, the matrices $\frac{1}{K}\mathbf{D}_\Sigma^{-1}$ and $\frac{1}{I}\mathbb{1}_I$ define also a weighting and a metric for the space of the categories respectively. MCA consists in searching a matrix $\hat{\mathbf{Z}}$ with a lower rank S as close as possible to the disjunctive table \mathbf{Z} in the sense defined by these metrics. Let $\mathbf{M}_{I \times J}$ denote the matrix where each row is equal to the vector of the means of each column of \mathbf{Z} . MCA consists in performing the SVD of the matrix triplet $(\mathbf{Z} - \mathbf{M}, \frac{1}{K}\mathbf{D}_\Sigma^{-1}, \frac{1}{I}\mathbb{1}_I)$ [33] which is equivalent to writing $(\mathbf{Z} - \mathbf{M})$ as

$$\mathbf{Z} - \mathbf{M} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top \quad (14)$$

where the columns of $\mathbf{U}_{I \times J}$ are the left singular vectors satisfying the relationship $\mathbf{U}^\top \text{diag}(\mathbf{1}/\mathbf{I}, \dots, \mathbf{1}/\mathbf{I})\mathbf{U} = \mathbb{1}_J$; columns of $\mathbf{V}_{J \times J}$ are the right singular vectors satisfying the relationship $\mathbf{V}^\top \frac{1}{K}\mathbf{D}_\Sigma^{-1}\mathbf{V} = \mathbb{1}_J$ and $\mathbf{\Lambda}_{J \times J}^{1/2} = \text{diag}(\lambda_1^{1/2}, \dots, \lambda_J^{1/2})$ is the diagonal matrix of the singular values.

The S first principal components are given by $\hat{\mathbf{U}}_{I \times S}\hat{\mathbf{\Lambda}}_{S \times S}^{1/2}$, the product between the first columns of \mathbf{U} and the diagonal matrix $\mathbf{\Lambda}^{1/2}$ restricted to its S first elements. In the same way, the S first loadings are given by $\hat{\mathbf{V}}_{J \times S}$. $\hat{\mathbf{Z}}$ defined by:

$$\hat{\mathbf{Z}} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^\top + \mathbf{M} \quad (15)$$

is the best approximation of \mathbf{Z} , in the sense of the metrics, with the constraint of rank S (Eckart-Young theorem [34]). Equation (15) is called *reconstruction formula*.

Note that, contrary to \mathbf{Z} , $\hat{\mathbf{Z}}$ is a fuzzy disjunctive table in the sense that its cells are real numbers and not only zeros and ones as in a classic disjunctive table. However, the sum per variable is still equal to one [35]. Most of the values are contained in the interval $[0, 1]$ or close to it because $\hat{\mathbf{Z}}$ is as close as possible to \mathbf{Z} which contains only zeros and ones, but values out of this interval can occur.

Performing MCA requires $J - K$ parameters corresponding to the terms useful for the centering and the weighting of the categories, $IS - S - \frac{S(S+1)}{2}$ for the centered and orthonormal left singular vectors and $(J - K)S - S - \frac{S(S+1)}{2}$ for the orthonormal right singular vectors, for a total of $J - K + S(I - 1 + (J - K) - S)$ independent parameters. This number of parameters increases linearly with the number of values in the data set.

3.2 Single imputation using MCA

[36] proposed an iterative algorithm called “iterative MCA” to perform single imputation using MCA. The main steps of the algorithm are as follows:

1. initialization $\ell = 0$: recode \mathbf{X} as disjunctive table \mathbf{Z} , substitute missing values by initial values (the proportions) and calculate \mathbf{M}^0 and \mathbf{D}_Σ^0 on this completed data set.

2. step ℓ :

- (a) perform the MCA, in other words the SVD of $\left(\mathbf{Z}^{\ell-1} - \mathbf{M}^{\ell-1}, \frac{1}{K} (\mathbf{D}_\Sigma^{\ell-1})^{-1}, \frac{1}{I} \mathbb{1}_I\right)$ to obtain $\hat{\mathbf{U}}^\ell$, $\hat{\mathbf{V}}^\ell$ and $\left(\hat{\mathbf{\Lambda}}^\ell\right)^{1/2}$;
- (b) keep the S first dimensions and use the reconstruction formula (15) to compute the fitted matrix:

$$\hat{\mathbf{Z}}_{I \times J}^\ell = \left(\hat{\mathbf{U}}_{I \times S}^\ell \left(\hat{\mathbf{\Lambda}}_{S \times S}^\ell \right)^{1/2} \left(\hat{\mathbf{V}}_{J \times S}^\ell \right)^\top \right) + \mathbf{M}_{I \times J}^{\ell-1}$$

and the new imputed data set becomes $\mathbf{Z}^\ell = \mathbf{W} * \mathbf{Z} + (\mathbb{1} - \mathbf{W}) * \hat{\mathbf{Z}}^\ell$ with $*$ being the Hadamard product, $\mathbb{1}_{I \times J}$ being a matrix with only ones and \mathbf{W} a weighting matrix where $w_{ij} = 0$ if z_{ij} is missing and $w_{ij} = 1$ otherwise. The observed values are the same but the missing ones are replaced by the fitted values;

- (c) from the new completed matrix \mathbf{Z}^ℓ , \mathbf{D}_Σ^ℓ and \mathbf{M}^ℓ 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 $\sum_{ij} (\hat{z}_{ij}^{\ell-1} - \hat{z}_{ij}^\ell)^2 \leq \varepsilon$, with ε equals to 10^{-6} for example.

The iterative MCA algorithm consists in recoding the incomplete data set as an incomplete disjunctive table, randomly imputing the missing values, estimating the principal components and loadings from the completed matrix and then, using these estimates to impute missing values according to the reconstruction formula (15). The steps of estimation and imputation are repeated until convergence, leading to an imputation of the disjunctive table, as well as to an estimate of the MCA parameters.

The algorithm can suffer from overfitting issues, when missing values are numerous, when the relationships between variables are weak, or when the number of observations is low. To overcome these issues, a regularized version of it has been proposed [36]. The rationale is to remove the noise in order to avoid instabilities in the prediction by replacing the singular

values $\left(\sqrt{\hat{\lambda}_s^\ell}\right)_{1 \leq s \leq S}$ of step (2.b) by *shrunk* singular values $\left(\frac{\hat{\lambda}_s^\ell - \sum_{s=S+1}^{J-K} \frac{\lambda_s}{J-K-S}}{\sqrt{\hat{\lambda}_s^\ell}}\right)_{1 \leq s \leq S}$. In this way, singular values are thresholded with a greater amount of shrinkage for the smallest ones. Thus, the first dimensions of variability take a more significant part in the reconstruction of the data than the others. Assuming that the first dimensions of variability are made of information and noise, whereas the last ones are made of noise only, this behaviour is then satisfactory. Geometrically, the regularization makes the individual closer to the center of gravity. Concerning the cells of $\hat{\mathbf{Z}}$, the regularization makes the values closer to the mean proportions and consequently, these values are more often in the interval $[0, 1]$.

The regularized iterative MCA algorithm enables us to impute an incomplete disjunctive table but not an initial incomplete data set. A strategy to go from the imputed disjunctive table to an imputed categorical data set is required. We also suggest the use of the coin flipping approach. Let us note that for each set of dummy variables coding for one categorical variable, the sum per row is equal to one, even if it contains imputed values. Moreover, most of the imputed cells are in the interval $[0, 1]$ or are close to it. Consequently, modifications of these cells are not often required.

3.3 MI using MCA

To perform MI using MCA, we need to reflect the uncertainty concerning the principal components and loadings. To do so, we use a non-parametric bootstrap approach based on the specificities of MCA. Indeed, as seen in Section 3.1, MCA enables us to assign a weight to each individual. This possibility to include a weight for the individual is very useful when the same lines of the data set occur several times. Instead of storing each replicate, a weight proportional to the number of occurrences of each line can be used, allowing the storage only of the lines that are different. Thus, a non-parametric bootstrap, such as the one used for the MI using the normal distribution, can easily be performed simply by modifying the weight of the individuals: if an individual does not belong to the bootstrap replicate, then its weight is null, otherwise, its weight is proportional to the number of times the observation occurs in the replicate. Note that individuals with a weight equal to zero are classically called *supplementary individuals* in the MCA framework [33].

Thus, we define a MI method called multiple imputation using multiple correspondence analysis (MIMCA). First, the algorithm consists in drawing M sets of weights for the individuals. Then, M single imputations are performed: at first, the regularized iterative MCA algorithm is used to impute the incomplete disjunctive table using the previous weighting for the individuals; Next, coin flipping is used to obtain categorical data and mimic the distribution of the categorical data. At the end, M imputed data sets are obtained and any statistical method can be applied on each one. In detail, the MIMCA algorithm is written as follows:

1. Reflect the variability on the set of parameters of the imputation model: draw I values with replacement in $\{1, \dots, I\}$ and define a weight r_i for each individual proportional to the number of times the individual i is drawn.
2. Impute the disjunctive table according to the previous weighting:
 - (a) initialization $\ell = 0$: recode \mathbf{X} as a disjunctive table \mathbf{Z} , substitute missing values by initial values (the proportions) and calculate \mathbf{M}^0 and \mathbf{D}_Σ^0 on this completed data set.

(b) step ℓ :

- i. perform the SVD of $\left(\mathbf{Z}^{\ell-1} - \mathbf{M}^{\ell-1}, \frac{1}{K} (\mathbf{D}_{\Sigma}^{\ell-1})^{-1}, \mathbf{diag}(r_1, \dots, r_I)\right)$ to obtain $\hat{\mathbf{U}}^{\ell}$, $\hat{\mathbf{V}}^{\ell}$ and $(\hat{\mathbf{\Lambda}}^{\ell})^{1/2}$;
- ii. keep the S first dimensions and compute the fitted matrix:

$$\hat{\mathbf{Z}}^{\ell} = \left(\hat{\mathbf{U}}^{\ell} \left(\hat{\mathbf{\Lambda}}_{shrunk}^{\ell} \right)^{1/2} \left(\hat{\mathbf{V}}^{\ell} \right)^{\top} \right) + \mathbf{M}^{\ell-1}$$

where $(\hat{\mathbf{\Lambda}}_{shrunk}^{\ell})^{1/2}$ is the diagonal matrix containing the shrunk singular values and derive the new imputed data set $\mathbf{Z}^{\ell} = \mathbf{W} * \mathbf{Z} + (\mathbf{1} - \mathbf{W}) * \hat{\mathbf{Z}}^{\ell}$

- iii. from the new completed matrix \mathbf{Z}^{ℓ} , $\mathbf{D}_{\Sigma}^{\ell}$ and \mathbf{M}^{ℓ} are updated.

(c) step (2.b) is repeated until convergence.

3. Mimic the distribution of the categorical data set using coin flipping on \mathbf{Z}^{ℓ} :

- (a) if necessary, modify suitably the values of \mathbf{Z}^{ℓ} : negative values are replaced by zero, and values higher than one are replaced by one. Then, for each set of dummy variables coding for one categorical variable, scale in order to verify the constraint that the sum is equal to one.
- (b) for imputed cells coding for one missing value, draw one category according to a multinomial distribution.

4. Create M imputed data sets: for m from 1 to M alternate steps 1, 2 and 3.

3.4 Properties of the imputation method

MI using MCA is part of the family of joint modelling MI methods, which means that it avoids the runtime issues of conditional modelling. Most of the properties of the MIMCA method are directly linked to MCA properties. MCA provides an efficient summary of the two-way associations between variables, as well as the similarities between individuals. The imputation benefits from these properties and provides an imputation model sufficiently complex to apply then an analysis model focusing on two-way associations between variables, such as a main effects logistic regression model. In addition, like the MI using the normal distribution, MIMCA uses draws from a multinomial distribution with parameter θ (obtained by the disjunctive table) specific to each individual and depending on the observed values of the other variables. Lastly, because of the relatively small number of parameters required to perform MCA, the imputation method works well even if the number of individuals is small. These properties have been highlighted in previous works on imputation using principal components methods [10, 37].

Since these two methods, MIMCA and the multiple imputation with the normal distribution, provide several imputations of the disjunctive table, and then use the same strategy to go from the disjunctive table to the categorical data set, they seem very close. However, they differ on many other points.

The first one is that the imputation of the disjunctive table by MCA is a deterministic imputation, replacing a missing value by the most plausible value given by the estimate of the principal components and the estimate of the loadings. Then, coin flipping is used to mimic the distribution of the categorical data. On the contrary, the multiple imputation

based on the normal distribution uses stochastic regressions to impute the disjunctive table, that is to say, a Gaussian noise is added to the conditional expectation given by the observed values. Then, coin flipping is used, adding uncertainty a second time.

The second difference between the two methods is the covariance of the imputed values. Indeed, the matrix $\widehat{\mathbf{Z}}^\ell$ contains the reconstructed data by the iterative MCA algorithm and the product $\widehat{\mathbf{Z}}^{\ell^\top} \widehat{\mathbf{Z}}^\ell$ provides the covariance matrix of this data. The rank of it is S . On the contrary, the rank of the covariance matrix used to perform imputation using the normal distribution is $J - K$ (because of the constraint of the sum equal to one per variable). Consequently, the relationships between imputed variables are different.

The third difference is the number of estimated parameters. Indeed, although the imputation by the normal distribution requires a extremely large number of parameters when the number of categories increases, the imputation using MCA requires a number of parameters linearly dependent to the number of cells. This property is essential from a practical point of view because it makes it very easy to impute data sets with a small number of individuals.

4 Simulation study

As mentioned in the introduction, the aim of MI methods is to obtain an inference on a quantity of interest ψ . Here, we focus on the parameters of a logistic regression without interaction, which is a statistical method frequently used for categorical data. At first, we present how to make inference for the parameters from multiple imputed data sets. Then, we explain how we assess the quality of the inference built, that is to say, the quality of the MI methods. Finally, the MI methods presented in Sections 2 and 3 are compared through a simulation study based on real data sets. It thus provides more realistic performances from a practical point of view. The code to reproduce all the simulations with the R software [38], as well as the data sets used, are available on the webpage of the first author.

4.1 Inference from imputed data sets

Each MI method gives M imputed data sets as outputs. Then, the parameters of the analysis model (for instance the logistic regression) as well as their associated variance are estimated from each one. We denote $\left(\widehat{\psi}_m\right)_{1 \leq m \leq M}$ the set of the M estimates of the model's parameters and we denote $\left(\widehat{Var}\left(\widehat{\psi}_m\right)\right)_{1 \leq m \leq M}$ the set of the M associated variances. These estimates have to be pooled to provide a unique estimate of ψ and of its variance using Rubin's rules [4].

This methodology is explained for a scalar quantity of interest ψ . The extension to a vector is straightforward, proceeding in the same way element by element. The estimate of ψ is simply given by the mean over the M estimates obtained from each imputed data set:

$$\hat{\psi} = \frac{1}{M} \sum_{m=1}^M \widehat{\psi}_m, \quad (16)$$

while the estimate of the variance of $\hat{\psi}$ is the sum of two terms:

$$\widehat{Var}(\hat{\psi}) = \frac{1}{M} \sum_{m=1}^M \widehat{Var}\left(\widehat{\psi}_m\right) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^M \left(\widehat{\psi}_m - \hat{\psi}\right)^2. \quad (17)$$

The first term is the within-imputation variance, corresponding to the sampling variance. The second one is the between-imputation variance, corresponding to the variance due to missing values. The factor $(1 + \frac{1}{M})$ is due to the fact that $\hat{\psi}$ is estimated from a finite number of imputed tables.

Then, the 95% confidence interval is calculated as:

$$\hat{\psi} \pm t_{\nu, .975} \sqrt{\widehat{Var}(\hat{\psi})}$$

where $t_{\nu, .975}$ is the .975 critical value of the Student's t -distribution with ν degrees of freedom estimated as suggested by [39].

4.2 Simulation design from real data sets

The validity of MI methods are often assessed by simulation [25, p.47]. We design a simulation study using real data sets to assess the quality of the MIMCA method. Each data set is considered as a population data and denoted \mathbf{X}_{pop} . The parameters of the logistic regression model are estimated from this population data and they are considered as the true coefficients ψ . Then, a sample \mathbf{X} is drawn from the population. This step reflects the sampling variance. The values of the response variable of the logistic model are drawn according to the probabilities defined by ψ . Then, incomplete data are generated completely at random to reflect the variance due to missing values [40]. The MI methods are applied and the inferences are performed. This procedure is repeated T times.

The performances of a MI method are measured according to three criteria: the bias given by $\frac{1}{T} \sum_{t=1}^T (\hat{\psi}_t - \psi)$, the median (over the T simulations) of the confidence intervals width as well as the coverage. This latter is calculated as the percentage of cases where the true value ψ is within its 95% confidence interval.

A coverage sufficiently close to the nominal level is required to consider that the inference is correct, but it is not sufficient, the confidence interval width should be as small as possible.

To appreciate the value of the bias and of the width of the confidence interval, it is useful to compare them to those obtained from two other methods. The first one consists in calculating the criteria for the data sets without missing values, which we named the “Full data” method. The second one is the listwise deletion. This consists in deleting the individuals with missing values. Because the estimates of the parameters of the model are obtained from a subsample, the confidence intervals obtained should be larger than those obtained from multiple imputation.

A single imputation method (named *Sample*) is added as a benchmark to understand better how MI methods benefit from using the relationships between variables to impute the data. This single imputation method consists in drawing each category according to a multinomial distribution $\mathcal{M}(\theta, 1)$, with θ defined according to the proportion of each category of the current variable.

4.3 Results

The methods described in this paper are performed using the following R packages: *cat* [41] for MI using the saturated loglinear model, *Amelia* [19, 20] for MI using a normal distribution, *mi* [42] for MI using the DPMPM method, *mice* [30, 43] for the FCS approach using iterated logistic regressions and random forests. This package will also be used to pool the results from the imputed data sets. The tuning parameters of each MI method are chosen according to their default values implemented in the R packages. Firstly, the tuning parameter of the

MIMCA method, that is to say, the number of components, is chosen to provide accurate inferences. Its choice will be discussed later in Section 4.3.3.

The MI methods are assessed in terms of the quality of the inference as well as the time consumed from data sets covering many situations. The data sets differ in terms of the number of individuals, the number of variables, the number of categories per variable, the relationships between variables.

The evaluation is based on the following categorical data sets. For each data set a categorical response variable is available.

- *Saheart*: This data set [44] provides clinical attributes of $I_{pop} = 462$ males of the Western Cape in South Africa. These attributes can explain the presence of a coronary heart disease. The data set contains $K = 10$ variables with a number of categories between 2 and 4.
- *Galetas*: This data set [45] refers to the preferences of $I_{pop} = 1192$ judges regarding 11 cakes in terms of global appreciation and in terms of color aspect. The data set contains $K = 4$ variables with two that have 11 categories.
- *Sbp*: The $I_{pop} = 500$ subjects of this data set are described by clinical covariates explaining their blood pressure [46]. The data set contains $K = 18$ variables that have 2 to 4 categories.
- *Income*: This data set, from the R package *kernlab* [47], contains $I_{pop} = 6876$ individuals described by several demographic attributes that could explain the annual income of an household. The data set contains $K = 14$ variables with a number of categories between 2 and 9.
- *Titanic*: This data set [48] provides information on $I_{pop} = 2201$ passengers on the ocean liner *Titanic*. The $K = 4$ variables deal with the economic status, the sex, the age and the survival of the passengers. The first variable has four categories, while the other ones have two categories. The data set is available in the R software.
- *Credit*: German Credit Data from the UCI Repository of Machine Learning Database [49] contains $I_{pop} = 982$ clients described by several attributes which enable the bank to classify themselves as good or bad credit risk. The data set contains $K = 20$ variables with a number of categories between 2 and 4.

The simulation design is performed for $T = 200$ simulations and 20% of missing values generated completely at random. The MI methods are performed with $M = 5$ imputed data sets which is usually enough [4].

4.3.1 Assessment of the inferences

First of all, we can note that some methods cannot be applied on all the data sets. As explained previously, MI using the loglinear model can be applied only on data sets with a small number of categories such as *Titanic* or *Galetas*. MI using the normal distribution encounters inversion issues when the number of individuals is small compared to the number of variables. That is why no results are provided for MI using the normal distribution on the data sets *Credit* and *Sbp*. The others MI methods can be applied on all the data sets.

For each data set and each method, the coverages of all the confidence intervals of the parameters of the model are calculated from T simulations (see Table 2 in Appendix A for

more details on these models). All the coverages are summarized with a boxplot (see Figure 2). The results for the bias and the confidence interval width are presented in Figure 4 and 5 in Appendix B.

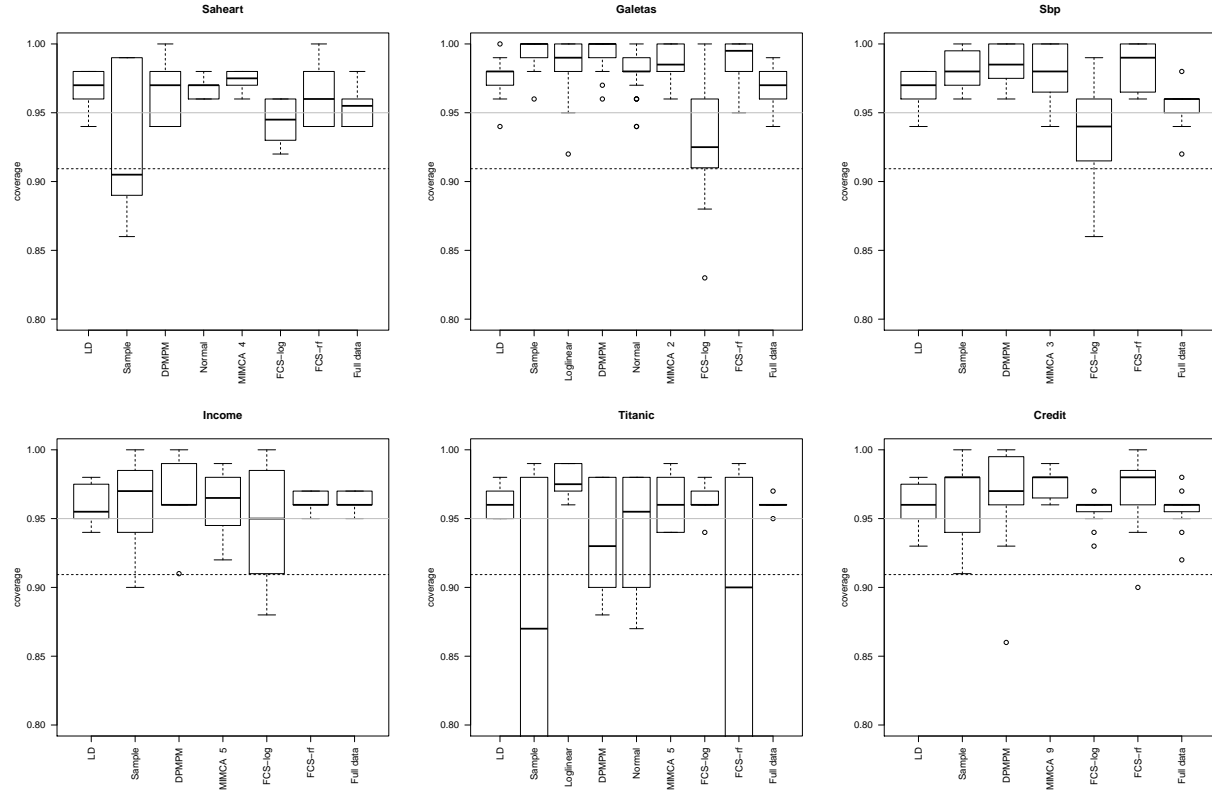


Figure 2: Distribution of the coverages of the confidence intervals for all the parameters, for several methods (Listwise deletion, Sample, Loglinear model, Normal distribution, DPMPM, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) and for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). The horizontal dashed line corresponds to the lower bound of the 95% confidence interval for a proportion of 0.95 from a sample of size 200 according to the Agresti-Coull method [50]. Coverages under this line are considered as undesirable.

As expected, MI using the loglinear model performs well on the two data sets where it can be applied. The coverages are close to the nominal levels, the biases are close to zero, and the confidence interval widths are small.

MI using the non-parametric version of the latent class model performs quite well since most of the quantities of interest have a coverage close to 95%. However, some inferences are incorrect from time to time such as on the data set *Credit* or *Titanic*. This behaviour is in agreement with the study of [18] which also presents some unsatisfactory coverages. [16] note that this MI model can have some difficulties in capturing the associations among the variables, particularly when the number of variables is high or the relationships between variables are complex, that can explain the poor coverages observed. Indeed, on the data set *Credit*, the number of variables is the highest among the data sets considered, while on the data set *Titanic*, the relationships between variables can be described as complex, in the sense that the survival status of the passengers is linked to all the other variables, but these are not closely connected. Moreover, the very poor coverages for the method *Sample* indicates that the imputation model has to take into account these relationships to provide

confidence intervals that reach the nominal rate.

MI using the normal distribution can be applied on three data sets only. On these data sets, the coverages can be too small (see *Titanic* in Figure 2). This highlights that despite the fact that this method is still often used in practice to deal with incomplete categorical data, it is not suitable and we do not recommend using such a strategy. However, [6] showed that this method could be used to impute mixed data (*i.e.* with continuous and categorical data) but only continuous variables contain missing values.

The FCS using logistic regressions encounters difficulties on the data sets with a high number of categories such as *Galetas* and *Income*. This high number of categories implies a high number of parameters for each conditional model that may explain the undercoverage on several quantities.

The FCS using random forests performs well and the method encounters difficulties only on the *Titanic* data set. This behaviour can be explained by the step of subsampling variables in the imputation algorithm (Section 2.4), *i.e.*, each tree is built with potentially different variables and with a smaller number than $(K - 1)$. In the *Titanic* data set, the number of variables is very small and the relationships between the variables are weak and all the variables are important to predict the survival response. Thus, it introduces too much bias in the individual tree prediction which may explain the poor inference. Even if, in the most practical cases, MI using random forests is very robust to the misspecification of the parameters, on this data set, the inference could be improved in increasing the number of explanatory variables retained for each tree.

Concerning MI using MCA, all the coverages observed are satisfying. The confidence interval width is of the same order of magnitude than the other MI methods. In addition, the method can be applied whatever the number of categories per variables, the number of variables or the number of individuals. Thus, it appears to be the easiest method to use to impute categorical data.

4.3.2 Computational efficiency

MI methods can be time consuming and the running time of the algorithms could be considered as an important property of a MI method from a practical point of view. Table 1 gathers the times required to impute $M = 5$ times the data sets with 20% of missing values.

	Saheart	Galetas	Sbp	Income	Titanic	Credit
Loglinear	NA	4.597	NA	NA	0.740	NA
DPMPM	20.050	17.414	56.302	143.652	10.854	24.289
Normal	0.920	0.822	NA	26.989	0.483	NA
MIMCA	5.014	8.972	7.181	58.729	2.750	8.507
FCS log	20.429	38.016	53.109	881.188	4.781	56.178
FCS forests	91.474	112.987	193.156	6329.514	265.771	461.248

Table 1: Time consumed (in seconds) to impute data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit), for different methods (Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests). The imputation is done for $M = 5$ data sets. Calculation has been performed on an Intel®Core™2 Duo CPU E7500, running Ubuntu 12.04 LTS equipped with 3 GB ram. Some values are not provided because all methods cannot be performed on each data set.

First of all, as expected, the FCS method is more time consuming than the others based on a joint model. In particular, for the data set *Income*, where the number of individuals and variables is high, the FCS using random forests requires 6,329 seconds (*i.e.* 1.75 hours), illustrating substantial running time issues. FCS using logistic regressions requires 881 seconds, a time 6 times higher than MI using the latent class model, and 15 times higher than MI method using MCA. Indeed, the number of incomplete variables increases the number of conditional models required, as well as the number of parameters in each of them because more covariates are used. In addition, the time required to estimate its parameters is non-negligible, particularly when the number of individuals is high. Then, MI using the latent class model can be performed in a reasonable time, but this is at least two times higher than the one required for MI using MCA. Thus, the MIMCA method should be particularly recommended to impute data sets of high dimensions.

Having a method which is not too expensive enables the user to produce more than the classical $M = 5$ imputed data sets. This could lead to a more accurate inference.

4.3.3 Choice of the number of dimensions

MCA requires a predefined number of dimensions S which can be chosen by cross-validation [36]. Cross-validation consists in searching the number of dimensions S minimizing an error of prediction. More precisely, missing values are added completely at random to the data set \mathbf{X} . Then, the missing values of the incomplete disjunctive table \mathbf{Z} are predicted using the regularized iterative MCA algorithm. The mean error of prediction is calculated according to $\frac{1}{\text{Card}(\mathcal{U})} \sum_{(i,j) \in \mathcal{U}} (z_{ij} - \hat{z}_{ij})^2$, where \mathcal{U} denotes the set of the added missing values. The procedure is repeated k times for a predefined number of dimensions. The number of dimensions retained is the one minimizing the mean of the k mean errors of prediction. This procedure can be used whether the data set contains missing values or not.

To evaluate how the choice of S impacts on the quality of the inferences, we perform the MIMCA algorithm varying the number of dimensions around the one provided by cross-validation. Figure 3 presents how this tuning parameter influences the coverages in the previous study. The impacts on the width of the confidence intervals are reported in Figure 6 and the ones on the bias in Figure 7 in Appendix B.

Except for the data set *Titanic*, the coverages are stable according to the number of dimensions retained. In particular, the number of dimensions suggested by cross-validation provides coverages close to the nominal level of the confidence interval. In the case of the data set *Titanic*, the cross-validation suggests retaining 5 dimensions, which is the choice giving the smallest confidence intervals, while giving coverages close to 95%. But retaining less dimensions leads to worse performances since the covariates are not closely related (Section 4.3.1). Indeed, these covariates can not be well represented within a space of low dimensions. Consequently, a high number of dimensions is required to reflect the useful associations to impute the data. *Titanic* illustrates that underfitting can be problematic. The same comment is made by [15] who advise choosing a number of classes sufficiently high in the case of MI using the latent class model. However, overfitting is less problematic because it increases the variance, but it does not skip the useful information.

5 Conclusion

This paper proposes an original MI method to deal with categorical data based on MCA. The principal components and the loadings that are the parameters of the MCA enables the

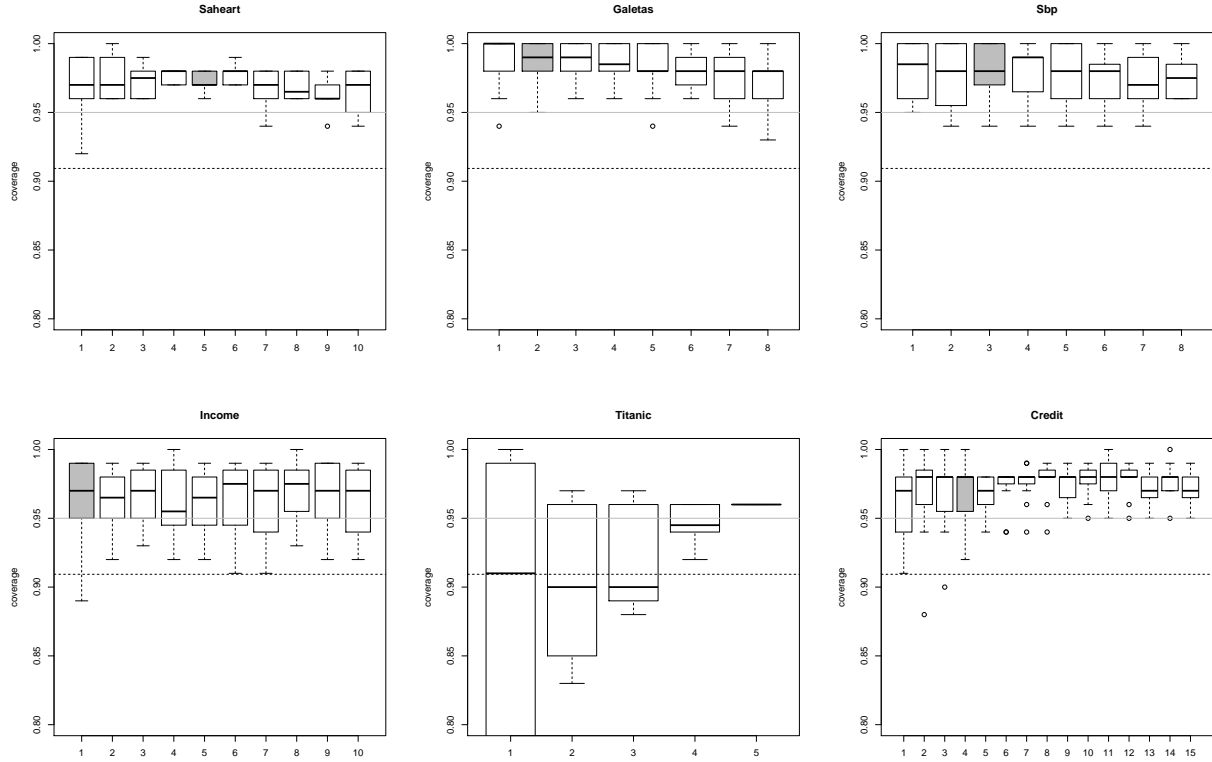


Figure 3: Distribution of the coverages of the confidence intervals for all the parameters for the MIMCA algorithm for several numbers of dimensions and for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). The results for the number of dimensions provided by cross-validation are in grey. The horizontal dashed line corresponds to the lower bound of the 95% confidence interval for a proportion of 0.95 from a sample of size 200 according to the Agresti-Coull method [50]. Coverages under this line are considered as undesirable.

imputation of data. To perform MI, the uncertainty on these parameters is reflected using a non-parametric bootstrap, which results in a specific weighting for the individuals.

From a simulation study based on real data sets, this MI method has been compared to the other main available MI methods for categorical variables. We highlighted the competitiveness of the MIMCA method to provide valid inferences for an analysis model requiring two-way associations (such as logistic regression without interaction, or a homogeneous log-linear model, proportion, odds ratios, etc).

We showed that MIMCA can be applied to various configurations of data. In particular, the method is accurate for a large number of variables, for a large number of categories per variables and when the number of individuals is small. Moreover, the MIMCA algorithm performs fairly quickly, allowing the user to generate more imputed data sets and therefore to obtain more accurate inferences (M between 20 and 100 can be beneficial [25, p.49]). Thus, MIMCA is very suitable to impute data sets of high dimensions that require more computation. Note that MIMCA depends on a tuning parameter (the number of components), but we highlighted that the performances of the MI method are robust to a misspecification of it.

Because of the intrinsic properties of MCA, MI using MCA is appropriate when the

analysis model contains two-way associations between variables such as logistic regression without interaction. To consider the case with interactions, one solution could be to introduce to the data set additional variables corresponding to the interactions. However, the new variable "interaction" is considered as a variable in itself without taking into account its explicit link with the associated variables. It may lead to imputed values which are not in agreement with each others. This topic is a subject of intensive research for continuous variables [51, 52].

In addition, the encouraging results of the MIMCA to impute categorical data prompt the extension of the method to impute mixed data. The first research in this direction [37] has shown that the principal components method dedicated to mixed data (called Factorial Analysis for Mixed Data) is efficient to perform single imputation, but the extension to a MI method requires further research.

References

- [1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.
- [2] X. L. Meng and D. B. Rubin. Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm. *Journal of the American Statistical Association*, 86(416):899–909, December 1991.
- [3] P. D. Allison. Handling missing data by maximum likelihood. In *SAS global forum*, pages 1–21, 2012.
- [4] D. B. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, 1987.
- [5] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley series in probability and statistics, New-York, 1987, 2002.
- [6] J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997.
- [7] D.W. van der Palm, L.A. van der Ark, and J.K. Vermunt. A comparison of incomplete-data methods for categorical data. *Statistical methods in medical research*, 2014. in press.
- [8] S. Van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76:1049–1064, 2006.
- [9] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [10] V. Audigier, F. Husson, and J. Josse. Multiple imputation for continuous variables using a Bayesian principal component analysis. *ArXiv e-prints*, January 2014.
- [11] J. Josse and F. Husson. Multiple imputation in PCA. *Advances in data analysis and classification*, 5:231–246, 2011.
- [12] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.
- [13] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. A Wiley-Interscience publication. Wiley, New York, 1992.
- [14] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:805–811, 1987.
- [15] J. K. Vermunt, J. R. van Ginkel, L. A. van der Ark, and K. Sijtsma. Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociological Methodology*, Vol 38, 38:369–397, 2008.
- [16] D. Vidotto, M. C. Kapteijn, and Vermunt J.K. Multiple imputation of missing categorical data using latent class models: State of art. *Psychological Test and Assessment Modeling*, 2014. in press.
- [17] D. B. Dunson and C. Xing. Nonparametric Bayes modeling of multivariate categorical data. 104(487):1042–1051, September 2009.

- [18] Y. Si and J.P. Reiter. Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38:499–521, 2013.
- [19] J. Honaker, G. King, and M. Blackwell. *Amelia II: A Program for Missing Data*, 2014. R package version 1.7.2.
- [20] J. Honaker, G. King, and M. Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011.
- [21] P. D. Allison. *Missing Data*. Thousand Oaks, CA: Sage, 2002.
- [22] C. A. Bernaards, T. R. Belin, and J. L. Schafer. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26(6):1368–1382, mar 2007.
- [23] H. Demirtas. Rounding strategies for multiply imputed binary data. *Biometrical journal*, 51(4):677–88, 2009.
- [24] R. M. Yucel, Y. He, and A. M. Zaslavsky. Using calibration to improve rounding in imputation. *The American Statistician*, 62:125–129, 2008.
- [25] S. Van Buuren. *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC, 1 edition, 2012.
- [26] J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 1974.
- [27] L. L. Doove, S. Van Buuren, and E. Dusseldorp. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104, 2014.
- [28] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, 179(6):764–774, March 2014.
- [29] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.
- [30] S. Van Buuren and C. G. M. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.
- [31] M. J. Greenacre and J. Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC, 2006.
- [32] L. Lebart, A. Morineau, and K. M. Werwick. *Multivariate Descriptive Statistical Analysis*. Wiley, New-York, 1984.
- [33] M. J. Greenacre. *Theory and applications of correspondence analysis*. Academic Press, London, 1984.
- [34] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936.

- [35] M. Tenenhaus and F. W. Young. 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, 1985.
- [36] J. Josse, M. Chavent, B. Liqueur, and F. Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29:91–116, 2012.
- [37] V. Audigier, F. Husson, and J. Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, pages 1–22, 2014. in press.
- [38] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.
- [39] J. Barnard and D. B. Rubin. Small Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, 86:948–955, 1999.
- [40] J. P. L. Brand, S. van Buuren, K. Groothuis-Oudshoorn, and E. S. Gelsema. A toolkit in sas for the evaluation of multiple imputation methods. *Statistica Neerlandica*, 57(1):36–45, 2003.
- [41] T. Harding, F. Tusell, and J. L. Schafer. *cat: Analysis of categorical-variable datasets with missing values*, 2012. R package version 0.0-6.5.
- [42] A. Gelman, J. Hill, Y. Su, M. Yajima, M. Grazia Pittau, B. Goodrich, and Y. Si. *mi: Missing Data Imputation and Model Checking*, 2013. R package version 0.9-93.
- [43] S. Van Buuren and K. Groothuis-Oudshoorn. *mice*, 2014. R package version 2.22.
- [44] J. Rousseauw, J. du Plessis, A. Benade, P. Jordann, J. Kotze, P. Jooste, and J. Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64:430–436, 1983.
- [45] Applied Mathematics Department, Agrocampus Rennes, France. galetas data set. Available on <http://math.agrocampus-ouest.fr/infogluceDeliverLive/digitalAssets/74258-galetas.txt>.
- [46] GlaxoSmithKline, Toronto, Ontario, Canada. Blood pressure data set. Available on <http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMainF.htm>.
- [47] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.
- [48] R. Dawson and MacG. J. The ‘unusual episode’ data revisited. *Journal of Statistics Education*, 3, 1995.
- [49] M. Lichman. UCI machine learning repository, 2013.
- [50] A. Agresti and B. A. Coull. Approximate Is Better than ”Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126, May 1998.
- [51] S. R. Seaman, J. W. Bartlett, and I. R. White. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC medical research methodology*, 12(1):46, 2012.

- [52] J. W. Bartlett, S. R. Seaman, I. R. White, and J. R. Carpenter. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical Methods in Medical Research*, 2014.

Appendix

A Simulation design: analysis models and sample characteristics

Data set	number of individuals	number of variables	sample size	logistic regression model	number of quantities of interest
Saheart	462	10	300	CHD = FAMHIST + TOBACCO + ALCOHOL	30
Galetas	1192	4	300	GALLE = GRUPO	6
Sbp	500	18	200	SBP = SMOKE + EXERCISE + ALCOHOL	12
Income	6876	14	1500	INCOME = SEX	8
Titanic	2201	4	300	SURV = CLASS+AGE+SEX	6
Credit	982	20	300	CLASS = CHECK- ING_STATUS + DURATION + CREDIT_HISTORY + PURPOSE	11

Table 2: Set of the sample characteristics and of the analysis models used to perform the simulation study (Section 4.2) for the several data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit).

B Simluation study: complementary results

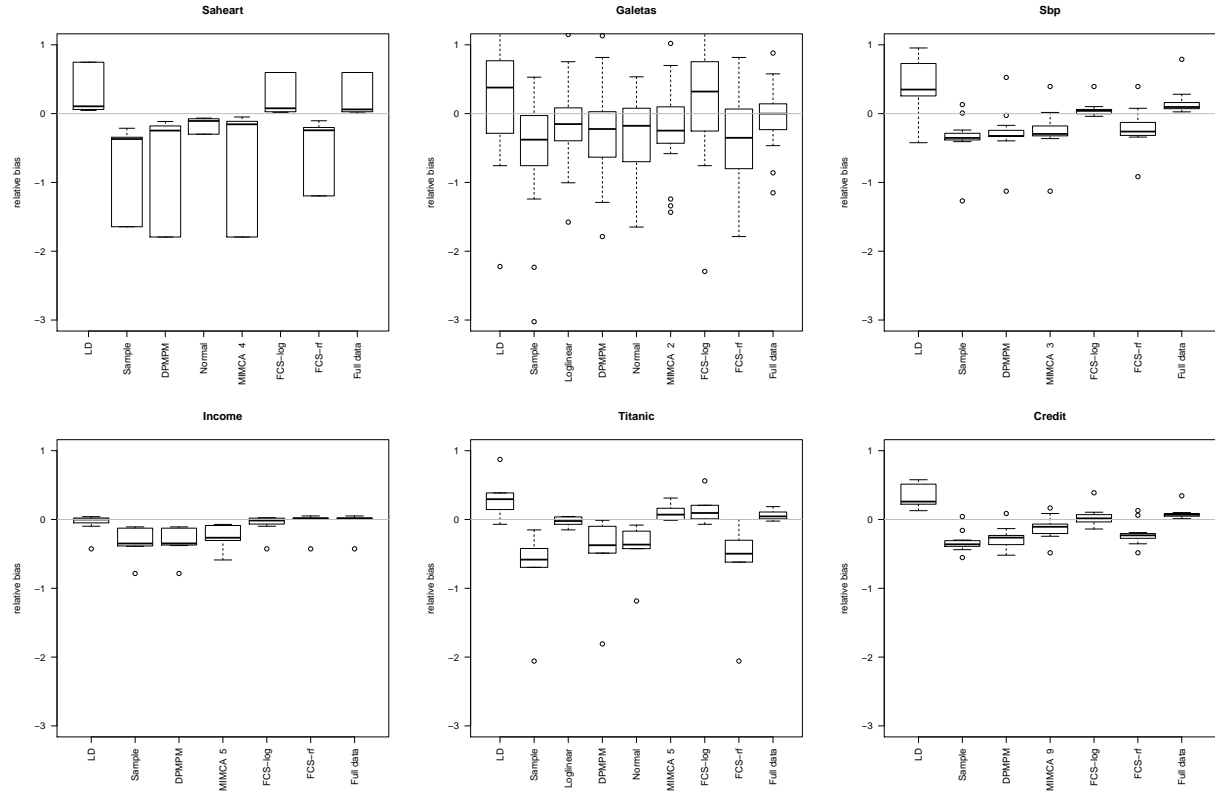


Figure 4: Distribution of the relative bias (bias divided by the true value) over the several quantities of interest for several methods (Listwise deletion, Sample, Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the relative bias observed for one coefficient.

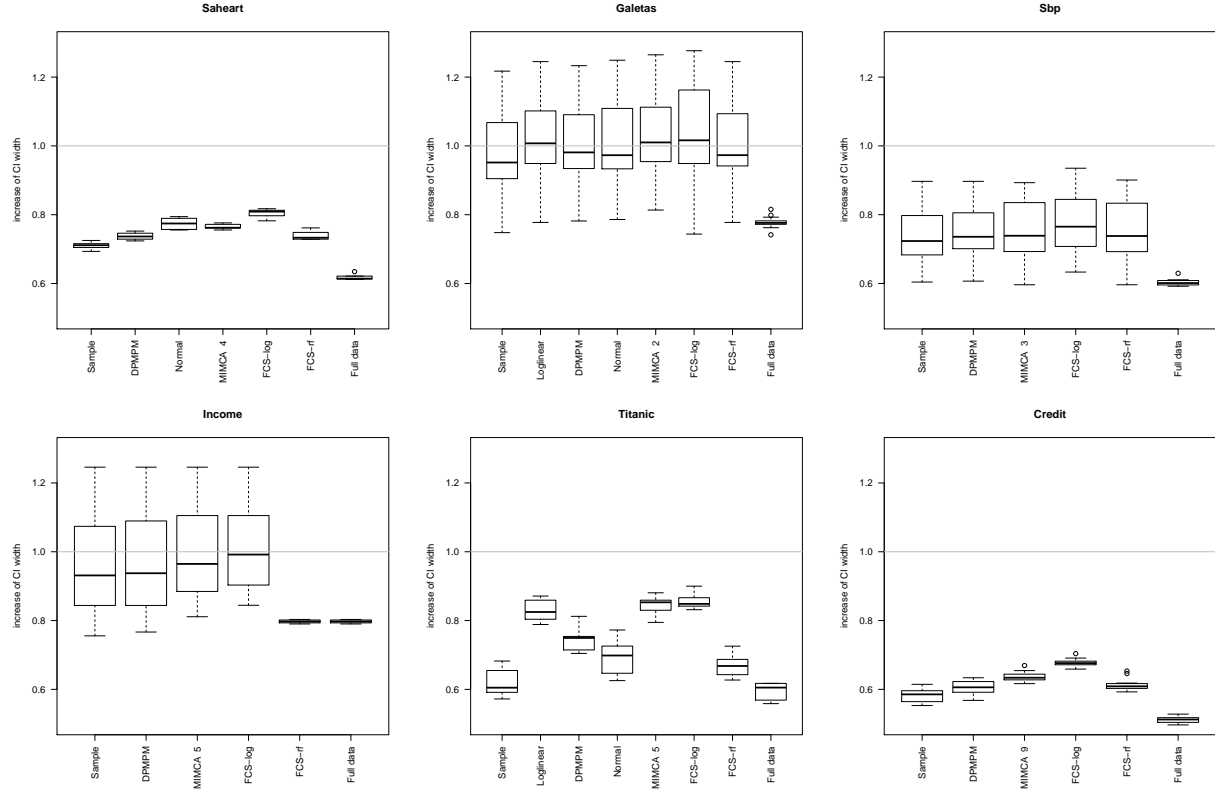


Figure 5: Distribution of the median of the confidence interval for the several quantities of interest for several methods (Sample, Loglinear model, DPMPM, Normal distribution, MIMCA, FCS using logistic regressions, FCS using random forests, Full data) for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the median of the confidence interval observed for one coefficient divided by the one obtained by Listwise deletion. The horizontal dashed line corresponds to a ratio of 1. Points over this line corresponds to confidence interval higher than the one obtain by listwise deletion.

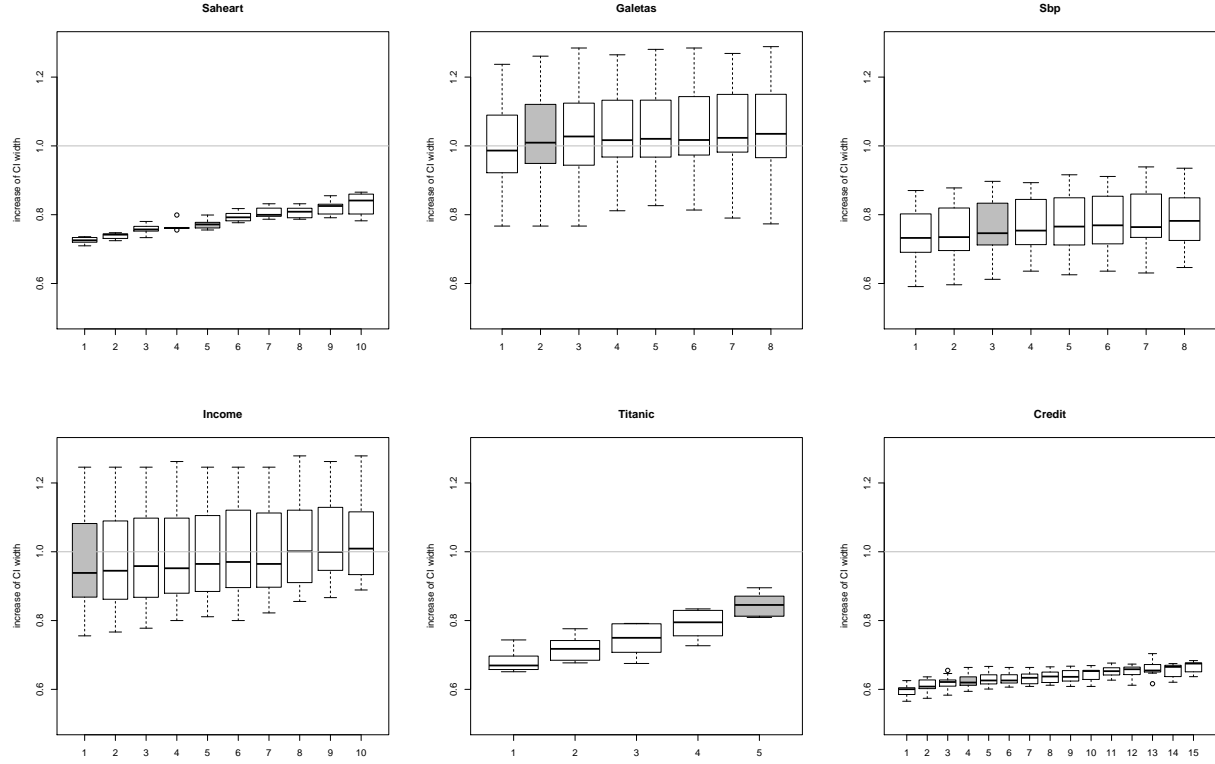


Figure 6: Distribution of the median of the confidence interval for the several quantities of interest for the MIMCA algorithm for several numbers of dimensions for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the median of the confidence interval observed for one coefficient divided by the one obtained by Listwise deletion. The horizontal dashed line corresponds to a ratio of 1. Points over this line corresponds to confidence interval higher than the one obtain by listwise deletion. The results for the number of dimensions provided by cross-validation are in grey.

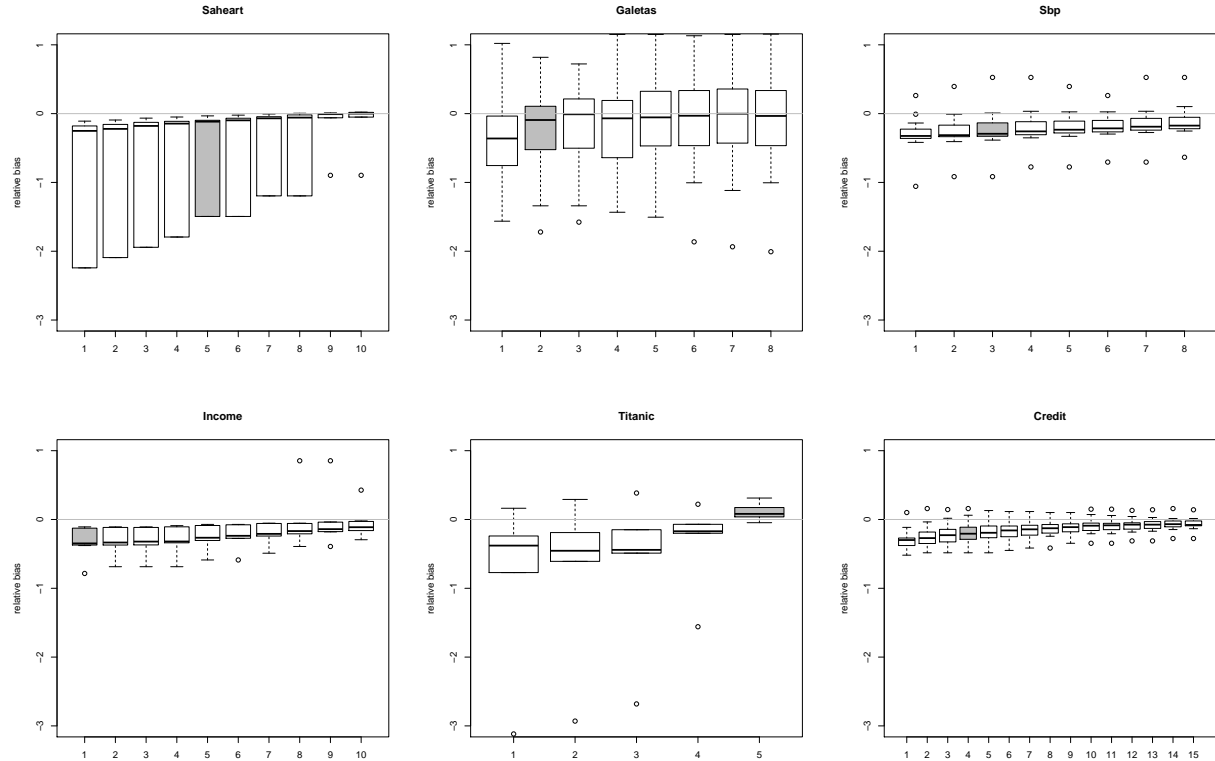


Figure 7: Distribution of the relative bias (bias divided by the true value) over the several quantities of interest for the MIMCA algorithm for several numbers of dimensions for different data sets (Saheart, Galetas, Sbp, Income, Titanic, Credit). One point represents the relative bias observed for one coefficient. The results for the number of dimensions provided by cross-validation are in grey.