Multiple Imputation using Regularised Iterative Multiple Correspondence Analysis

Johané Nienkemper-Swanepoel*
Stellenbosch University, Stellenbosch, South Africa – jnienkie@gmail.com

Michael J von Maltitz
University of the Free State, Bloemfontein, South Africa – vmaltitzmj@ufs.ac.za

Abstract

The occurrence of non-response in survey data is a prevalent problem, often leading to invalid inferences and inefficient estimates. The application of a regularised iterative multiple correspondence analysis (RIMCA) algorithm in single imputation (SI) has been suggested for the handling of missing categorical data in survey analysis (Josse, Chavent, Liquet & Husson 2012). An adapted version of this algorithm is applied as a multiple imputation (MI) technique in this paper and compared to the published results. A comparison is drawn between the performance of SI and MI making use of RIMCA for both simulated and survey data. It was found that the MI procedure allowed for better estimates and wider confidence intervals (as expected from a valid imputation procedure).

Key terms: incomplete ordinal categorical data, multiple imputation, multiple correspondence analysis, principal component analysis, regularised iterative multiple correspondence analysis.

1 Introduction

Missing values are a common occurrence in the analysis of survey data. Missing data entries may result in a biased sample when the mechanism that causes data to become missing acts as a second round of sampling resulting in a final sample not representative of the population in question.

Missing data occurs for various reasons, ranging from the capturing of data to the handling of data. Researchers believe that data entries become missing because of a random process, referred to as the distribution of missingness (Little & Rubin 2002). Three missingness mechanisms can occur: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). The MAR mechanism classifies missing values that are dependent on the observed values in the data set and independent of the other missing values that occur. MCAR is an extension of the MAR mechanism, since in this case the missing values are independent of all variables in the data set, observed and missing. Values that are missing because of the MNAR mechanism will at least be dependent on missing data. Both the MAR and MCAR mechanisms are classified as ignorable non-responses whereas MNAR is referred to as non-ignorable (Buhi, Goodson & Neilands 2008; Schafer & Graham 2002). Ignorable non-responses enable the researcher to ignore the cause of missingness and therefore simplify the procedures for the analysis of missing data (Buhi *et al*. 2008).

The method chosen to handle the missing values in a data set will determine the validity of results and analysis; therefore, it is of utmost importance to always have the sample data reflect the population that the sample is drawn from in order to obtain accurate inferences. Generally three method classes can be defined for the handling of missing values: deletion, reweighting and imputation, the latter being preferred for several reasons (see Buhi *et al.* 2008; Schafer & Graham 2002). Imputation techniques consist of single imputation (SI) and multiple imputation (MI). SI replaces each missing value with one plausible value in order to fill the data set to its original size, whereas MI replaces multiple plausible values for each missing data entry resulting in several complete data sets to analyse. The success of MI lies in the incorporation of the uncertainty that arises from imputing missing values into the overall inferences, therefore achieving realistic variances whilst maintaining relationships that may occur between variables. This paper attempts to develop another branch of MI, investigating the applicability of a regularised iterative multiple correspondence analysis (RIMCA) algorithm to multiply impute missing values in categorical data sets.

The SI RIMCA procedure developed by Josse *et al.* (2012) experiences two problems, namely the uncertainty of the choice of retained dimensions in the dimension-reduction algorithm, and the problem

that the imputed values have inherent uncertainties which are not modelled in the SI method. Both of these problems are solved in the adaptation of RIMCA in SI to MI.

Three measures of uncertainty should be incorporated for a valid MI procedure (Rubin 2003). These are, *firstly*, uncertainty arises in choosing the distribution of the missingness mechanism, *secondly*, uncertainty in the imputation model and the parameter values used to create the imputations, and *thirdly*, residual uncertainty occurs when drawing imputed values. The incorporation of the uncertainty measures in the RIMCA algorithm will be discussed in section 2.4.

2 Methodology

2.1 MCA as weighted PCA of a triplet

The RIMCA algorithm is based on multiple correspondence analysis (MCA). However, it is necessary to perform MCA as a weighted principal component analysis (PCA), a continuous multivariate data technique, since, during the algorithm, non-categorical data will be created where there was missing data originally.

In order to illustrate MCA as a weighted PCA, a data set with I individuals and J categorical variables $v_j, j = 1, \dots, J$ with k_j categories, is considered.

MCA is presented as the PCA of a triplet, $(\mathbf{Z}, \mathbf{M}, \mathbf{D})$, as follows (Josse *et al.* 2012):

$$\left(I\mathbf{X}\mathbf{D}_{\Sigma}^{-1}, \frac{1}{IJ}\mathbf{D}_{\Sigma}, \frac{1}{I}\mathbb{I}_I\right).$$

The first term of the triplet, $\mathbf{Z}$, represents the data, the second term, $\mathbf{M}$, represents the metric and the third term, $\mathbf{D}$, represents the row masses (Josse, Chavent, Liquet & Husson 2011).

The diagonal matrix of the column margins of the indicator matrix, $\mathbf{X}$, is given by $\mathbf{D}_{\Sigma} = diag\left((I_k)_{k=1,\dots,K}\right)$. The matrix $\mathbf{M} = \frac{1}{IJ}\mathbf{D}_{\Sigma}$ is used to compute the distances between the rows. The diagonal matrix $\mathbf{D} = \frac{1}{I}\mathbb{I}_I$ corresponds to the row masses, where $\mathbb{I}_d$ is the identity matrix of size d. For information regarding the expansion of the data matrix, $\mathbf{Z}$, and performing PCA on a triplet, see Josse *et al.* (2012). Generalised singular value decomposition (GSVD) is used in order to obtain specific estimates from the data matrix. For information on this procedure.

2.2 RIMCA

The RIMCA algorithm consists of three steps: initialisation, reconstruction and iteration. These steps are given in Josse *et al.* (2012), but are reiterated here for clarification. Consider a data set with I individuals, J categorical variables, each variable $j = 1, \dots, J$ with k_j categories, $K = \sum_{j=1}^{J} k_j$.

i) Initialisation ($\ell = 0$)

The data matrix is transformed to an indicator matrix, $\mathbf{X}^0$, of dummy variables consisting of zeros and ones. The missing values are substituted by proportioned initial values, $\frac{I_k}{I}$, which is a mean imputation for continuous variables (referred to as the missing fuzzy average method) in which missing values are substituted by the proportion observed in each category (Van der Heijden & Escofier 2003). A constraint is imposed over the row margins per variable to add up to one, in order to satisfy the barycentric relations required for correspondence analysis, consequently multiple correspondence analysis. The column margins of the now completed indicator matrix is obtained and is expressed as the diagonal matrix entries of $\mathbf{D}_{\Sigma}^{0} = diag(\left(I_k^0\right)_{k=1,\dots,K})$, where I_k^0 is the column margin of column k.

ii) Reconstruction

The second step reconstructs the data and imputes plausible values to the missing values. These plausible values are decimal values between zero and one, adding up to one for a single variable. These imputed decimal values will henceforth be known as the imputed fuzzy values, since they do not indicate the imputed category, but rather the degree of membership of the observation to each possible category. The imputed fuzzy values are based on the MCA axes and components, providing plausible values with respect to the observed data.

Firstly, MCA is performed on the now completed indicator matrix, $\mathbf{X}^{\ell-1}$, which is weighted PCA on the triplet: $\left(I\mathbf{X}^{\boldsymbol{\ell-1}}\left(\mathbf{D}_{\Sigma}^{\ell-1}\right)^{-1}, \frac{1}{IJ}\mathbf{D}_{\Sigma}^{\ell-1}, \frac{1}{I}\mathbb{I}_I\right)$.

The estimates $\hat{\mathbf{F}}^{\ell}$ and $\hat{\mathbf{U}}^{\boldsymbol{\ell}}$, which are the matrices of eigenvectors used for the decomposition of the data matrix $\mathbf{Z}^{\ell}$, are obtained from the GSVD of the following: $\left(I\mathbf{X}^{\ell-1}\left(\mathbf{D}_{\Sigma}^{\ell-1}\right)^{-1} - \mathbf{1}_I\mathbf{1}_K'\right) \times \sqrt{\frac{\mathbf{D}_{\Sigma}^{\ell-1}}{IJ}}$

Secondly, $\mathbf{Z}^{\ell}$ is reconstructed using a pre-determined number of dimensions, S, are retained in the following reconstruction: $\hat{z}_{ik}{}^{\ell} = 1 + \sum_{s=1}^{S} \frac{\hat{f}_{is}^{\ell}}{\|\hat{\mathbf{f}}_s^{\ell}\|}\left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)\hat{u}_{ks}^{\ell}$,

where $\| \hat{\mathbf{f}}_s^{\ell} \|$ is calculated according to the Hilbert-Schmidt norm and λ_s represents the eigenvalue of rank s, which is also the variance of each component $\mathbf{f}_s$.The variance is estimated by the mean of the last eigenvalues: $\hat{\sigma}^2 = \frac{1}{K-J-S}\sum_{s=S+1}^{K-J}\lambda_s$

Now the indicator matrix is updated by allocating the associated values and then replacing the fuzzy initial values with imputed values obtained from the reconstruction. The associated values are obtained by using the margins of step $\ell - 1$ in the following way:

$\hat{x}_{ik}^{\ell} = \frac{1}{I}\left(1 + \sum_{s=1}^{S}\frac{\hat{f}_{is}^{\ell}}{\|\hat{\mathbf{f}}_s^{\ell}\|}\left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)\hat{u}_{ks}^{\ell}\right)\mathbf{D}_{\Sigma}^{\ell-1}$; expressed in matrix notation: $\hat{\mathbf{X}}^{\ell} = \frac{1}{I}\hat{\mathbf{Z}}^{\ell}\mathbf{D}_{\boldsymbol{\Sigma}}^{\ell-1}$

The indicator matrix is then updated using: $\mathbf{X}^{\ell} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^{\ell}$, where $\mathbf{W}$ is a weight matrix indicating missing values with a zero and observed values with a one. According to Josse *et al.* (2012) the weight matrix enables the minimisation of the reconstruction error over all non-missing values in a data set, while ignoring the missing values.

Thirdly, the column margins, I_k^{ℓ}, of the imputed indicator matrix, $\mathbf{X}^{\ell}$, are calculated. This results in the updated diagonal matrix of column margins, $\mathbf{D}_{\Sigma}^{\ell}$.

iii) Iteration

The iteration step is concerned with the repetition of the reconstruction step until the difference between the imputed indicator matrix from one repetition to the next falls below a pre-determined threshold, Ɛ, fixed at 10^{-6}. The change is measured by $\sum_{ik}\left(\hat{x}_{ik}^{\ell-1} - \hat{x}_{ik}^{\ell}\right)^2 \leq \varepsilon$ (Josse *et al*. 2012).

In SI, the final categorical dataset is obtained by replacing the fuzzy imputed values with category values (‘0’s or ‘1’s). This is done per variable for each observation; the category with the largest fuzzy value will be allocated a ‘1’ in the indicator matrix, thereby allocating the most plausible category values with respect to a degree of membership (Josse *et al*. 2012).

2.3 From SI to MI

The methodology followed by Josse *et al.* (2012) is adopted and altered with respect to the three uncertainty measures required by MI.

The RIMCA algorithm is proposed for MAR and MCAR values, therefore the missing values are considered as ignorable. Thus the ignorable non-responses allow the researcher to ignore the distribution of missingness (Buhi *et al.* 2008), covering the first uncertainty required by MI.

The second required uncertainty, uncertainty in the model, is initially incorporated by allowing for random starting points. Randomly generated Uniform(0,1) initial values are allocated for the category value of a particular variable, still placing a constraint over the category values per variable to add up to one, in order to satisfy the barycentric relations required for MCA.

The number of dimensions to retain in the reconstruction algorithm will not be fixed *a priori*, in contrast to the procedure by Josse *et al.* (2012). This represents additional model uncertainty. All possible reconstruction dimension limit choices can be used in order to generate imputed data sets. Thus a range of final data sets are obtained with varying degrees of under- and overfitting. This solves the

first problem of RIMCA in SI, namely dimension choice. The number of multiple data sets to use in MI (number of reconstruction dimension limit choices) is recommended to be a modest number between two and ten (Rubin 1987). In this paper ten multiple data sets are randomly chosen from the possible S dimension choices that are available.[1]

A final adaptation incorporating both model uncertainty and uncertainty when drawing imputed values, the third required uncertainty, is then applied. Five final data sets will be drawn for each of the fuzzy indicator matrices built from each of the ten converged reconstructions. The category randomly assigned for a missing datum is drawn with probability equal to the degree of membership of that observation to each category in the missing datum. Thus each original incomplete data set will result in 50 imputed data sets capturing both model and imputed value uncertainty. Drawing multiply from the final fuzzy values solves the second problem of the SI procedure mentioned in Section 1, namely that the imputed values have inherent uncertainties that were not accounted for in SI.

One drawback of the category assignment should be noted. Since the final fuzzy values are fixed after the algorithm converges, the parameters of the imputation model are also fixed. This means that imputations are drawn conditionally on estimates of the parameters, so the MI variance might be underestimated.

Overall, however, all three uncertainties required for valid MI are accounted for in this paper's adapted RIMCA algorithm.

2.4 Analysing the data

In order to establish which imputation procedure produces the best results, the RIMCA algorithm is applied as both a SI and MI technique to the same simulated and real data sets. Means and confidence intervals obtained from both RIMCA procedures are compared with those from the incomplete data (and the original data in the simulation study). Rubin's rules (Rubin 1987) will be used for the calculation of the descriptive statistics obtained by MI, while the confidence intervals for the means of the singly imputed data sets will be constructed using Student's t-distribution.

3 Simulation study

3.1 Data

The simulation protocol followed by Josse *et al.* (2012) is replicated in this paper. Complete data sets are generated from a multivariate Normal distribution consisting of 100 observations in ten variables with different correlation structures (0.4 and 0.8). The complete data sets are made incomplete by inserting different percentages (MCAR: 10% and 30%; MAR: 8% and 16%) of missing values with random and non-random patterns using both MCAR and MAR mechanisms. Overall, 16 different data scenarios are simulated.

3.2 Results

The bias and mean square error (MSE) of the variable means are obtained from the complete-case analysis (CC), SI and MI procedures over 1000 simulations for each of the 16 different data scenarios. Each variable's bias and MSE in each scenario are ranked, and the summary of all the method rankings obtained by each variable in each scenario are provided in Table 3.1. Rank 1 indicates the smallest bias and MSE among the three procedures, consequently rank 3 indicates the largest bias and MSE.

Table 3.1 Summary of 1000 simulations

BIAS				MSE			
RANK	SI	CC	MI	RANK	SI	CC	MI
1	0	43	80	1	0	28	95
2	0	80	43	2	0	95	28
3	123	0	0	3	123	0	0

3.3 Discussion

RIMCA in MI resulted in smaller bias and MSE than RIMCA in SI over 1000 simulations and 16 data scenarios (Table 3.1). This confirms the superiority of the MI procedure over the SI procedure. Since some of the missingness mechanisms are MCAR, we expect to see that CC rankings are in some cases better than SI and MI.

[1] Sensitivity analyses were performed in order to confirm whether the results obtained from RIMCA were sensitive towards a random selection of dimensions. The analyses confirmed that the results remained relatively stable over the various selections.

Looking at the detailed analysis results on a single simulation under each scenario (not tabulated in this paper) there are several notable features. The confidence intervals obtained from MI are wider than the SI confidence intervals in almost all cases, with the exception of two cases in which a slightly wider confidence interval was obtained from the SI procedure. The wider confidence intervals provided by MI confirm the successful incorporation of additional uncertainty. It is found that MI performs slightly better in MAR incomplete data with a low correlation structure, contrary to literature stating that RIMCA in SI is expected to perform better in data with a high correlation structure (Josse *et al.* 2012). In select cases it was found that the SI confidence intervals did not contain the true mean, resulting in inaccurate estimation.

For all simulated data sets it is found that the CC estimates and MI estimates are extremely close to each other. The MI confidence intervals are narrower than the confidence intervals provided by CC analysis. The narrow confidence intervals can be partially explained by perhaps the small amount of between-variance that was evident, as well as the increased sample size. In any case, a slightly underestimated variance is expected, because of the fixed parameter values in the final step of the RIMCA MI procedure.

The main advantage of the MI procedure remains that, in combining the multiple data sets, valid inferences are attained incorporating additional variance caused by the missing values.

4 Application

4.1 Data

We now consider the user satisfaction survey of craft operators on the *Canal des Deux Mers*, in the South of France, undertaken by *Voes Navigables de France*, the same data set analysed by Josse *et al.* (2012). The questionnaire consists of the responses of 1232 individuals to 14 questions with two or three possible categories, with a total of 35 categories. In this data set 9.07% of the data is missing, and non-response occurs in 42.5% of the respondents.

4.2 Results

The comparison between the results obtained from RIMCA in MI and SI is given by means of Table 4.1. The confidence intervals, estimated means and standard errors of the incomplete real data are given as the complete-case (CC) analysis.

Table 4.1 Confidence interval widths, means and standard errors obtained from complete-case analysis, RIMCA in SI and RIMCA in MI

Real data	Confidence Interval Width			Mean			Standard Error		
Variable	CC	SI	MI	CC	SI	MI	CC	SI	MI
1*	0.0738	0.0714	**0.0734**	1.4303	1.4131	1.4301	0.0188	0.0182	0.0187
2*	0.0988	0.0961	**0.0980**	1.9852	1.9156	1.9839	0.0252	0.0245	0.0250
3*	0.0641	0.0623	**0.0637**	1.2928	1.2833	1.2926	0.0163	0.0159	0.0163
4*	0.0500	0.0492	**0.0502**	1.2676	1.2622	1.2675	0.0127	0.0125	0.0128
5*	0.0495	0.0485	**0.0494**	1.2575	1.2516	1.2577	0.0126	0.0124	0.0126
6*	0.0416	0.0407	**0.0414**	1.1610	1.1575	1.1611	0.0106	0.0104	0.0106
7*	0.0567	0.0557	**0.0565**	1.4741	1.4602	1.4746	0.0144	0.0142	0.0144
8*	0.0567	0.0544	**0.0562**	1.4095	1.3856	1.4082	0.0144	0.0139	0.0143
9*	0.0764	0.0674	**0.0754**	1.9383	1.9456	1.9390	0.0195	0.0172	0.0192
10*	0.0626	0.0517	**0.0606**	1.4017	1.3084	1.4033	0.0159	0.0132	0.0154
11*	0.0640	0.0530	**0.0622**	1.4491	1.3401	1.4506	0.0163	0.0135	0.0158
12*	0.0785	0.0534	**0.0731**	2.0155	2.0106	2.0157	0.0200	0.0136	0.0186
13*	0.0827	0.0808	**0.0827**	2.1938	2.1891	2.1937	0.0211	0.0206	0.0211
14	0.0635	**0.0714**	0.0632	1.7150	1.4131	1.7153	0.0162	0.0156	0.0161

CC – complete-case analysis, SI – single imputation, MI – multiple imputation, * – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI), wider confidence interval given in **bold**

4.3 Discussion

The estimated means for the SI and MI procedures are similar with a few slight deviations. The confidence intervals for MI are slightly wider for all of the variables, with the exception of variable 14, where the SI confidence interval is wider.

RIMCA in MI adds uncertainty to the imputation procedure, resulting in wider confidence intervals for the mean in this application. The performance of SI and MI is very similar in the case of this specific data set. The similarity of the two procedures (SI and MI) might be due to the small number of missing values in the data (9%) and also because of a high correlation between the variables (meaning that even less information is missing).

The post-MI means are closer to the CC means than the SI means are, but the importance of this result is debatable. The MI procedure is better replicating the relationships present in the incomplete data, but we cannot tell whether these relationships are true for the population.

5 Conclusion

The RIMCA algorithm was performed, in both SI and MI methods on the same data sets, simulated and real. The simulation study consisted of 16 different data sets, varying with regard to correlation structure, missingness mechanisms and percentage of missing values in the data.

From the simulation and real data application, it was found that in almost all of the cases the confidence intervals provided by MI were wider than those from SI, which confirms the added uncertainty when multiple data sets are imputed. Within the simulation study, the mean estimates obtained from MI were also closer to the true mean values than the estimates provided by SI were.

In summary, this paper shows that in the case of incomplete ordinal data, the RIMCA MI algorithm should be chosen over the RIMCA SI algorithm and CC analysis in order to obtain completed data sets for further analysis.

References

Buhi ER, Goodson P & Neilands TB. 2008. Out of Sight, Not Out of Mind: Strategies for Handling Missing Data. *American Journal of Health Behaviour* 32(1):83–92.

Josse J, Chavent M, Liquet B & Husson F. 2011. Handling missing values with regularized iterative multiple correspondence analysis. *ICC conference notes.* University of St Andrew. United Kingdom.

Josse J, Chavent M, Liquet B & Husson F. 2012. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification* 29:91–116.

Little RJA & Rubin DB. 2002. *Statistical analysis with missing data*, 2nd ed. Wiley-Interscience, John Wiley & Sons, Inc. Publication. United States of America.

Rubin DB. 1978. Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section*. Washington D.C., pp. 20–34.

Rubin DB. 1987. Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc. United States of America.

Rubin DB. 1996. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association* 91(434):473–489.

Rubin DB. 2003. Discussion on Multiple Imputation. *International Statistical Review* 71(3):619–625.

Schafer JL & Graham JW. 2002. Missing data: our view of the state of the art. *American Psychological Association, Inc.* 7(2):147–177.

Van der Heijden PGM & Escofier B. 2003. Multiple correspondence analysis with missing data. In Escofier B (ed). *Analyse des correspondances. Recherches au Coeur de I'analyse des donnees.* Rennes: Presses Universitaire de Rennes – Societe Francaise de Statistque, pp. 153–170.