Comparison of Alternative Imputation Methods for Ordinal Data

Federica Cugnata[*] Silvia Salini[†]

Abstract

In this paper, we compare alternative missing imputation methods in the presence of ordinal data, in the framework of CUB (Combination of Uniform and (shifted) Binomial random variable) models. Various imputation methods are considered, as are univariate and multivariate approaches. The first step consists of running a simulation study designed by varying the parameters of the CUB model, to consider and compare CUB models as well as other methods of missing imputation. We use real datasets on which to base the comparison between our approach and some general methods of missing imputation for various missing data mechanisms.

Keywords: Missing data; CUB models; single imputation

1 Introduction

In this paper, we consider the CUB (Combination of Uniform and (shifted) Binomial random variables) model [Piccolo, 2003] for the analysis of ordinal

[*] Dipartimento di Economia e Statistica Cognetti de Martiis, Università degli Studi di Torino

[†] Corresponding Author: Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano, Via Conservatorio, 7 - 20122 Milan, Italy. E-mail: silvia.salini@unimi.it

variables. We decided to use the CUB model because on one hand, it allowed us to generate different distributions of ordinal variables, and on the other, it allowed us to interpret them in the specific context of customer satisfaction. In CUB models, the answers to ordinal response items in a questionnaire are interpreted as the result of a cognitive process, where the judgement is intrinsically continuous but is expressed in a discrete way within a pre-fixed scale of m categories. The rationale of this approach stems from the interpretation of the final choices of respondents as a result of two components; a personal *feeling* and some intrinsic *uncertainty* in choosing the ordinal value of the response [Iannario, 2012]. The first component is expressed by a shifted binomial random variable. The second component is expressed by a uniform random variable. The two components are linearly combined in a mixture distribution.

To compare different methods of missing imputation, two simulation studies are done. The first one is designed by varying the CUB model parameters, and the second simulates missing values in a Likert structure. Two real datasets with similar structures have been used in simulation studies. Three missing data mechanisms, namely *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR), are considered and the imputation methods are applied and compared in terms of proportion of correct classification and in terms of CUB model parameter estimation.

The paper is organized as follows. Following the Introduction, Section 2 presents a classification of missing imputation methods. Section 3 is devoted to CUB models. Section 4 presents two simulation studies and the relevant results. Section 5 deals with real datasets. Section 6 draws the conclusions.

2 Missing data imputation for ordinal data

Various approaches can be followed in the treatment of missing values [Little and Rubin, 1987, 2002]. In brief, it is possible to distinguish between *i.*

strategies which allow a complete dataset to be created (complete-case analysis or listwise deletion, available-case analysis, weighting procedures, and imputation-based procedures), and *ii.* strategies which allow direct analyses using model-based procedures; models are specified for the observed data, and inferences are based on likelihood or Bayesian analysis. Moreover, the numerous studies in the literature on missing data highlight that for both approaches, there are numerous procedures and methods of missing imputation, which are often difficult to classify. In this paper, a classification of the various procedures and methods will be proposed, followed by some specific proposals for the imputation of ordinal data.

The most common procedures for imputation of missing data can be classified as:

a. **Univariate**: methods that substantially use information from the distribution of the variable from which the variable itself is missing (i.e., mean, median, mode, random imputation, etc.).

b. **Multivariate**: methods that use the observed pattern for one or more related variables to estimate by means of a model, in which the variable is missing (i.e., linear and nonlinear regression models).

Another common classification of methods is:

a. **Single imputation** (SI), which imputes one value for each missing item.

b. **Multiple imputation** (MI), which imputes more than one value for each missing item to allow for the appropriate assessment of imputation uncertainty. Each set of imputations is used to create a complete dataset, which is analysed by complete-data methods; the results are then combined to produce appropriate estimates that incorporate missing-data uncertainty.

In multiple imputation, each missing value is replaced with multiple imputed values, creating several simulated complete datasets. Rubin [1987] presented the method for combining results from a data analysis performed s times, once for each s imputed dataset, to obtain a single set of results.

Most of the literature on missing data has focused on quantitative data. Less attention has been paid to the treatment of missing imputation methods for ordinal data, although ordinal variables occur in many fields. Existing methods for ordinal data are generally an adaptation of techniques originally designed for quantitative variables. Galati et al. [2012] studied bias arising from rounding categorical variables following multivariate normal (MVN) imputation. Three methods that assign imputed values to categories based on fixed reference points are compared using different scenarios: crude rounding, projected distance-based rounding, and distance-based rounding (DBR).They concluded that these simple methods are generally unsatisfactory for rounding categorical variables following imputation under an MVN model.

Mattei et al. [2012] give a useful and comprehensive review of missing data and imputation methods and present an example from the context of customer satisfaction. They start with a basic discussion of *missing-data patterns*, describing which values are observed in the data matrix and which are missing; and *missing-data mechanisms*, which concern the relationship between missingness and the values of variables in the data matrix. Second, they review four classes of approaches to handling missing data, with a focus on MI, which they believe is the most generally useful approach for survey data, including customer satisfaction data. Third, a simple MI analysis is conducted for the ABC ACSS Survey data[1], and theresults are compared to those from alternative missing-data methods.

Ferrari et al. [2011], in the specific context of qualitative variables, proposed a procedure based on an iterative algorithm where sequentially missing categories for one element are replaced with the corresponding values

[1] The ABC Company has conducted an Annual Customer Satisfaction Survey (ACSS) since 2001,to gather information on its touch points and interactions with customers, through a questionnaire consisting of 81 questions.

observed for the most similar element from a complete dataset (R package ForImp). They employed nonlinear principal component analysis to build statistical indicators. They carried out a simulation study in which they applied the forward method to a real dataset and compared the results of their single multivariate imputation method to other univariate imputation methods. Their iterative method may be extended to other explanatory multivariate techniques.

Stekhoven and Bühlmann [2012] proposed an iterative non-parametric imputation method for mixed-type data, essentially based on random forest (R package missForest). By averaging over many unpruned classification or regression trees, random forest intrinsically constitutes a multiple imputation scheme. Using the built-in out-of-bag error estimates of random forest, they were able to estimate the imputation error without the need for a test set. Evaluation was performed on multiple datasets from a diverse selection of biological fields, with artificially introduced missing values ranging from 10% to 30%. They showed that missForest can successfully handle missing values, particularly in datasets including different types of variables. In their comparative study, missForest outperformed other methods of imputation, especially in datasettings where complex interactions and nonlinear relations were suspected. Additionally, missForest was found to exhibit attractive computational efficiency and was able to cope with high-dimensional data. The idea of using regression and classification trees to input missing values is not new: Iacus and Porro [2007] proposed random recursive partitioning (RRP). This method generates a proximity matrix, that can be used in non-parametric matching problems such as hot-deck missing data imputation and average treatment effect estimation. RRP is a Monte Carlo procedure that randomly generates non-empty, recursive partitions of the data and calculates the proximity between observations as the empirical frequency in the same cell of these random partitions over all the replications.

White et al. [2010] consider multiple imputation. They highlight that the automated procedures widely available in standard software, may hide many assumptions and possible difficulties in the specific context of categorical variables and may give severely biased results. They propose bootstrap methods, penalized regression methods and a new argumentation procedure to solve this problem.

In this paper, we also consider the use of CUB models to inputate missing values for both univariate and multivariate procedures.

3 CUB models

CUB models are a class of statistical models introduced by Piccolo [2003] for the specific purpose of interpreting and fitting ordinal responses. An application of CUB models on marginal ranks can be found in D'Elia and Piccolo [2005a]. In CUB models, ratings are interpreted as the result of two main factors: the personal *feeling* of the subject towards the item and some intrinsic *uncertainty*. Let R be a random variable that assumes m possible categories, $r = 1, 2, 3, \dots, m$. Formally, the probability distribution of the CUB model is given by:

$$P_r(R = r) = \pi \binom{m-1}{r-1} \xi^{m-r}(1-\xi)^{r-1} + (1-\pi)\frac{1}{m}, \qquad r = 1, 2, \dots, m. \tag{1}$$

Since the distribution is well defined when parameters are $\pi \in (0, 1]$ and $\xi \in [0, 1]$, the parametric space is the (left open) unit square:

$$\Omega(\pi, \xi) = \{(\pi, \xi) : \ 0 < \pi \leq 1, \ 0 \leq \xi \leq 1\}.$$

Iannario [2010] proved that such a model is identifiable for any $m > 3$.
The first component is a shifted binomial random variable; ξ is inversely related to the *feeling* of the respondent towards the item: ξ increases when respondents choose low ratings, and vice versa.

The second component is a uniform random variable; π is inversely related to the *uncertainty* in the final judgement. If the respondents manifest a great propensity for extreme indecision in the choice, $\pi \longrightarrow 0$. When the respondent manifests a minimum propensity for extreme indecision and the choice is more resolute and determined mostly by *feeling*, then $\pi \longrightarrow 1$ [Iannario, 2012].

To improve the performance of this structure, an extension of the `CUB` model with covariates has been proposed [Iannario, 2007, Piccolo and D'Elia, 2008]. If p and q covariates are introduced to explain *uncertainty* and *feeling*, respectively, we will denote such a structure as a $\texttt{CUB}(p, q)$ model. The general formulation of a $\texttt{CUB}(p, q)$ model is modelled by two components:

1. A *stochastic component*:

$$Pr(R_i = r \mid \boldsymbol{y}_i; \boldsymbol{w}_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r} (1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right),$$

r = 1,2,...,m; for any $i = 1, 2, \ldots, n$.

2. Two *systematic components*:

$$\pi_i = \frac{1}{1+e^{-\boldsymbol{y}_i \boldsymbol{\beta}}}; \qquad \xi_i = \frac{1}{1+e^{-\boldsymbol{w}_i \boldsymbol{\gamma}}}; \qquad i = 1, 2, \ldots, n,$$

where $\boldsymbol{y}_i = (1, y_{i1}, y_{i2}, ..., y_{ip})'$ and $\boldsymbol{w}_i = (1, w_{i1}, w_{i2}, ..., w_{iq})'$ denote the covariates of the i-th subject, selected to explain π_i and ξ_i respectively. $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_q)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_q)'$ are parameter vectors.

Asymptotic statistical inference for CUB models, an effective EM procedure for maximum likelihood estimators, has been developed and implemented by Piccolo [2006], and related software is freely available [Iannario, 2012]. The simulation routine `simcub()` [Iannario and Piccolo, 2009], can be used to simulate from a given CUB distribution.

When only one variable contains missing values, we can estimate the CUB model based on the subset of the complete data, and then simulate from this CUB distribution to impute each missing value.

When more than one variable has missing data, imputation typically requires an iterative method of repeated imputations. On the basis of the iterative robust model-based imputation proposed by Templ et al. [2011], we propose a CUB approach-based iterative algorithm (iCUB), where, in each step of the iteration, one variable is used as a response variable and the remaining variables serve as the covariates in the CUB models. The proposed iterative algorithm consists specifically of the following steps:

Step 1 Initialize the missing values using a simple imputation technique.

Step 2 Sort the variables according to the original amount of missing values. We now assume that the variables are already sorted, i.e. $M(\boldsymbol{x}_1) \geq M(\boldsymbol{x}_2) \geq ... \geq M(\boldsymbol{x}_v)$ where $M(\boldsymbol{x}_j)$ denotes the number of missing cells in variable $\boldsymbol{x}_j$. Set $I = \{1, ..., v\}$.

Step 3 Set $l = 1$.

Step 4 Denote $mis_l \in \{1, ..., n\}$ the indices of the observations that are originally missing in variable $\boldsymbol{x}_l$, and $obs_l = \{1, ..., n\} \backslash mis_l$ the indices corresponding to the observed cells of $\boldsymbol{x}_l$. Let $\boldsymbol{X}^{obs_l}_{I\backslash\{l\}}$ and $\boldsymbol{X}^{mis_l}_{I\backslash\{l\}}$ denote the matrices with the variables corresponding to the observed and missing cells of $\boldsymbol{x}_l$, respectively.

Based on the subset of the observed cells of $\boldsymbol{x}_l$, estimate the CUB model

$$Pr(\boldsymbol{x}_l^{i\in obs_l} = r \mid \boldsymbol{X}^{i\in obs_l}_{I\backslash\{l\}}) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r}(1-\xi_i)^{r-1} + (1-\pi_i)\left(\frac{1}{m}\right)$$

$r = 1, 2, \ldots, m.$

$$\pi_i = \frac{1}{1+e^{-\boldsymbol{X}^i_{I\backslash\{l\}}\boldsymbol{\beta}}}; \qquad \xi_i = \frac{1}{1+e^{-\boldsymbol{X}^i_{I\backslash\{l\}}\boldsymbol{\gamma}}}; \qquad i \in obs_l\,,$$

We can use a model selection to choose the best model.

Step 5 Estimate the `CUB` model coefficients with the corresponding model in Step 4, and replace each missing value $\boldsymbol{x}_l^{mis_l}$ by a random number generared by a `CUB` model with the estimated `CUB` model coefficients.

Step 6 Carry out Steps 4-5 in turn for each $l = 2, ..., v$.

Step 7 Repeat Steps 3-6 until the imputed values stabilize, i.e. until

$$\textstyle\sum_i (\hat{x}_{l,i} - \widetilde{x}_{l,i})^2 < \delta \qquad \text{for all } i \in mis_l \text{ and } l \in I$$

for a small constant δ, where $\hat{x}_{l,i}$ is the i-th imputed value of the current iteration, and $\widetilde{x}_{l,i}$ is the i-th imputed value from the previous iteration.

The R function for iCUB and the related functions can be downloaded from here: http://users.unimi.it/salini/iCUB.zip.

4 Simulation study

Two simulation studies were conducted. The first considered the imputation for only one variable with covariates and the second considered the imputation for more variables with a Likert structure without covariates, as in Ferrari et al. [2011]. In all cases, the number of Monte Carlo replications was 1.000.

4.1 Imputation for one variable with covariates

In the first simulation, we considered a variable Y generated by a `CUB`(0,0) model and two covariates: X_1 generated by normal distribution $N(y, 0.16)$ and X_2 generated by a `CUB` model with Y as a covariate to explain feeling. The variable Y was generated by a `CUB`(0,0) model with a different number of possible categories ($m = 5, 7, 9$) and for varying parameters over the admissible parameters space, $\pi = 0.1, 0.2, 0.3, \dots, 1$ and $\xi = 0, 0.1, 0.2, 0.3, \dots, 0.99$.

The missing values were selected only in Y using two different missing data patterns:

a) missing completely at random (MCAR),

b) missing not at random (MNAR), in which only the low categories are omitted.

We repeated the experiment for three sample sizes ($n = 200, 500, 1.000$) and three different amounts of missing values (v=5%, 10%, and 20%).

To evaluate the imputation method performance, we considered the percentage of cases correctly imputed and the bias of the estimates of the CUB parameters.

The methods compared in this simulation study were: median imputation (ME), random imputation (RA), CUB(0,0), polytomous ordered logistic regression (PO), CUB(p,q), forward imputation (FO) [Ferrari et al., 2011] and miss-Forest (MF) [Stekhoven and Bühlmann, 2012].

In Figure 1 and Figure 2, for case a) MCAR and for b) MNAR respectively, the value of ξ is plotted in the horizontal axis and the percentage of cases correctly imputed is plotted in the vertical axis, for $\pi = 0.1, 0.5, 0.9$.

The first thing to be noticed, in both cases, is that when uncertainty is high ($\pi = 0.1$), MF, FO, and CUBpq methods behave better than all other methods considered for all levels of feeling. When the missing values are not at random (case b), the polytomous regression, which, like MF, FO, and CUBpq, consider the covariates too, is better than univariate methods median, random and, CUB00. When uncertainty decreases ($\pi = 0.5, 0.9$), the performance of the models MF, FO, and, CUBpq remains better than the other models and, in the case of missing values not at random, changes with

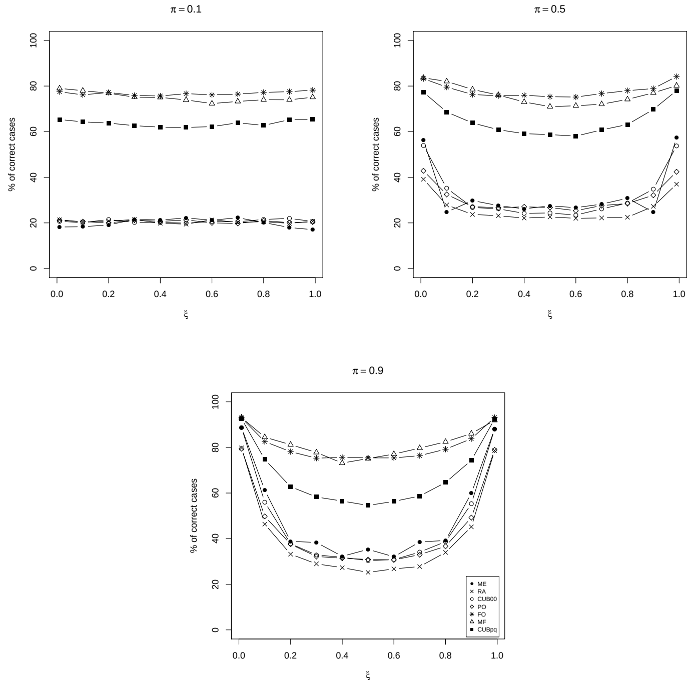


Figure 1: Imputation for one variable m = 5 categories, MCAR v = 5%: Percentage of cases correctly imputed.

the variation of ξ: it improves with increasing ξ (a decrease of feeling) when the missing values are in low categories. In cases where $\pi = 0.9$, that is where there is little uncertainty, and the missing values are not at random, median sometimes behaves better than CUBpq.

When m increases, the performance of all methods worsens a little while maintaining the same pattern of Figure 1 and Figure 2. The same happens when v increases. When n increases, results are stable; therefore, we decided not to report them.

In Figure 3 and Figure 4, for case a) MCAR and for b) MNAR respectively, the box-plots of the bias of the estimates of the ξ parameter are reported. We compared the estimates obtained using different methods of

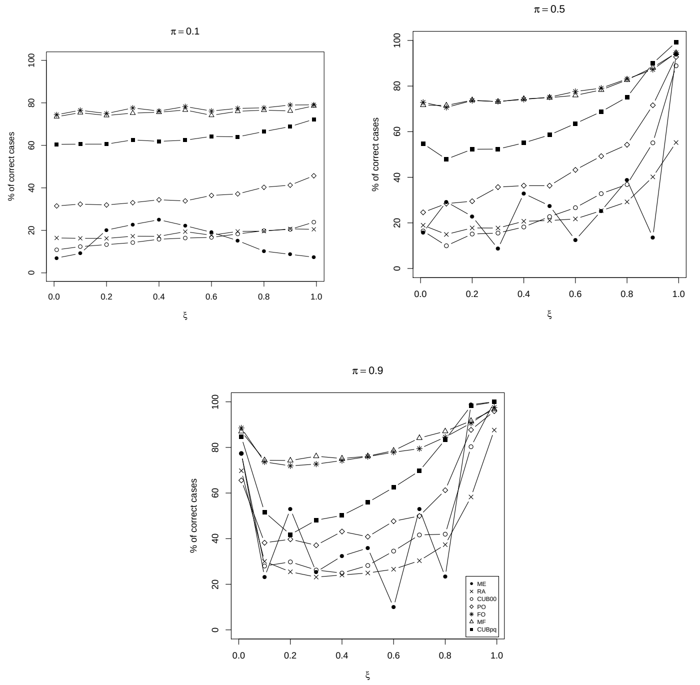
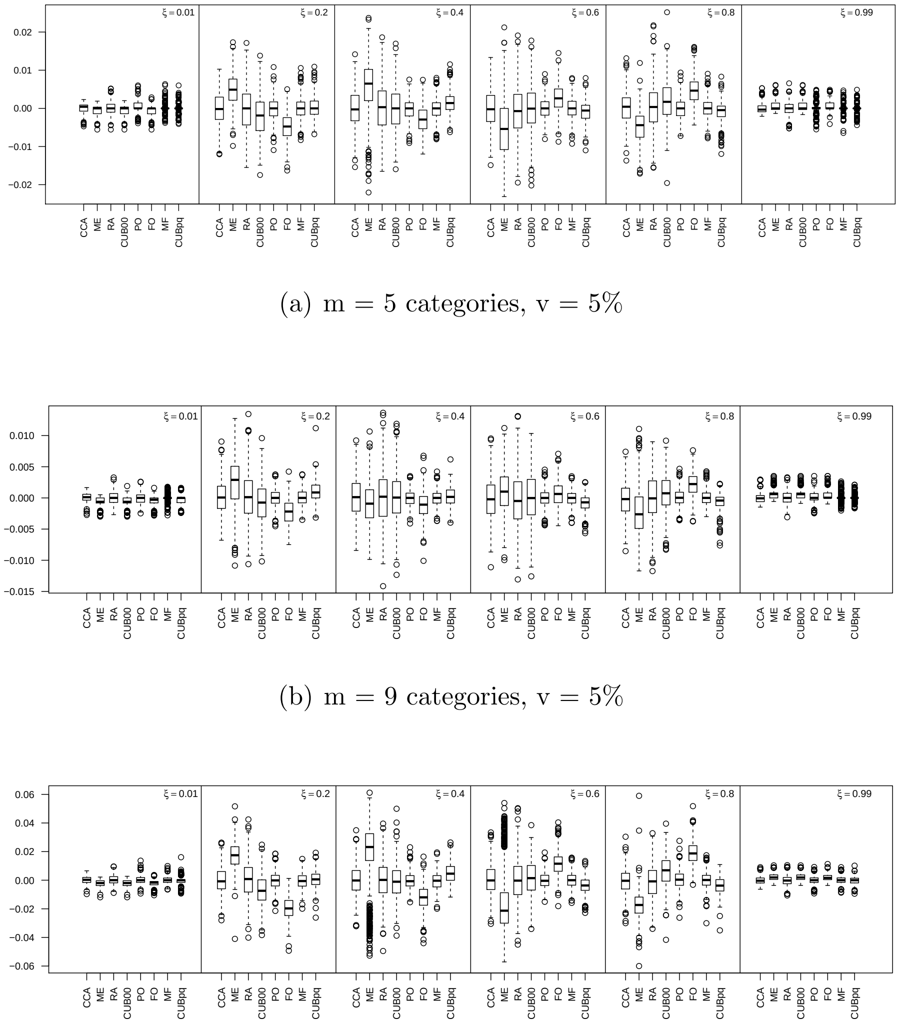


Figure 2: Imputation for one variable m = 5 categories, MNAR v = 5%: Percentage of cases correctly imputed.

imputation as well as complete-case analysis (CCA), in which we ignored incomplete cases.

It was immediately observed that all univariate methods are biased, with the most biased being the median. Among the multivariate methods in some cases, FO has a greater bias than others. The bias is generally reduced when ξ is very small or very large. There is no large variability of results when π varies, so we have not reported the results. If the number of missing v increases, then this increases the variability of the bias. There is, however, no significant change as a result of the m changes.



(a) m = 5 categories, v = 5%

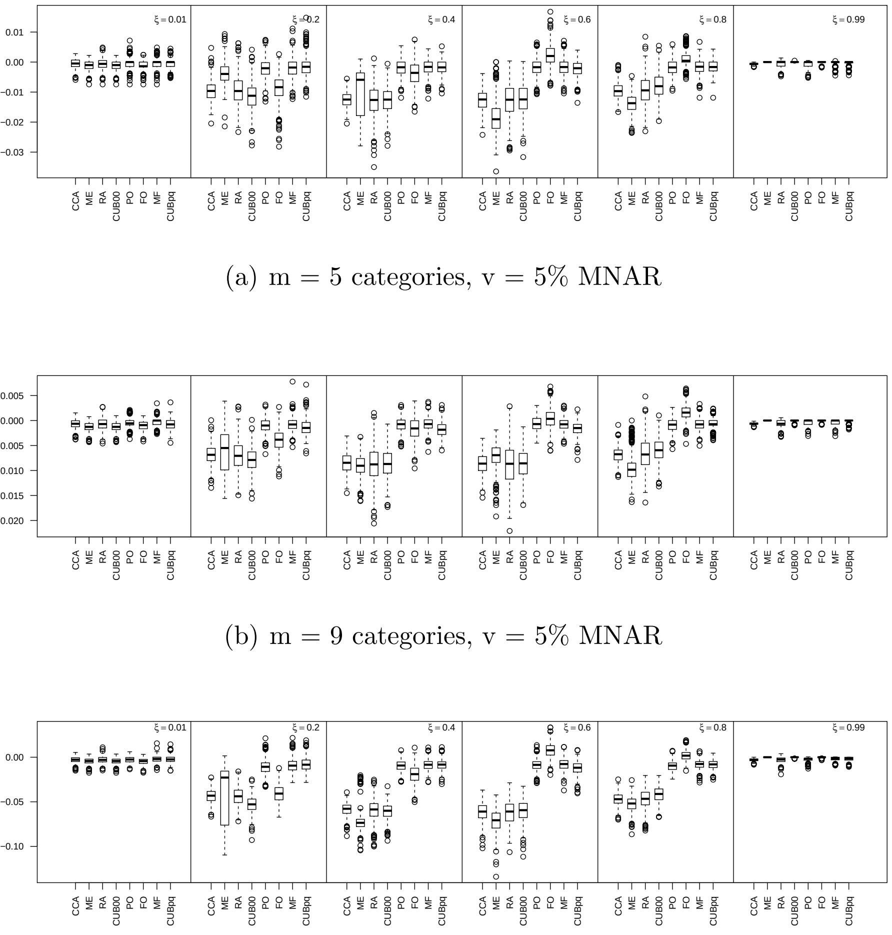
(b) m = 9 categories, v = 5%

(c) m = 5 categories, v = 20%

Figure 3: Imputation for one variable, MCAR: bias of the estimates of the parameter ξ.

4.2 Imputation for more variables with a Likert structure

In this simulation, the missing values existed in more than one variable, and, following the approach of Ferrari et al. [2011], we use all the variables to



(a) m = 5 categories, v = 5% MNAR

(b) m = 9 categories, v = 5% MNAR

(c) m = 5 categories, v = 20% MNAR

Figure 4: Imputation for one variable: bias of the estimates of the parameter ξ.

predict the missing values on the others. The multivariate ordinal variable $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ is generated, following Ferrari and Barbiero [2012] and using the R package GenOrd [Barbiero and Ferrari, 2013]. This approach is able to generate multivariate ordinal variables with the required marginal distributions and correlations. A sample is drawn from a standard multivariate normal rv with correlation matrix R^N and then discretized to yield a sample of ordinal data with assigned marginal distributions by employing a quantile approach. The matrix R^N, ensuring the prescribed correlation matrix R^D on the target variables, is computed through a recursive algorithm. We consider five ordinal categories ($m = 5$) and three different correlation coefficients ($\rho(Y_i, Y_j) = 0.3, 0.5, 0.8$), to assess the effect of different correlations on final results.

We consider two missing data mechanisms:

a) missing completely at random (MCAR),

b) missing not at random (MNAR), in which lowest category are more often omitted.

To evaluate the performance of the imputation method, we considered the mean and the standard deviation of percentage of cases correctly imputed. Table 1 shows the percentage of correct cases in case of MCAR e MNAR with a rate of missing values equal to 5%. The multivariate models obviously impute better than univariate, in particular for low values of ρ. To compare the results in the case MCAR and MNAR, one can observe that the univariate methods worsen in the case of MNAR and those multivariates improve instead, especially when ρ is high.

The procedure was also repeated for a missing rate equal to 10%, and 20% but the amount of missing values seems to only minimally affect the performance of all methods.

To verify our simulation results, we selected various real datasets. The first two examples were built to produce a situation similar to our simulation studies. The first dataset contains the ranking of nine serious problems that could arise in a large metropolitan area. We considered the 2006 wave and

Table 1: % correct cases.

MCAR

	ρ		
Method	0.3	0.5	0.8
ME	25.051	24.697	25.051
	(2.051)	(2.097)	(2.949)
RA	20.857	20.491	19.977
	(2.343)	(2.491)	(2.823)
CUB00	20.251	21.28	20.457
	(1.949)	(2.52)	(2.657)
PO	25.291	29.12	45.749
	(1.991)	(2.88)	(1.851)
FO	23.543	29.28	46.731
	(1.143)	(2.72)	(3.269)
MF	26.149	33.189	52.366
	(2.149)	(2.211)	(3.434)
CUBpq	26.549	33.749	49.291
	(2.651)	(2.151)	(2.009)

MNAR

	ρ		
Method	0.3	0.5	0.8
ME	15.795	15.597	15.816
	(1.395)	(1.203)	(0.184)
RA	20.149	19.824	20.157
	(1.451)	(1.024)	(1.357)
CUB00	16.565	16.573	16.496
	(1.365)	(2.973)	(1.104)
PO	25.08	32.621	54.139
	(2.72)	(1.621)	(2.661)
FO	24.947	32.363	53.643
	(2.253)	(2.037)	(3.557)
MF	17.923	27.149	58.048
	(1.477)	(2.549)	(2.752)
CUBpq	23.259	32.613	57.872
	(2.459)	(2.413)	(2.128)

some covariates of the respondents. See D'Elia and Piccolo [2005b] and Iannario [2007] for more details on the dataset. The values of each variable, as shown in the paper by D'Elia and Piccolo [2005b], can be modeled effectively with a `CUB` model. The fact that data are rankings does not appear to be relevant if one is interested in the construction of univariate models for each emergency, because they are estimated for variables with respect to the marginal analysis of multivariate distribution. Considering the covariates, we applied the same approach to this dataset as in the first simulation study. The second dataset comes from a typical questionnaire completed by airline passengers to evaluate their flight.

The questionnaire contains variables such as *overall experience*, *likelihood to repurchase*, *likelihood to recommend* and *value for money.* There are further questions grouped by topic: *overall booking, check-in, departure, cabin environment* and, *meal.* The evaluation of each item is based on a seven-points scale (from 1 = extremely dissatisfied to 7 = extremely satisfied). Covariates related to the flight and covariates related to the passenger are present. We applied the same approach to this dataset as in the second simulation study.

For these two examples, we considered three different cases of missing patterns, selecting 10% of the available rows each time:

A) missing at random (MCAR)

B) missing in the low categories (MNAR)

C) missing associated to some values of the covariates (MAR)

4.3 Dataset: Emergency in Metropolitan Area

The dataset *Emergency in Metropolitan Area* contains 419 observations. The variables are 1. *Political Patronage*, 2. *Organized Crime*, 3. *Unemployment*, 4. *Pollution*, 5. *Public Health*, 6. *Petty Crimes*, 7. *Immigration*, 8. *Street and*

Waste, 9. *Traffic Transport*. The estimation of the CUB model parameters for the nine variables is reported in Figure 5.

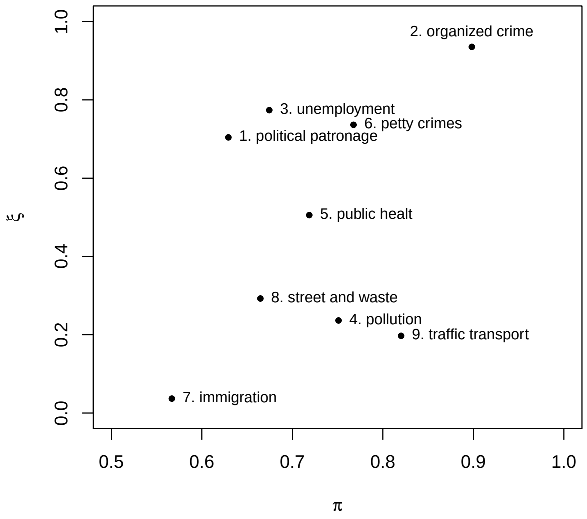


Figure 5: Emergency in Metropolitan Area: CUB models parameters.

The number of generated missing cases is 10% of the total rows. Table 2 reports the percentage of correct cases for the same method used in the first simulation studies. The first observation is that, if we consider the percentage of cases correctly classified, the median tended to work well for these data, in particular in case A, where missing values were randomly selected. In the other cases, the performance of the median was lower. The CUB models exceeded the level of the other models and improved slightly for MCAR and MAR in cases where there was more uncertainty and the level of feeling was high (for example, variable 7, *Immigration*, and variable 9, *Traffic Transport*). This is consistent with the result of the first simulation study shown in Figure 1.

Table 2: Emergency in Metropolitan Area. % of correct cases.

		ME	RA	CUB00	PO	FO	MF	CUBpq
Case A	1	19.55	16.55	18.71	16.41	12.11	23.76	19.58
	2	55.80	38.81	49.53	37.19	40.90	35.63	49.50
	3	22.82	17.54	18.83	15.10	19.00	21.46	18.62
	4	21.39	17.07	19.78	18.87	19.38	21.58	20.92
	5	24.03	14.20	19.09	15.28	18.66	21.64	20.61
	6	22.60	17.62	21.16	16.86	15.15	20.86	20.73
	7	15.11	28.66	41.41	31.91	26.67	15.68	39.85
	8	25.88	16.52	19.75	15.25	16.06	22.46	19.06
	9	20.31	18.61	22.04	19.62	19.43	19.63	23.25
Case B	1	4.82	17.48	19.61	17.34	10.44	19.20	21.34
	2	100.00	51.34	65.89	49.63	62.22	16.34	67.04
	3	27.01	18.69	20.27	17.40	21.97	15.19	20.57
	4	20.90	13.58	11.32	13.92	13.67	22.16	12.42
	5	21.33	14.53	16.58	17.82	23.71	19.33	17.96
	6	32.59	20.10	22.95	17.22	21.37	13.30	21.29
	7	0.00	8.91	2.84	11.26	10.64	17.96	9.57
	8	2.02	9.33	12.29	11.70	13.47	21.51	13.17
	9	0.00	12.59	10.04	14.17	13.69	20.96	11.80
Case C	1	14.50	15.25	19.25	17.13	11.54	21.26	19.30
	2	51.90	37.38	47.24	30.17	34.52	32.71	43.80
	3	27.61	17.61	18.62	17.37	19.43	22.65	19.65
	4	20.65	16.52	18.43	19.86	15.05	21.19	18.96
	5	21.62	15.75	18.98	17.96	15.41	23.70	19.50
	6	19.74	16.42	19.47	17.04	16.10	21.32	20.59
	7	12.77	32.78	48.58	38.22	33.32	18.81	51.72
	8	22.87	15.74	16.81	14.15	16.16	16.74	16.24
	9	17.58	17.38	20.68	21.98	17.04	16.80	22.81

From a model-based point of view, it might be interesting to evaluate the bias in the estimators of the parameters of the `CUB` models in the dataset completed by different methods. Figure 6 shows the estimates of π and ξ in the different datasets and for the four cases for variable 9, *Traffic Transport*. It is immediately evident, that the median sometimes produces biased estimates for π, and the same happens for the other variables. In cases B and C, where the missing values are not at random and are in lower and higher categories, respectively, all the estimators for ξ are obviously biased: the (complete case available) CCA have been created by changing the initial distributions.

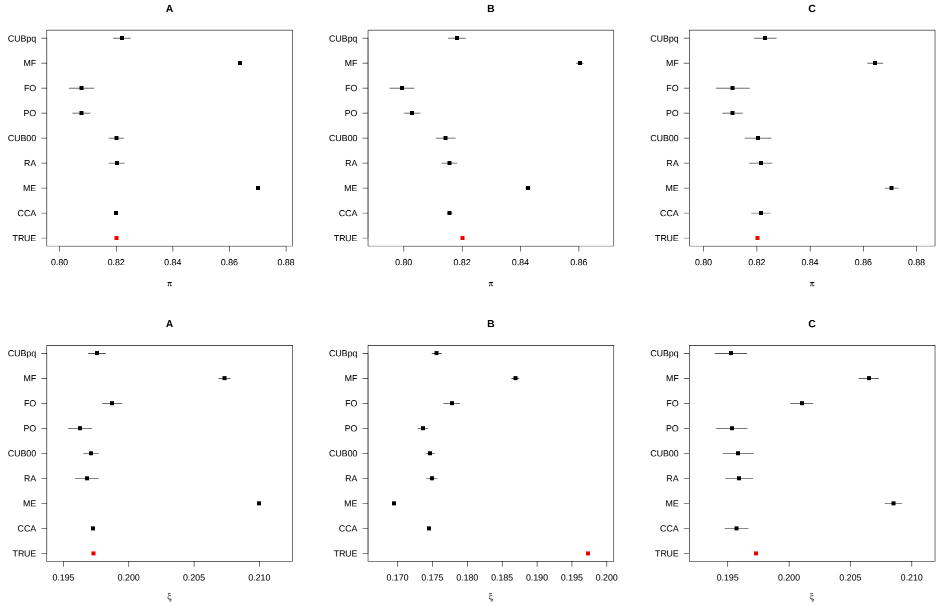


Figure 6: Estimation of the parameters of the `CUB` models, Variable 9. Traffic Transport.

4.4 Dataset: Airline Industry

The dataset *Airline Industry* contains $n = 558$ valid questionnaires collected in 2010. The variables of satisfaction are 1. *Booking*, 2. *Check-in*, 3. *Departure*, 4. *Cabin environment* ,and 5. *Meal*. The estimation of the parameters of the `CUB` model for the five variables is reported in Figure 7.

The number of generated missing cases is 10% of the total rows. Table 3 reports the percentage of correct cases for the same methods used in the second simulation study.

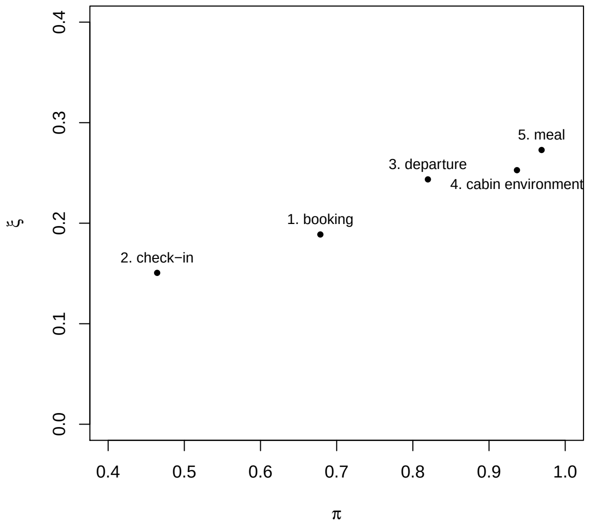


Figure 7: Airline Industry: CUB model parameters.

Table 3: Dataset: Airline Industry: % of correct cases.

		ME	RA	CUB00	PO	FO	MF	CUBpq
Case A	1	41.85	24.55	30.06	31.37	29.48	35.14	33.51
	2	34.88	20.84	24.14	36.94	34.53	32.56	37.25
	3	39.61	23.40	31.40	34.13	42.38	43.90	37.29
	4	37.12	25.66	32.89	34.44	43.03	44.55	35.50
	5	32.32	27.29	34.11	36.98	44.05	48.47	38.25
Case B	1	21.61	14.82	17.71	16.73	18.01	22.98	18.75
	2	10.04	10.94	11.86	20.80	18.27	20.98	21.01
	3	8.04	15.09	16.76	24.94	22.59	29.59	21.79
	4	1.80	19.72	18.03	26.12	30.33	36.01	21.92
	5	0.00	20.69	21.49	28.37	31.16	45.88	24.48
Case C	1	48.03	28.81	35.27	32.11	34.09	39.81	38.64
	2	31.09	21.67	25.37	36.89	36.78	37.34	38.70
	3	31.35	23.18	31.68	35.80	40.75	46.57	37.49
	4	35.59	27.19	31.99	33.36	42.12	43.19	35.94
	5	28.96	28.07	33.00	32.53	41.21	48.12	38.17

As shown in Figure 6, in this dataset the values of the CUB model parameters for variables 4, *Cabin environment*, and 5, *Meal.* fall in the case of little uncertainty and high feeling.

In these cases, the `CUB`pq model performs worse than the other multivariate models. On the contrary, for variable 2, *Check-in* in which uncertainty is present (value of π is low), `CUB`pq is the best solution for the four cases. Moreover, generally, `CUB`pq in both cases of missing at random and missing not at random approaches PO.

From a model-based point of view, it might be interesting to evaluate the bias in the estimators of the parameters of the `CUB` models in the dataset completed by different methods. Figure 8 and Figure 9 show the estimates of π and ξ for the three cases of missing patterns for variable 1, *Booking*, and variable 3, *Departure*. We also report the estimates obtained with the true dataset (TRUE) and with available-case analysis (ACA) which uses only complete data on the variable that is considered.

In this case, MF, being based on an algorithmic approach, always produces estimates that seem more biased for π with respect to the other multivariate estimators. In this case, as in the previous one, the median is completely biased with respect to the estimators for ξ, and in some cases, all the univariate estimators are biased as well. Moreover, MF, except in case B where missing values are concentrated in lower categories, is the most biased of all the multivariate estimators.

5 Conclusion

As is well known, the imputation for missing ordinal data is more complex than it is for continuous data. Proposals that work well are found in the literature, particularly in *forward imputation* (FO) and *missForest* (MF). When the `CUB` model is the preferred model for data analysis, a further opportunity exists to use `CUB` models for imputation. We performed two different

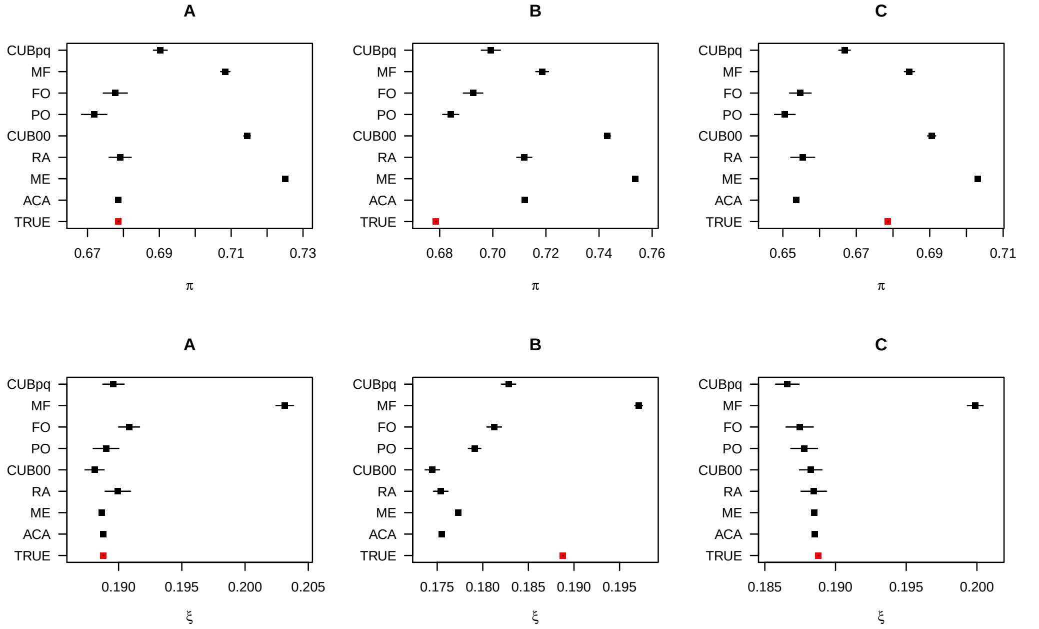


Figure 8: Estimation of the parameters of the `CUB` models, Variable 1, *Booking.*

simulation studies and tested the results on two different real datasets that reflected the characteristics of the simulation studies. When missing values were present only for one variable and covariates related to it were available, the multivariate methods performed better than univariate ones. In cases where there is little uncertainty, the observations are highly concentrated on a few values and the relationship with the covariates is not very strong, the median method may be the best method according to the criteria of correct attribution of the cases. From a model-based point of view, however, we also verified that, as expected, imputation with the median produces biased estimators of the parameters. When data have the classical Likert-scale structure and the missing values are present for some ordinal variables, then, in addition to the classic covariates, ordinal variables may be used in

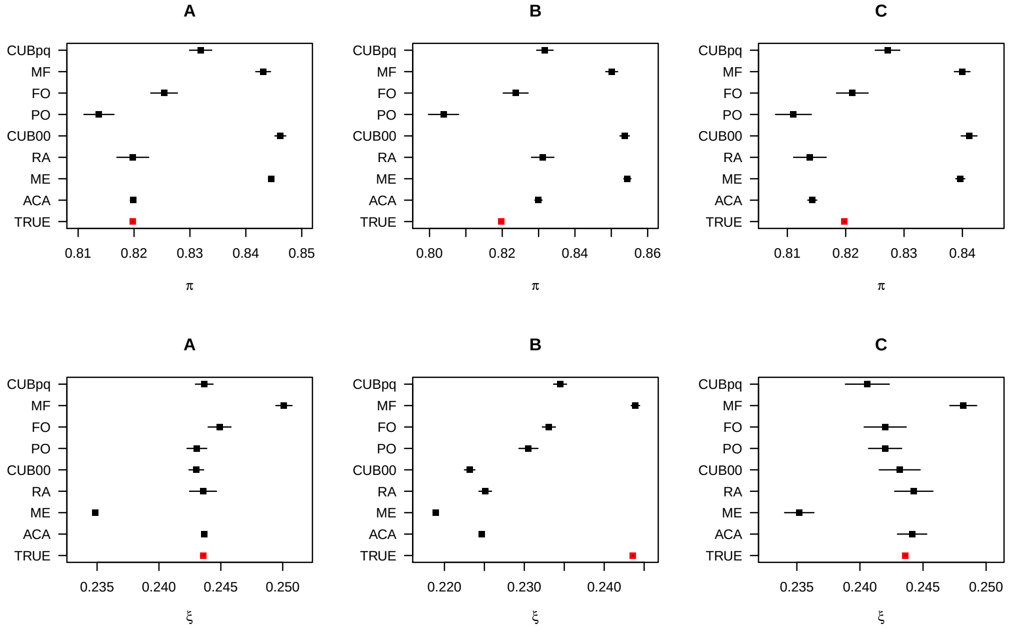


Figure 9: Estimation of the parameters of the CUB models, Variable 3, *Departure*.

the multivariate imputation methods. The simulation study shows that in general *forward imputation* (FO), and *missForest* (MF) perform better than the other multivariate methods. However, when the uncertainty is high, the CUBpq model approach seems better. Reviewing all the results, simulations and applications suggest that the method missForest, in agreement with the conclusions of the authors Stekhoven and Bühlmann [2012], performs best and is also more computationally efficient. The authors, however, proceeding in an optical complete-case, did not raise the question of the properties of the estimators obtained from their method. Our results seem to show a greater bias of the CUB model parameter estimates when using MF for imputation then when using other multivariate procedures. Circumventing this is a challenge that is worth investigating in the future. Another interesting aspect to note is that the multivariate methods work well, even in the case of missing not random, sometimes attaining the same performance as in the case with random missing values.

Acknowledgements

We thank MIUR PRIN MISURA (Multivariate models for risk assessment) for financial support of the project.
We thank also Domenico Piccolo, Maria Iannario and Alessandro Barbiero and the anonymous referees for their precious comments and suggestions.

References

Barbiero, A., Ferrari, P. A., 2013. Genord: simulation of ordinal and discrete variable with given correlation matrix and marginal distributions. R package, 1.1.0.

D'Elia, A., Piccolo, D., 2005a. A mixture model for preferences data analysis. Computational Statistics & Data Analysis 49 (3), 917 – 934.

D'Elia, A., Piccolo, D., 2005b. Uno studio sulla percezione delle emergenze metropolitane: un approccio modellistico. Quaderni di Statistica 7, 121 – 161.

Ferrari, P. A., Annoni, P., Barbiero, A., Manzi, G., 2011. An imputation method for categorical variables with application to nonlinear principal component analysis. Computational Statistics & Data Analysis 55 (7), 2410 – 2420.

Ferrari, P. A., Barbiero, A., 2012. Simulating ordinal data. Multivariate Behavioral Research 47 (4), 566–589.

Galati, J., Seaton, K., Lee, K., Simpson, J., Carlin, I., 2012. Rounding non-binary categorical variables following multivariate normal imputation: evaluation of simple methods and implications for practice. Journal of Statistical Computation and Simulation (ahead-of-print), 1–14.

Goodman, L. A., Kruskal, W. H., 1954. Measures of association for cross classifications. Journal of the American Statistical Association 49 (268), 732–764.

Iacus, S., Porro, G., 2007. Missing data imputation, matching and other applications of random recursive partitioning. Computational statistics & data analysis 52 (2), 773–789.

Iannario, M., 2007. A statistical approach for modelling urban audit perception surveys. Quaderni di Statistica 9, 149–172.

Iannario, M., 2010. On the identifiability of a mixture model for ordinal data. METRON 68 (1), 87–94.

Iannario, M., 2012. Hierarchical CUB models for ordinal variables. Communications in Statistics-Theory and Methods 41.16-17, 3110–3125.

Iannario, M., Piccolo, D., 2009. A program in R for cub models inference. (www.dipstat.unina.it/cub/).

Kenett, R., Salini, S., 2011. Modern Analysis of Customer Surveys: with Applications using R. Statistics in Practice. Wiley.

Little, R., Rubin, D., 1987. Statistical analysis with missing data. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Little, R., Rubin, D., 2002. Statistical analysis with missing data. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley.

Mattei, A., Mealli, F., Rubin, D. B., 2012. Statistical analysis with missing data. In: Kenett, R., Salini, S. (Eds.), Modern Analysis of Customer Surveys: with Applications using R. Statistics in Practice. Wiley.

Piccolo, D., 2003. On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Statistica 5, 86–104.

Piccolo, D., 2006. Observed information matrix for mub models. Quaderni di Statistica 8, 33 – 78.

Piccolo, D., D'Elia, A., 2008. A new approach for modelling consumers' preferences. Food Quality and Preference 19 (3), 247 – 259.

Rubin, D., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley.

Stekhoven, D., Bühlmann, P., 2012. Missforest–non-parametric missing value imputation for mixed-type data. Bioinformatics 28 (1), 112–8.

Templ, M., Kowarik, A., Filzmoser, P., 2011. Iterative stepwise regression imputation using standard and robust methods. Computational Statistics & Data Analysis 55 (10), 2793 – 2806.

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

White, I. R., Daniel, R., Royston, P., 2010. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Computational Statistics & Data Analysis 54 (10), 2267 – 2275.