LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Humera Razzak, Christian Heumann

Predictive performance of a hybrid technique for the multiple imputation of survey data

Technical Report Number 228, 2019
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Predictive performance of a hybrid technique for the multiple imputation of survey data

Humera Razzak*
Christian Heumann†
Department of Statistics, University of Munich

January 6, 2020

Abstract

We discuss the development of a multiple imputation (MI) method for analysing data from the Multiple Indicator Cluster Survey (MICS). A popular chained equations approach to MI called MICE fails to perform sometimes because of computational inefficiency, a complex dependency structure among categorical variables and high percentage of missing information in large scale survey data. On the other hand, a MI approach based on fully Bayesian joint modeling seems to perform very well for categorical variables having complex dependencies but requires transformation and other techniques to impute continuous variables. A hybrid approach is presented here where imputations for a large number of categorical variables are created under a fully Bayesian joint modeling MI technique and regular MICE is used to create imputations for continuous variables. This provides a flexible and practical hybrid MI approach to obtain complete data, which sometimes cannot be obtained when both MI approaches are applied separately. The method proposed is used to analyse data from the MICS 2014 survey women's data investigating the association between various factors and breastfeeding practices among women in Punjab. The relationship between the binary response (breastfeeding) and explanatory variables is modelled using generalized linear models (GLM's). The accuracy of a predictive model is assessed by the area under the receiver operating characteristic (ROC) curve, known as AUROC, and the results obtained under the proposed and existing MI methods are compared. The proposed method outperforms the MICE algorithms CART and PMM in most of the cases requiring less computational time and only minimal tuning by the analyst. The results obtained by the simulation study are supported by a real data example.

Key Words: Complex dependencies; Hybrid multiple imputation

1 Introduction

Many large scale complex surveys such as the Multiple Indicator Cluster Survey or MICS are conducted to recognize forces that contribute to the public health factors that interact at individual, family, community, population, and policy levels. Generally, MICS contains a large number of categorical variables with lots of categories, a complex dependency structure and missing values. For example, the data set of individual women from MICS 2014 used in the real data example has more than 60 per cent data missing on 44 background variables.

Missing data often implicates a biased or an inefficient analysis. Missing mechanisms are: (i) missing completely at random (MCAR), (ii) missing at random (MAR), (iii) missing not at random (MNAR)[1] (Little and Rubin, 2002). MCAR occurs if the probability of missing variable X does not depend on the values of any other variable in the data set (Bennett, 2001). This means that the value of the missing variable is unrelated to any other variable. For example, if the probability that the gender of the child is missing in a household database does not depend on any other variable of the database then MCAR holds. Although it is difficult to detect whether data are MCAR, however, Little (1988) provides a statistical test of MCAR. Schafer and Graham (2002) describe MCAR to be a special case of MAR. With MAR, the probability of having a missing data point in a certain variable is related

*Razzak@stat.uni-muenchen.de
†chris@stat.uni-muenchen.de

[1] MNAR is also called non-ignorable (Ankaia and Ravi, 2011) and not further used in the paper.

to atleast one other variable in the data set but is not related to the variable itself (Allison, 2002). MAR occurs if the probability that a variable X is missing depends on observed data set but not on the variable X itself. For example, if the probability that income of a person is missing depends on profession and age, then the missing data process is MAR. MNAR occurs if the probability that a variable X is missing depends on the variable X itself. For example, if the probability that income is missing dependes on the income itself (often the probability that income is missing is higher for low incomes than for higher incomes) then MNAR occurs.

It is critical to impute the data since multiply imputed data usually provides more accurate inference as compared to complete case analysis or single imputation (Abdella and Marwala, 2005, Little and Rubin, 2002), if the missing data is missing at random (MAR). In recent decades, lots of efforts have been made in the development of statistical methods to treat the problem of missing data. According to studies (Vach and Blettner, 1991 and Kleinbaum et al., 1981), the estimation of regression coefficients can be biased when ad hoc methods and complete case analysis for handling missing data are used. Various approaches based on the Expectation-Maximization (EM) algorithm (Little and Schluchter, 1985), a fully Bayesian analysis (Dellaportas and Smith, 1993), maximum likelihood (Vach and Schumacher, 1993), a mixture of independent multinomial distributions (Dunson and Xing, 2009) and weighted estimating equations (Robins et al., 1994) have been proposed. Multiple-imputation (MI) introduced by Rubin (1987) is nowadays considered as a gold standard to handle the missing data problem. MI replaces missing values in a data set by drawing random values from the predictive posterior distribution of the missing data given the observed data. MI creates M complete data sets. Inference of interest (e.g. mean, regression) can be run on each newly created imputed data set and estimates can be combined by using "Rubin's rules" (Rubin, 1987). One approach for MI is the so-called Fully conditional specification (FCS) model. FCS specifies univariate conditional distributions on a variable-by-variable basis, and draws sequentially missing values iteratively from the estimated conditional distributions. MI by chained equations (MICE) (Raghunathan et al., 2001, van Buuren and Groothuis-Oudshoorn, 2011) is such a fully conditional specification (FCS) approach to MI. The researcher can choose a suitable regression model for each variable, for example classification and regression trees (CART) (Breiman et al., 1984) for categorical variables, predictive mean matching (PMM) (Little, 1988) for continuous variables or just rely on the default method which e.g. uses logistic regression models for binary and PMM for continuous variables. Sometimes, problems of convergence and incompatibility arise when MICE is used for specifying univariate conditional distributions (Gelman and Speed, 1993). MICE fails to perform sometimes due to a complex dependency structure among the categorical variables and a high percentage of missing information which is typical for large scale survey data. Moreover, regression imputations are very time consuming. The R (R Core Team, 2018) package "mice" (van Buuren and Groothuis-Oudshoorn, 2011) implements MICE. The joint modeling (JM) specification is another approach used for MI. JM draws missing values simultaneously for all incomplete variables. JM involves specifying a multivariate distribution for the variables and draws imputations from their conditional distributions by the Markov Chain Monte Carlo (MCMC) methods (Schafer, 1997). Modeling variables of different types can make the specification of a joint distribution very difficult. The Dirichlet Process Infinite Mixtures of Products of Multinomials (DPMPM) is a full Bayesian JM approach (Dunson and Xing, 2009). Si and Reiter (2013) implement DPMPM to impute missing values for categorical variables. The R package "NPBayesImputeCat" by Quanli et al. (2018) implements the DPMPM approach for MI. The implemented DPMPM JM technique to handle missing values is therefore limited to categorical variables and requires transformations (or other tricks) for continuous variables.

The complex dependencies in the MICS data sets containing mixed type covariates (i.e. both categorical and continuous) can be difficult to be identified by the mentioned MI approaches. It has been shown that the MI approach based on DPMPM performs very well for categorical variables having complex dependencies but requires knowledge of complicated models to create the dependence structure between the continuous and the (possibly high) dimensional categorical variables (Murray and Reiter, 2016). These limitations sometimes create serious problems for researchers to obtain complete data sets with mixed type variables. Therefore, we need to develop methods for imputing mixed type data from large scale complex surveys which avoid difficulties of complicated models in high dimensions, combine existing well studied techniques to handle incomplete large scale complex data sets and which are computationally efficient.

We develop a Hybrid Multiple-Imputation (HMI) approach for handling data for the problem described above. We propose to apply the DPMPM MI approach to impute categorical variables having potentially complex dependencies and to use MICE to create imputations for the continuous variables after the categorical variables have been imputed beforehand. The HMI method enables us to utilize the good properties of the DPMPM MI approach and the simplicity of MICE to obtain complete data sets in the mixed data type situation in a flexible and practical manner.

The method proposed is used to analyse data from the MICS 2014 survey women's data. The association between various factors and breastfeeding practices among women in Punjab is investigated. The relationship between the binary response (breastfeeding) and explanatory variables is modelled using generalized linear models (GLM's). The accuracy of the predictive model is assessed by the area under the receiver operating characteristic (ROC) curve, known as AUROC. The predictive performance of the proposed and existing MI methods is compared under a large spectrum of data characteristics. The hybrid mechanism is described in section 2. In Section 3 and 4, cross validation and the measure of performance used for comparison are described. Through simulation studies, we evaluate two software packages used for implementing the hybrid procedure in section 5. Section 6 shows an applications of the proposed method for a real data set. Finally, we give concluding remarks.

2 Proposed hybrid architecture

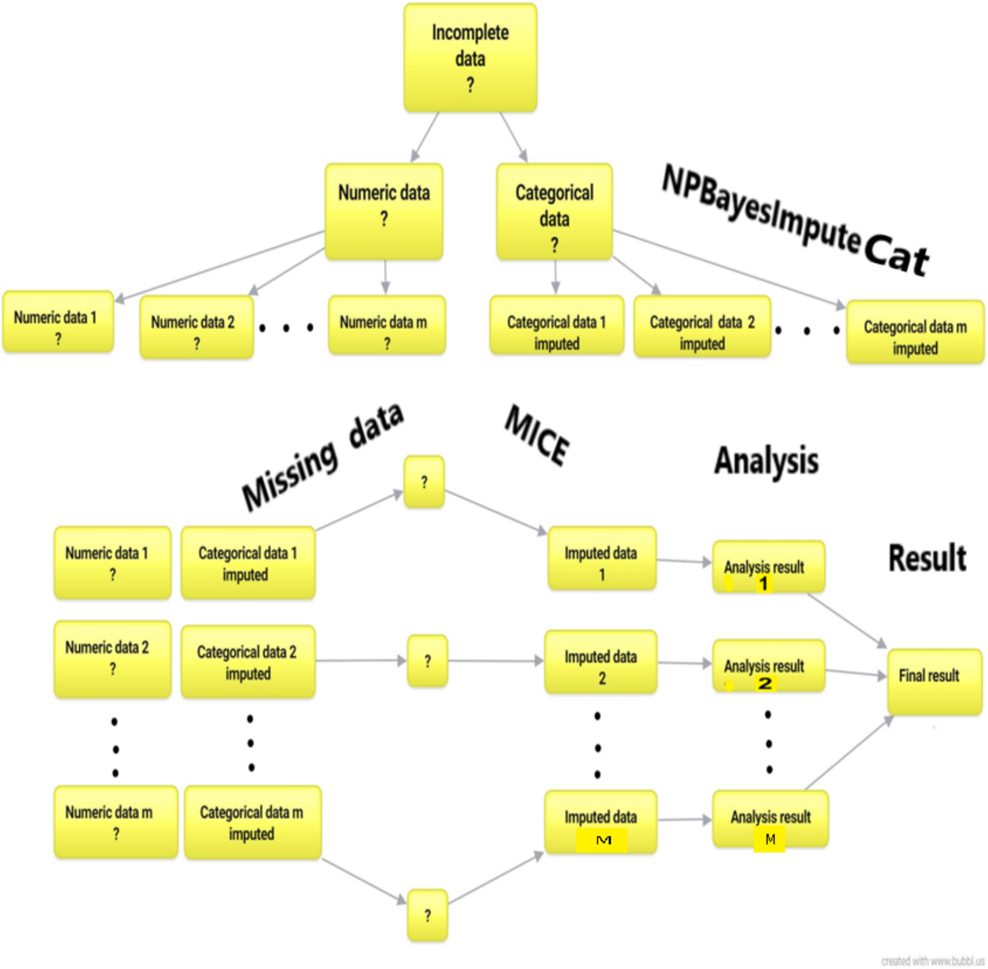


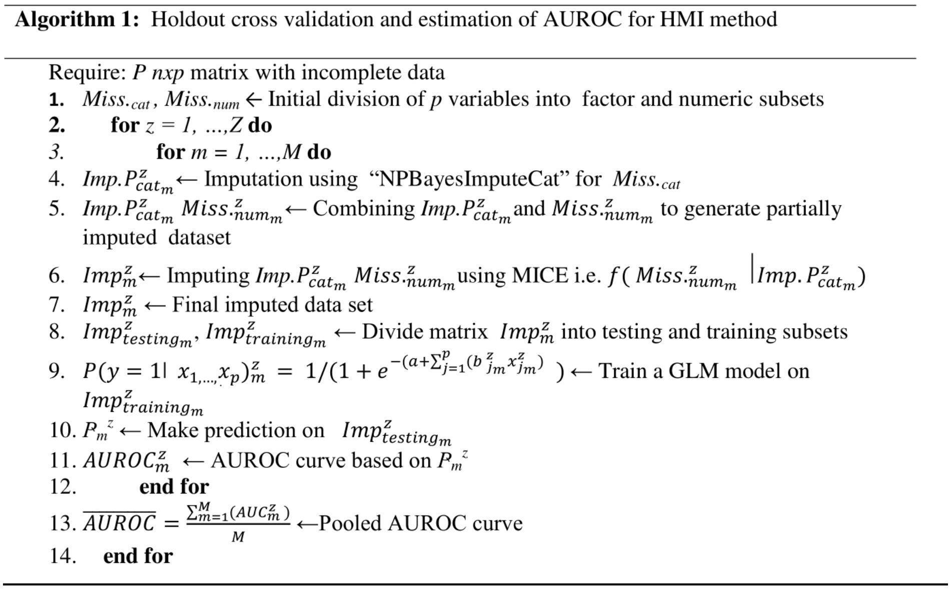
Figure 1: The schema of the hybrid imputation method

The proposed missing data imputation approach is a 3-stage approach. The dataflow diagram (Figure 1) presents the schema of the hybrid imputation method. Step 1: Only the categorical variables ($Imp._{cat}$) are imputed utilizing the R package NPBayesImputeCat (Quanli et al., 2018) which uses a fully Bayesian joint modeling approach. Step 2: The incomplete continuous variables ($Miss._{num}$) are combined with the already imputed categorical variables, $Imp._{cat}$, resulting in M incomplete data sets where values in the continuous variables may be missing and values in the categorical variables have been imputed. M incomplete data sets are made such that the rows of each $Miss._{num}$ data set correspond to the same rows of each $Imp._{cat}$ data set. Hence, one ensures that MI using chained equations for continuous variables uses the information of the imputed categorical covariates for the same unit. Step 3: MICE with various algorithms is used to yield M complete datasets. The R package mice (van Buuren and Groothuis-Oudshoorn 2011) is used for this purpose. The draws from the posterior predictive distribution of the incomplete continuous variables therefore depend on the (in the first step) imputed categorical variables. This process is repeated M times to generate multiple complete data sets. Two Hybrid MI based methods are H.CART and H.PMM. H.CART combines DPMPM with the CART and H.PMM combines DPMPM with PMM. For comparisons, CART, PMM and the Default method in MICE are used.

3 Cross validation

Holdout cross validation is used to assess the predictive performance of a logistic regression model used for the binary response. The logistic regression model [2]is used because the effect of various factors on a binary response (breastfeeding) is analysed later in the real data example. Train and test data sets are generated randomly using a 70% / 30% split. The basic reason to select this method is its simplicity.

4 Evaluation of Performance



Algorithm 1: Holdout cross validation and estimation of AUROC for HMI method

```
Require: $P$ $nxp$ matrix with incomplete data
$Miss._{cat}$, $Miss._{num}$ ← Initial division of $p$ variables into factor and numeric subsets
    for $z = 1, \ldots, Z$ do
        for $m = 1, \ldots, M$ do
$Imp.P^{z}_{cat_m}$ ← Imputation using "NPBayesImputeCat" for $Miss._{cat}$
$Imp.P^{z}_{cat_m}$ $Miss.^{z}_{num_m}$ ← Combining $Imp.P^{z}_{cat_m}$ and $Miss.^{z}_{num_m}$ to generate partially imputed dataset
$Imp^{z}_{m}$ ← Imputing $Imp.P^{z}_{cat_m}$ $Miss.^{z}_{num_m}$ using MICE i.e. $f(Miss.^{z}_{num_m} \mid Imp.P^{z}_{cat_m})$
$Imp^{z}_{m}$ ← Final imputed data set
$Imp^{z}_{testing_m}$, $Imp^{z}_{training_m}$ ← Divide matrix $Imp^{z}_{m}$ into testing and training subsets
$P(y = 1 \mid x_{1,\ldots,}x_p)^{z}_{m} = 1/(1 + e^{-(a+\sum_{j=1}^{p}(b^{z}_{jm}x^{z}_{jm})})$ ← Train a GLM model on $Imp^{z}_{training_m}$
$P_m{}^{z}$ ← Make prediction on $Imp^{z}_{testing_m}$
$AUROC^{z}_{m}$ ← AUROC curve based on $P_m{}^{z}$
            end for
$\overline{AUROC} = \frac{\sum_{m=1}^{M}(AUC^{z}_{m})}{M}$ ←Pooled AUROC curve
    end for
```

The area under the receiver operating characteristic (ROC) curve, known as AUROC, is used to compare different MI methods. For more detail see McNeil and Hanley (1984), Metz (1986), Swets (1979) and Wieand et al., (1989). Algorithm 1 describes how the AUROC curve is pooled[3].

[2] A special generalized linear model with link logit.

[3] The arithmetic mean is taken of M AUROC values obtained by M fitted GLM's

5 A small scale study

A small scale study is conducted to examine the impact of MI by our proposed method. The incomplete data is generated MAR to compare the methods in a realistic data situation. The number of categorical variables is kept higher than the number of continuous variables due to the fact that the simulation is aimed to be similar to the survey data. Table 1 represents a large spectrum of practially occurring data characteristics used for generating data according to a variety of settings. Series of simulations are run varying the correlation among covariates, the number of imputations, different hybrid methods and the algorithms used in MICE.
Simulation study: Five (X_1, X_2, X_3, X_4 and X_5) dimensional correlated normal data is generated using the R package Binorm (Demirtas et al., 2014). The marginal distribution of $X_1, X_2, X_3 \sim Bernoulli(0.5), X_4 \sim N\ (80, 250)$ and $X_5 \sim N\ (80, 250)$. The correlation structure is given as:

$$\text{H}= \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}$$

Here, $\rho = 0.5$ and 0.7 stand for moderate and high correlations, respectively. The dichotomization of X_1, X_2 and X_3 is based on the following criteria

$$P(X_i = 1) = P(X_i \leq \mu_i) = 0.5.$$

Where $i = 1, 2, 3$ and 0.5 is the mean value of X_i. A population consists of $N = 1000$ observations is generated. By defining and standardizing $\mu_y = \beta_1 X_{i1} + ... + \beta_p X_{ip}$, $\theta = \beta_{true} = (2, 2, 2, 2, 2), p = 5, i = 1...N$. We generate the covariate dependent binary response y using the probability

$$\pi = \frac{1}{[1+\exp(a - b\mu_y)]}.$$

Where $a = -1$ and $b = -8$. By using the following probability, it is ensured that the missing mechanism is MAR in each variable:

$$\text{p}= 1 - \frac{\epsilon^{(-0.5-\mu_y)}}{(1+\epsilon^{(-0.5-\mu_y)})}.$$

The probability defined above yields about 20% of the observations in X_i and y to be missing (at random). R version 3.0.1 is used to perform all calculations. The packages mice, version 2.17 and NPBayesImputeCat, version 0.1 are used to perform MICE for continuous data and Non-Parametric Bayesian Multiple Imputation for categorical variables, respectively.

Table 1: Simulation settings

Perameters	Notations	Values
Population size	N	1000
No. of covariates	p	5
No.imputations	Imp.	$2, 5, 10$
Correlation	ρ	$0.5, 0.7$
Prior specifications	a_α, b_α	$0.25, 0.25$
Missing mechanism		MAR
Algorithms		CART, PMM, Default, DPMPM
No. of mixture componenta	k	80
No.simulations	Z	$50, 200$

Various numbers of imputations ($M = 2, 5, 10$) are generated using five MI methods for moderately and highly correlated simulated data. Numbers of imputations are small to facilitate beginners because manuals and descriptions for statistical software often use small number of imputations in examples whereas, large number of imputations is made for better estimates. A total of 200 simulations were made for each method. The binary response is modeled using GLM's depending on various categorical and continuous covariates. Predictive performance of the GLM's for binary response is compared using pooled AUROC curves after cross validation. The actual times taken for MI using all methods for high and moderate correlated data sets are displayed in Tables 2 and 3 respectively. Median values of pooled AUROC curves for all MI methods and different correlations are shown in Table 4. Since no noticeable differences in the posterior distributions of y are observed for different prior specifications in the similar study by Si and Reiter (2013), we limited the examination of different vague prior specifications for a_α and b_α to $(a_{\alpha=0.25}, b_{\alpha=0.25})$. The maximum number of mixture components k is set to 80 in all simulation runs. The AUROC values for moderate and high correlated, cross validated complete data sets are 98 per cents. These values can be used as benchmark (theoretical AUROC) for comparison. For moderate correlation, the predictive performance of the Hybrid MI methods is low but at least comparable to the MICE MI methods (see Figure 2). Figure 3 shows that for the highly correlated data, the Hybrid MI methods perform better than PMM and Default. The performance of H.CART is slightly less than CART. The number of multiple imputations has no significant effect on the results in the simulation study. It is noticeable, that although there is no significant difference among computational time taken for two Hybrid MI methods and Default MI method, but this difference increases when comparison is made with PMM and CART.

Table 2: Similated data $\rho = 0.7$: The time to complete M multiple imputation by variants of MI across 200 simulations

Imp.	Default	CART	PMM	H.CART	H.PMM
2	15.12m	15.74m	25.52m	13.51m	15.50m
5	36.47m	38.08m	59.48m	30.99m	36.45m
10	1.13h	1.21h	1.83h	1.04h	1.17h

Note: m = minutes and h = hours to complete multiple imputation on this subset.

Table 3: Similated data $\rho = 0.5$: The time to complete M multiple imputation by variants of MI across 200 simulationss

Imp.	Default	CART	PMM	H.CART	H.PMM
2	12.19m	16.92m	25.46m	13.79m	15.65m
5	28.84m	41.09m	59.80m	32.63m	35.40m
10	53.35m	1.34h	1.85h	1.02h	1.15h

Note: m = minutes and h = hours to complete multiple imputation on this subset.

Table 4: Simulated data: Median values of the pooled AUROC curve for various MI methods across 200 simulations

	$\rho = 0.5$					$\rho = 0.7$				
Imp.	Default	CART	PMM	H.CART	H.PMM	Default	CART	PMM	H.CART	H.PMM
2	0.9511	0.9593	0.9507	0.9282	0.9270	0.9658	0.9720	0.9662	0.9715	0.9700
5	0.9533	0.9593	0.9505	0.9286	0.9272	0.9658	0.9718	0.9660	0.9714	0.9703
10	0.9509	0.9592	0.9504	0.9288	0.9273	0.9657	0.9718	0.9662	0.9713	0.9703

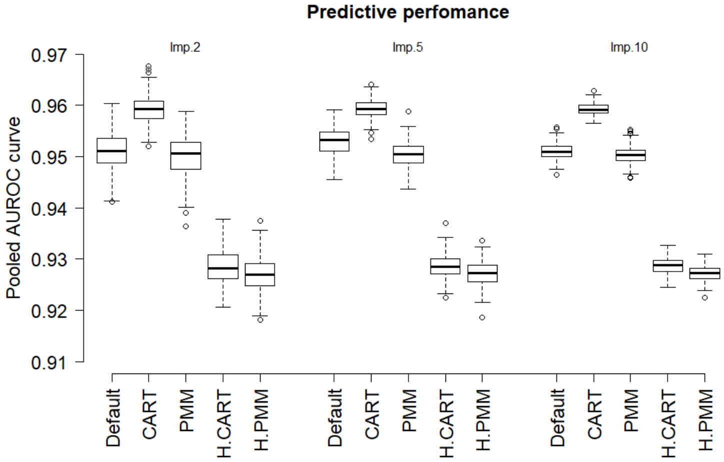


Figure 2: Simulation study: Boxplots of pooled AUROC under various MI methods for$\rho = 0.5$

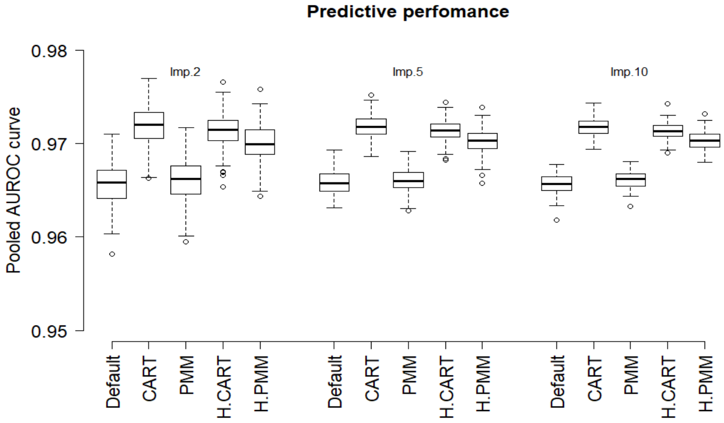


Figure 3: Simulation study: Boxplots of pooled AUROC under various MI methods for$\rho = 0.7$

6 Real data-based example: Imputation of MICS Background Variables

We use the MICS 2014 women's data as real data based example. This data contains more than 200 variables with 61286 observations around all districts of Punjab. Due to compatibility problems and for demonstration purposes, we include only forty four background variables in the analysis. Women's background characteristics like demographics, age, education, motherhood and recent births are included in this data set. The number of categorical variables is high as compared to continuous variables. According to WHO (2003), breastfeeding is important for the well-being of both child and a mother. MICS 2014 women's data can be used to determine the effect of various factors affecting feeding practices in Punjab. We treat item non response as MAR. Information on the global MICS may be obtained from mics.unicef.org and information about Bureau of Statistics, Punjab is available at bos.gop.pk. Fifty sampling simulations are run and $M = 5$ completed data sets are generated for each MI method. The binary response (Ever Breastfeed) compromising two categories (Yes / No) is modeled using the GLM's depending on various categorical and continuous covariates. The AUROC is pooled for each MI method after cross validation. Predictive performance of the GLM's for two hybrid methods is slightly less than the Default MI method and better than the remaining two, see Figure 4. Surprisingly, there is a great difference between the computational time required by the proposed and the MICE MI methods. It can be seen in Table 5 that the time taken by MICE methods is reduced from days to hours when the proposed methods are applied. The median values of the pooled AUROC curves for all methods can be seen in Table 6.

Table 5: Real data: The time to complete 5 multiple imputations by variants of MI across 50 simulationss

Imp.	Default	CART	PMM	H.CART	H.PMM
5	1.93d	1.88d	1.80d	10.78h	11.59h

Note: d = days and h = hours to complete multiple imputation on this subset.

Table 6: Real data: Median values of the pooled AUROC curve for various MI methods across 50 simulations

Imp.	Default	CART	PMM	H.CART	H.PMM
5	0.94	0.63	0.82	0.88	0.88

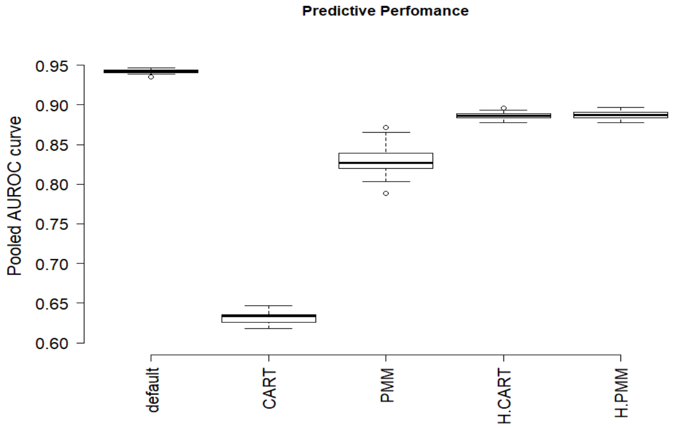


Figure 4: Real data : Boxplots of pooled AUROC obtained for 5 imputations under various MI methods

7 Concluding remarks

We proposed a computational efficient hybrid MI method. Our proposed method makes it possible to MI both types of variables (categorical with large numbers of outcomes and continuous) in survey data in the presence of complex dependencies. This method combines MI by chained equations and mixtures of multinomial. In this method, chained equations of MI continuous variables are made dependent on categorical variables MI by DPMPMs. This approach can prove to be very appropriate for a large number of variables with complex association structures especially coming from sample surveys. To implement this method no knowledge of complicated models is required. The dependence among continuous and categorical variables can be made through an easy engine. Better predictive performance with minimum computational time as compared to the existing methods is partly achieved in simulation studies. However, of note, one limitation of the proposed method is that the information available in the continuous variables is not used for imputing the categorical variables. The source of low rates of AUC for hybrid methods as compared to CART in simulation studies is still unknown. Further research for complex simulation studies, large-sample results or large number of imputations could be needed to find an answer.

8 References

M. Abdella and T. Marwala. The use of genetic algorithms and neural networks to approximate missing data in database. *IEEE 3rd International Conference on Computational Cybernetics*, 24: 207-212, 2005.

P.D. Allison. *Missing Data.* Thousand Oaks, CA: Sage Publications, 2002.

N. Ankaiah and V. Ravi. A Novel Soft Computing Hybrid for Data Imputation. In Proceedings of the 7th International Conference on Data Mining (DMIN), Las Vegas, USA, 2011.

L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone. *Classification and Regression Trees.* Wadsworth: Belmont, 1984.

D.A. Bennett. How can I Deal with Missing Data in my Study. *Australian and New Zealand Journal of Public Health*, 25:464 469, 2001.

P. Dellaportas and A.F.M. Smith. Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics*, 42:443-459, 1993.

D.B. Dunson and C. Xing. Nonparametric Bayes Modeling of Multivariate Categorical Data. *Journal of the American Statistical Association*, 104:1042-1051, 2009.

H. Demirtas, A. Amatya and B. Doganay . BinNor: An R Package for Concurrent Generation of Binary and Normal Data. *Communications in Statistics - Simulation and Computation*, 43(3):569 579, 2014. A. Gelman and T. P. Speed. Characterizing a Joint Probability Distribution by Conditionals. *Journal of the Royal Statistical Society*, 55(1), 185-88, 1993.

D.G. Kleinbaum, H. Morgernstern and L.L. Kupper. Selection Bias in Epidemiological Studies. *Am. J. Epidem.*, 113:452-463, 1981.

R.J.A. Little and M.D. Schluchter. Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. *Biometrika*, 72: 497-512, 1985.

R.J.A. Little. Missing-Data Adjustments in Large Surveys. *Journal of Business Economic Statistics*, 6: 287-296, 1988.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* Hoboken, NJ: John Wiley Sons, 2002.

B.J. McNeil and J.A. Hanley. Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Med Decis Making*, 4: 137-50, 1984.

C.E. Metz. ROC Methodology in Radiological Imaging. *Invest Radiol*, 21: 720-33, 1986.

J.S. Murray and J.P. Reiter. Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516), 1466-1479, 2016.

W. Quanli, M.V. Danial, J.P. Reiter and H. Jigchen. *NPBayesImputeCat: Non-Parametric Bayesian Multiple Imputation for Categorical Data*, 2018. Url: https://CRAN.R-project.org/package=NPBayesImputeCat. R package version 0.1.

J. M. Robins, L.P. Zhao, A. Rotnitzky and S. Lipsitz. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, 89: 846-866, 1994.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley series in probability and mathematical statistics. John Wiley Sons, New York, USA, 1987.

T.E. Raghunathan, J. M. Lepkowski, J. van Hoewyk, and P. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27 (1), 85-95, 2001.

R Core Team. R: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018. Url: https://
www.R-project.org.

J.A. Swets. ROC analysis applied to the evaluation of medical imaging techniques. *Invest Radiol*, 14:109-21, 1979.

J.L. Schafer. *Analysis of Incomplete Multivariate Data.* CRC Press, 1997. ISBN: 978-1-4398-2186-2.

J.L. Schafer and J.W. Graham. Missing data: Our view of the state of the art. *Psychological methods*, 7:147-177, 2002.

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.

M. Vach and M. Schumacher. Logistic regression with incompletely observed categorical covariates: a comparison of three approaches. *Biometrika*, 80:353-362, 1993.

W. Vach and M. Blettner. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values of confounding variables. *Am. J. Epidem.*, 134: 895-907, 1991.

S.van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45: 1-67, 2011. Doi: http://dx.doi. org/10.18637/jss.v045.i03.

S. Wieand, M.H. Gail, B.R. James and K.L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76: 585-92, 1989.

WHO World Health Organization. *Community-based Strategies for Breastfeeding Promotion and Support in Developing Countries.* Dept. of child and adolescent health and development, Geneva, World Health Organization, 2003.