Multiple Imputation of Missing Data Using Non-ignorable Data Augmentation Techniques in Industries

Ms. Lavanya V, Assistant Professor & Research Scholar, School of Science and Computer Studies, CMR University, Bengaluru. **Email:** lavanya.v@cmr.edu.in

Dr. T.A.Ashok Kumar, Professor, School of Science and Computer Studies, CMR University, Bengaluru
Email: ashokkumar.t@cmr.edu.in

Abstract

Data mining necessitates a pre-processing task to prepare and clean the data, ensuring its quality. Missing data values occur when no data is stored for a variable in an observation. Imputation is a popular method due to its conceptual simplicity and because it maintains the same number of observations as the complete dataset. The industry trusts this approach to manage large datasets effectively. Multiple imputations is a widely used technique for analyzing incomplete data, often assuming a missing-at-random mechanism, which means the response mechanism does not depend on the missing variable. However, assuming ignorable non-response can result in biased estimates if the missingness is actually non-ignorable.

In this research, we adopt the selection model approach, specifying both the response model and the respondents' outcome model to capture the joint model of the study variable and the response indicator. The proposed data augmentation algorithm utilizes the respondents' outcome model and incorporates a semi-parametric estimation. This multiple imputation method performs well if the specified response model is accurate.

In this paper, we propose a multiple imputation method for cases of non-ignorable non-response using data augmentation techniques, which are effective for datasets with a high proportion of missing data. Existing imputation methods often fail to meet analysis requirements due to low accuracy and poor stability, especially as the rate of missing data increases. Patterns of missing data can pertain to either cases or attributes. The global impact of the imputed data is assessed through several statistical tests, and it is found that the imputation value is high with the DarbouX variate, which fixes the infimum and supremum of the missing data.

Keywords: Imputation, Knowledge Transfer, missing data, data patterns, multiple imputations, Data Augmentation, Missing Random Mechanism, DarbouX Algorithm, Naive Bayesian.

Introduction

Non-response often leads to the assumption that missing data is ignorable, using the missing at random (MAR) mechanism. This assumption implies that the probability of missingness given the observed variables does not depend on the missing variable itself (Little & Rubin, 2002). While MAR is reasonable in many situations, there are cases where it is more realistic to assume non-ignorable nonresponse. For example, the 'not missing at random' (NMAR) assumption is

often preferred in the analysis of income surveys or election polls, as nonresponse for the study variable, given observed variables, is likely correlated with the unobserved values (Peress, 2010).

In such cases, specifying the joint modeling of the missing variable and the missing mechanism is necessary to obtain consistent statistical analysis results when non-ignorable missing data is present. In contrast, the MAR assumption does not require specifying the missing mechanism model. Greenlees et al. (1982) proposed a conditional mean imputation method, assuming a normal distribution for the outcome model and a logistic model for the response model. Qin et al. (2002) considered a semi-parametric likelihood approach by combining a nonparametric outcome model with a parametric response model. Chang and Kott (2008) and Kott and Chang (2010) estimated response model parameters using the calibration method without making assumptions about the outcome model.

The pattern-mixture model specifies two separate conditional distributions of the study variable given covariates for respondents and non-respondents. In an earlier study, Rubin (1977) proposed generating imputed values from the conditional distribution of the study variable for non-respondents. Rubin initially assumed two parametric normal outcome models for both respondents and non-respondents and then imputed missing values from the non-respondents' outcome model. The imputation is drawn from the posterior distribution of parameters of the non-respondents' outcome model based on prior information about the relationship between the two model parameters. Little and Wang (1996) considered a bi-variate normal pattern-mixture model with parameter restrictions between mixture models. Giusti and Little (2011) described a sensitivity analysis to assess the effect of non-ignorable nonresponse using the pattern-mixture model to avoid under-identification problems.

Multiple imputations has been applied in several settings under the NMAR assumption. Using the selection model approach, Durrant and Skinner (2006) proposed a multiple imputation method using a data augmentation algorithm with parametric outcome and response models. Galimard et al. (2016) proposed multiple imputation by chained equations (MICE) using Heckman's selection model. Rubin (1987) used the pattern-mixture approach to introduce a linear regression approximation with the closest predictor generated by an appropriate imputation model.

This approach was applied by van Buuren et al. (1999) to handle missing covariates in survival analysis. Similarly, Carpenter et al. (2007) generated imputed values using multiple imputation with the MAR model and then adjusted these values to fit the NMAR model. In this paper, we propose a multiple imputation method for non-ignorable non-response based on the selection model approach. Instead of specifying the outcome model for the hypothetical complete data, we specify the respondents' outcome model in the data augmentation algorithm for imputing missing values, following model assumptions by Riddles et al. (2016). Additionally, we incorporate a semi-parametric estimation of the respondents' outcome model.

The main advantage of our proposed method is that the respondents' outcome model is testable with the observed data. This makes the proposed method more robust to misspecification of the imputation model compared to previous studies, which either assume a non-testable outcome model or require sensitivity analysis to find a plausible imputation model. A simulation study

demonstrates that the proposed method remains robust to misspecification of the response model unless the fitted response model significantly deviates from the true response/nonresponse pattern.

Durant and Skinner's (2006) data augmentation method is summarized as a comparable method. We introduce the proposed data augmentation algorithm and present two simulation studies. Finally, the proposed method is applied to datasets from various industries.

2. Non-ignorable missing using Data Augmentation

2.1Parametric approach

An imputation method for a parametric outcome model is based on the relationship between the non-respondents' outcome model and the respondents' outcome model. The imputed values for non-respondents are generated by

$$y_i^* \sim \int (x|y, \delta = 0) \neq \int (x|y, \delta = 1)$$

Where $y_i^*\ y_i^*$ denotes the generated imputed value for unit i. However, we cannot directly use the rejection sampling approach to generate missing values because the odds of nonresponse probability $O(x1,y;\ \varphi)\ O(x1,y;\ \varphi)$ is not bounded above in general. One possible alternative approach which does not require iterative algorithm is to use the parametric fractional imputation proposed by Kim (2011). To implement the parametric fractional imputation idea, first draw L imputed values for uniti from the estimated respondent's outcome model, and then select a final impute value $y_i^*\ y_i^*$ with the probability proportional to the fraction weights $w_{ij}^*, j = 1,..,L$ $w_{ij}^*, j = 1,..,L$, where

$$w(\ ij)\hat{}* = O\ (x1i, y\ (j)\ i\ ;\ \varphi)/\ \sum Lj = 1\ O(x1i, y\ (j)\ i\ ;\ \varphi)$$

And y (j) i is the jth imputed value candidate generated from the estimated respondents' outcome model. See Kim (2011) for details. Propose a multiple imputation method using the data augmentation algorithm with a parametrically specified respondent's outcome model. The proposed data augmentation algorithm draws M multiple samples from the constructed distribution

2.2 Semi parametric approach

Semi-parametric approach the parametric respondents' outcome model given in the previous section is testable by usual goodness-of-fit test, but this approach can be still sensitive to model misspecification. In this section we propose a more robust semi-parametric approach

$$x^{*}_{i,semi} = \sum_{j=1}^{n} \delta_j K_h(y_j, y_i) O(y1_i, x_i^{(j)}; \emptyset) x_i^{(j)} / \sum_{j=1}^{n} \delta_j K_h(y_j, y_i) O(y1_i, x_i^{(j)}; \emptyset),$$

where $K_h\ K_h$ is a symmetric unimodal kernel function with a bandwidth h. Note that the uncertainty for both within variance over missing units and between variance over M repeated imputation for multiple imputation will not be correctly captured if the values of $x^{*}_{i,semi}\ x^{*}_{i,semi}$ are directly imputed on the missing unit.

We use residuals to recover the within variance and bootstrap sample for each imputed data set to adjust the between variance over repeated imputations

In each iteration, impute $x_i^{**}\ x_i^{**}$ on the missing unit $i, x_i^{**} = x^{*}_{i,semi} + \hat{e_i}\, i, x_i^{**} = x^{*}_{i,semi} + \hat{e_i}$, where $\hat{e_i}\ \hat{e_i}$ is randomly selected from the residuals $\hat{e_j}\ \hat{e_j}$, where $\hat{e_j} = x_j - \hat{x_{i,semi,j}} = 1, \ldots, n_R\ \hat{e_j} = x_j - \hat{x_{i,semi,j}} = 1, \ldots, n_R$

The overall multiple imputation is computed by

$$\hat{\mu y} = \frac{1}{M} \sum_{k=1}^{M} \hat{\mu y}^{(k)}$$

And the variance estimate of the mean estimator is calculated from the Rubin (1987)'s variance formula,

$$\hat{V y} = W_M + (1 + M^{-1}) B_M ,$$

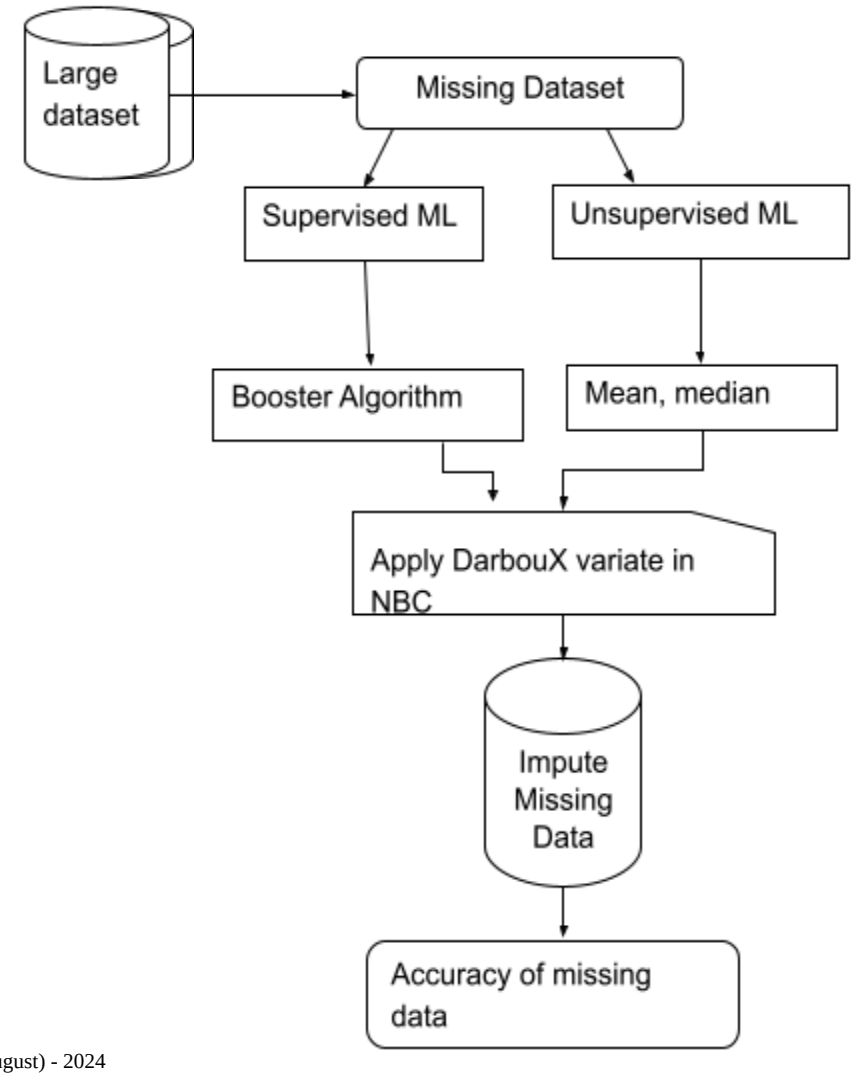
The variance estimation can be preferred to the linearized variance estimation, discussed in Riddles et al. (2016), due to complexity in computation. The Darboux theorem states that every defined group in R^n consist of a concurrent subgroup. For instance, a subgroup is a group that can be derived from another group by deleting any items without modifying the order of the resting items. Every bounded real sequence has a convergent subsequence. A subset of R is compact if and only if it is closed and bounded. The set S is rational and countable, and treat S as a bounded sequence from 0 to 1. Then it gives the following results for each statement. There is a convergent subsequence in S. Darboux theorem require an infinite construction, and it has no exception. The infinite construction is easier than the constructions in other proof. If (R_n) is a sequence of numbers in the closed segment [M, N], then it has a subsequence which converges to a point in [M, N].

Let's have an arbitrary point P, which is between the points M and N. Then observe the segment [M, P]. It may contain a finite number of members from the sequence (R_n) and it may contain an infinite number of them. If take the point P to be N, the segment [M, N] would contain an infinite

number of members from the sequence. If take the point P to be M, the segment [M, N] would contain at most only one point from the sequence. Let's introducing the set S= {P ϵ [M, N] | [M, P] contains a finite number of (R_n) members. M belongs to set S. If a point P belongs to S, it mean that [M, P] has a finite number of members from (R_n), and it will mean that any subset of [M, P] would also have only a finite number of members from (R_n). Therefore for any P that belongs to S, all the point between that P and M would also belongs to S.

The set S is actually a segment, starting at M and ending in some unknown location [M, N]. Now let's move to next step R= Sup(S) it means R is an accumulation point of (R_n). According to the special case R= M, and assume that R ϵ (M, N).Now we take an arbitrarily small ε. Observe the segment [M, R+ε]. R+ε cannot belong to S since it is higher than the supremum. Hence [M, R+ε] contains an infinite number of (R_n) members. Now the segment [M, R-ε]. R-ε must belong to S, since it is smaller than the supremum of the segment S. Thus [M, R-ε] contains a finite number of members from (R_n). But [M, R-ε] is a subset of [M, R+ε]. If the bigger set contains an infinite number of (R_n) members and its subset contains only a finite amount, the complement of the subset must contain an infinite number of members from (R_n). Proved that for every ε, the segment (R-ε, R+ε) contains an infinite number of members from the sequence. Construct a subsequence of (R_n) that converges to R. Take ε to be 1. Take any (R_n) member in (R-1, R+1) to be the first member.

Fig.1. Enhanced accuracy of Missing Data using DarbouX variates NBC.



This theorem proof that every bounded sequence of real numbers has a convergent subsequence, every bounded sequence in R^n has a convergent subsequence and every sequence in a closed and bounded set S in R^n has a convergent sub-sequence NBC technique is one of the widely used missing data treatment methods. The basic idea of NBC is first to define the attribute to be imputed, called imputation attribute and then, to construct NBC using imputation attribute as the class attribute. Other attribute in the dataset are used as the training subset. In addition to NBC the DarbouX variates is used to fix the infimum and supremum in the data sequence. Hence the imputation problem is becoming a problem of classified data sequence. Finally, the NBC along with the DarbouX variates is used to estimate and replace the missing data in imputation attribute.

Experimental Results

Experimental datasets were obtained from the University dataset of the UCI Repository. Table 1 describes the dataset, which features multivariate data characteristics with categorical integer attributes, containing 265 instances and 13 attributes. The primary objective of the experiments conducted in this work is to analyse the classification performance of machine learning algorithms. Datasets without missing values were used, and a few values were randomly removed at rates ranging from 5% to 25%. In these experiments, missing values were artificially introduced at varying rates across different attributes.

Data Set Characteristics:	**Multivariate**	**Number of Instances:**	**265**
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	13
Associated Tasks:	Classification	**Missing Values?**	Yes

Table 1 Dataset Used for Analysis

The following diagram represents the classification of missing value Imputation of original dataset using supervised machine learning techniques like Naïve Bayesan, Booster Algorithm, NBC-DarbouX variate and unsupervised machine learning techniques like Mean, Median and STD.

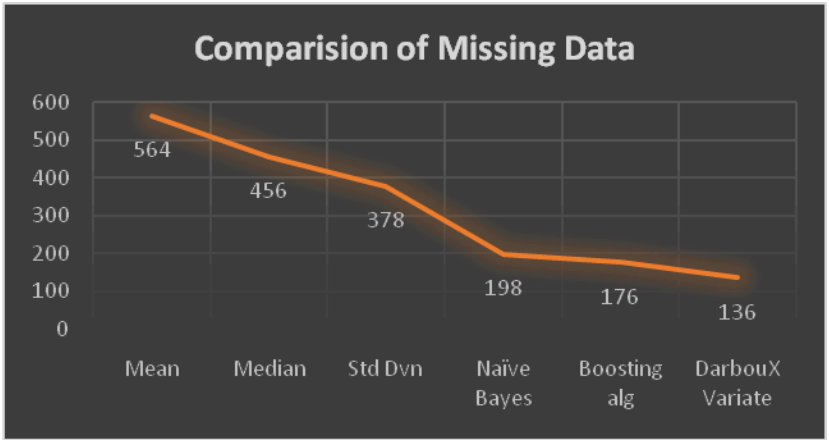


Fig.2. Missing Values imputation in Original Dataset

The figure below represents the percentage rates of missing values using both supervised and unsupervised techniques at rates of 5%, 10%, 15%, 20%, and 25%. It also compares the supervised techniques—NBC, Boosting Algorithm, NBC-DarbouX variate—with the unsupervised techniques—Mean, Median, and Standard Deviation—across different rates of missing values for all attributes.

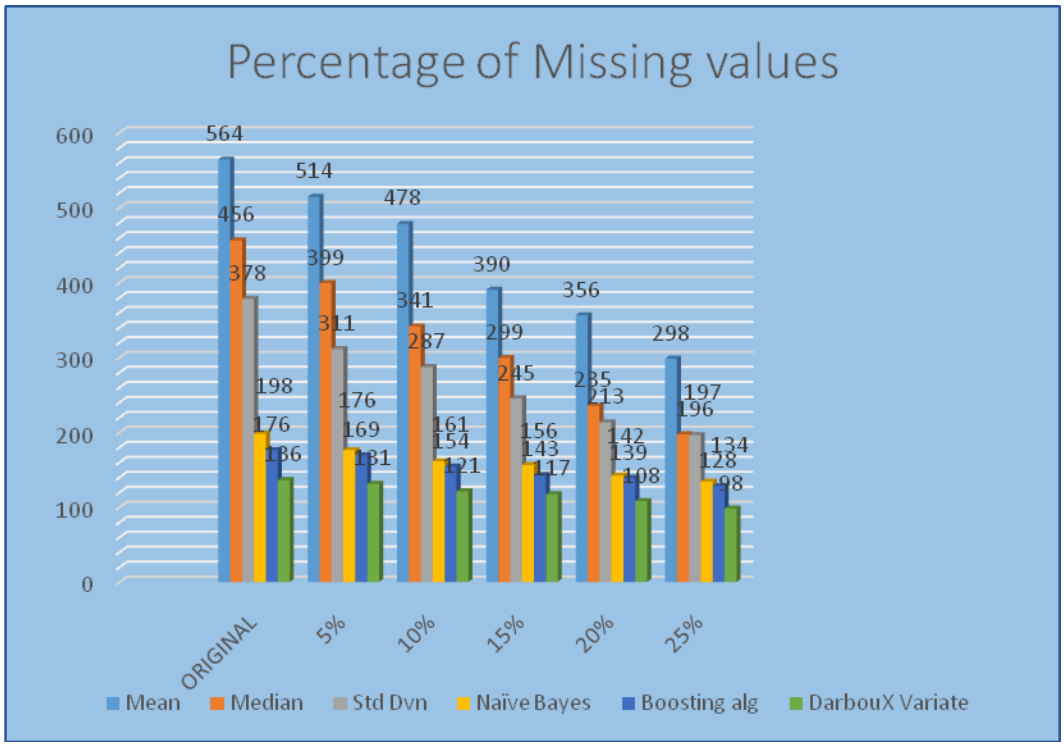


Fig 3. Percentage Rates of Missing Values

Conclusion

Nonignorable nonresponse models typically require strong assumptions for both the selection model approach and the pattern-mixture model approach. In this paper, we propose a multiple imputation method that specifies the respondents' outcome model and the response model with a nonresponse instrument variable. The proposed data augmentation algorithm combines fractional imputation and multiple imputation. Since the imputation step of the data augmentation algorithm is essentially equivalent to the expectation step of Kim (2011)'s parametric fractional imputation, we can easily impute missing values using probability sampling proportional to the fractional weights.

For a normal respondents' outcome model and a logistic response model, the main advantage of the proposed multiple imputation method is that the respondents' outcome model can be evaluated using a goodness-of-fit test or non-parametrically estimated. If the specified response model does not significantly deviate from the true response mechanism, the proposed multiple imputation estimators are relatively robust in terms of bias and coverage.

According to previous discussions, the DarbouX variate Naive Bayesian imputation classifier consists of two processes. Process 1 involves stating the imputation of elements and the imputation sequence. Process 2 involves applying the DarbouX variate – NBC to assign missing values. The Naive Bayesian classifier assigns the missing value in the first imputation element of the sequence and then updates the database for subsequent imputations. The DarbouX variate helps construct the classification model with infimum and supremum bounds; however, it cannot be systematically improved and does not automatically select suitable features like a boosted tree. The performance of the DarbouX variate depends on the correctness of the element selection in the database.

The main drawbacks of the Bayes classifier are its strong feature independence assumption and the issue of zero probability estimates when there are no occurrences of a class label and a certain element value together. According to the conditional independence assumption, multiplying all probabilities will yield zero, affecting the posterior probability estimate. This drawback is addressed by applying DarbouX variates in NBC to fix the infimum and supremum of the data sequence.

Considering the model specification of the respondents' outcome model and the response model, the proposed semi-parametric multiple imputation estimator is preferable in real data analysis. Additionally, the proposed method works well even for ignorable missing data, with only a minor cost in efficiency.

References

1. Box, G. E. P., & Tiao, G. G. (1992). Bayesian Inference in Statistical Analysis. Wiley.

2. Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputations under missing at random: A weighting approach. Statistical Methods in Medical Research.
3. Chang, T., & Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. Biometrika, 95, 555–571.
4. Durrant, G. B., & Skinner, C. J. (2006). Using data augmentation to correct for non-ignorable non-response when surrogate data are available: An application to the distribution of hourly pay. Journal of the Royal Statistical Society: Series A, 169, 605–623.
5. Galimard, J.-E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. Statistics in Medicine, 35, 2907–2920.
6. Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398–409.
7. Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 41, 337–348.
8. Giusti, C., & Little, R. J. A. (2011). An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. Journal of Official Statistics, 27, 211–229.
9. Greenlees, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. Journal of the American Statistical Association, 77, 251–261.
10. Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica, 47, 153–161.
11. Im, J., & Kim, S. (2017). Multiple Imputations for nonignorable missing data. Journal of the Korean Statistical Society, 46(4), 583–592.
12. Kim, J. K., & Yu, C. L. (2011). A semi-parametric estimation of mean functional with non-ignorable missing data. Journal of the American Statistical Association, 106, 157–165.
13. Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. Biometrika, 98, 119–132.
14. Kott, P. S., & Chang, T. (2010). Using calibration weighting to adjust for nonresponse and coverage errors. Journal of the American Statistical Association, 105, 1265–1275.
15. Little, R. J. A. (1995). Modeling the drop-out mechanism in longitudinal studies. Journal of the American Statistical Association, 90, 1112–1121.
16. Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis With Missing Data. Wiley.
17. Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. Journal of the American Statistical Association, 105, 1418–1430.
18. Qin, J., Leung, D & Shao, J. (2002). Estimation with survey data under non=ignorable nonresponse or informative sampling. Journal of the American Statistical Association, 97, 193–200.
19. Riddles, M. K., Kim, J. K & J. (2016). Propensity-score-adjustment method for non-ignorable nonresponse. Journal of Survey Statistics and Methodology, 215–245.
20. Rubin, D. B. (1977). Formalizing subjective notions about the effect of non-respondents in sample surveys. Journal of the American Statistical Association, 72, 538–543.
21. Rubin, D. B. (1987). Multiple Imputations for Nonresponse in Surveys. Wiley.