Uncertain Data Analysis with Regularized XGBoost

G.V. Suresh
Research Scholar, JNTU-Hyderabad, India.
E-mail: vijaysuresh.g@gmail.com

Dr. E. Sreenivasa Reddy
Professor, ANU-Guntur, India.
E-mail: edra92@gmail.com

Received September 20, 2021; Accepted December 17, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19245

Abstract

Uncertainty is a ubiquitous element in available knowledge about the real world. Data sampling error, obsolete sources, network latency, and transmission error are all factors that contribute to the uncertainty. These kinds of uncertainty have to be handled cautiously, or else the classification results could be unreliable or even erroneous. There are numerous methodologies developed to comprehend and control uncertainty in data. There are many faces for uncertainty i.e., inconsistency, imprecision, ambiguity, incompleteness, vagueness, unpredictability, noise, and unreliability. Missing information is inevitable in real-world data sets. While some conventional multiple imputation approaches are well studied and have shown empirical validity, they entail limitations in processing large datasets with complex data structures. In addition, these standard approaches tend to be computationally inefficient for medium and large datasets. In this paper, we propose a scalable multiple imputation frameworks based on XGBoost, bootstrapping and regularized method. XGBoost, one of the fastest implementations of gradient boosted trees, is able to automatically retain interactions and non-linear relations in a dataset while achieving high computational efficiency with the aid of bootstrapping and regularized methods. In the context of high-dimensional data, this methodology provides fewer biased estimates and reflects acceptable imputation variability than previous regression approaches. We validate our adaptive imputation approaches with standard methods on numerical and real data sets and shown promising results.

Keywords

Uncertainty, Missing Data, Multiple Imputation, Bootstrapping, Regularized Method, and XGBOOST.

Introduction

Real-world data sets frequently contain information derived from several sources, each of which uses a different measuring unit, data encoding, or data structure, resulting in a plethora of inaccuracies. In particular, they are frequently accompanied by several characteristics that have only been partially noticed. In the recent past, significant work has been made on the subject of "missing data," with the majority of the effort focusing on statistical inference. Multiple imputations (MI), first introduced by Rubin (1978) Liu Schafer JL (1997). Have received increasing recognition in dealing with missing data as it can reduce bias and represent the uncertainty of missing values. Instead of a single value being substituted for a missing value in an incomplete dataset, a list of probable values is substituted for each missing value *Sengupta S., Das A.K. (2012).* Afterwards, analysis can be carried out individually on entire imputed datasets, with the results being pooled to produce reliable inference Liu Schafer JL (1997). One main flaw of traditional MI implementations is that they fail to capture complex relations among variables automatically (*Gupta, M.K., Chandra (2020).* Another disadvantage of existing MI frameworks is the excessive computation time for large datasets *Liu Xingyi. Fillingetal. 2009) (K. Lavanya et al, 2012).* When there are reasonably large amounts of missing values present across many variables, this problem can become unmanageable. Applying Machine Learning (ML) techniques to multiple imputation can help to tackle the computational bottleneck, but the validity of imputed values obtained by current ML-based implementations is questionable *Liu Xingyi* et al, 2010). Additionally, measures like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used to verify XGBoost-based imputation models, which likewise try to restore the true data rather than representing uncertainty in the absence of missing values. Regularized regression has been proposed as a logical solution to the problem of developing imputation models in the presence of large amounts of high-dimensional data, and in this case it appears to be a logical solution to the problem of developing imputation models when dealing with large amounts of high-dimensional data. Regularized estimators come in a variety of shapes and sizes, with the Ridge Penalty being one of the most popular. Lasso Penalty, Adaptive Lasso, and the EasticNet (EN) Penalty Large amounts of data necessitated the use of regularized approaches and bootstrap combinations to fit the imputation model. This section describes theories and concepts. In Section 3, described about proposed MI with xgboost, Regularized Regression and Bootstrap. The Section 4 provided results of the proposed method over numerical and real data sets. In section 5, mentioned conclusion.

Related Theories and Concepts

A. Categories of Missing Values

Missing data is classified into two categories: ignorable and non-ignored. Ignore the case where the probability of missing data is determined by the data that is visible rather than the data that is missing. The amount of missing data, rather than the amount of observed data, determines the probability of missing data at non-ignorable. Let $X=(x_{ij}):(n\times k)$ rectangular data set without missing values. If $M=(m_{ij})$ then $m_{ij}=1$ if x_{ij} is missing and $m_{ij}=0$ if x_{ij} is present. Further in line to ignorable and non-ignorable missing data mechanisms classified into three types MCAR, MAR and NMAR.

- **Missing completely at random (MCAR):** Mechanism for missing data that is independent by data values X (Missing- X_{miss} - or Observed- X_{obs}).
- **Missing at random (MAR):** The mechanism of missingness is only dependent on X_{obs} , not on X_{miss} .
- **Not missing at random (NMAR):** The mechanism of missingness is dependent on X_{miss} .

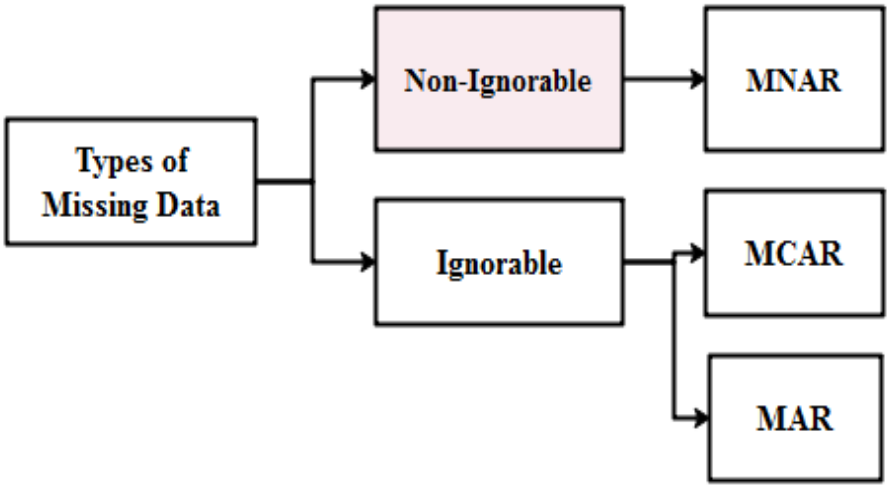

Fig. 1 Classification of Missing Data

Various techniques are available for treating missing data; a few techniques are described below[18]. Missing data treatment techniques can be classified into three classes, Traditional Approaches and Modern Approaches as shown in figure 2.

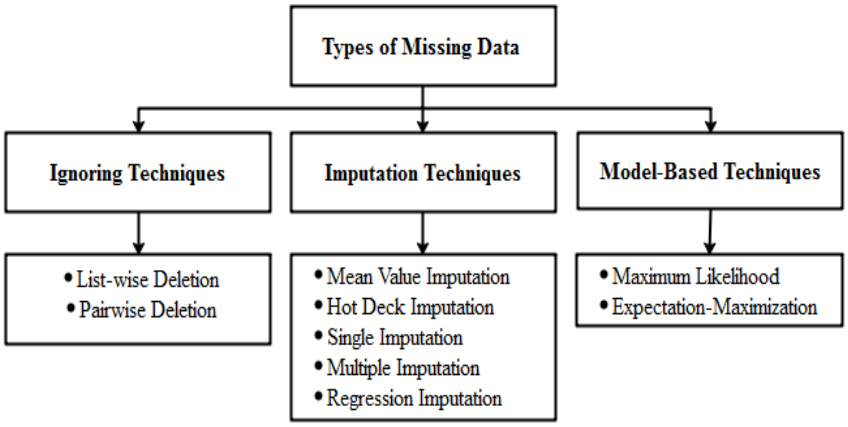

Fig. 2 Methods for handling Missing Data

In addition, the incidence of missing data can result in major difficulties for researchers. In fact, improper treatment of missing data during data analysis might result in bias being discovered and unclear conclusions being formed from a research study, as well as limiting the generalizability of the research findings. When there is a lack of data in DM, three sorts of difficulties are commonly encountered:

- Efficiency loss.
- Complexity in handling and analysing the data.
- Unfairness caused by discrepancies between incomplete and complete data.
- Efficiency loss.
- Complexity in handling and analysing the data.
- Unfairness resulting from differences between missing and complete data.

B. Multiple Imputation(MI)

Using Multiple Imputation (MI), we can create M imputed datasets and fit our complete data model to each of them, resulting in estimates $\hat{\theta}_m$ **m=1,2..,M** as well as corresponding within imputation variance estimates $\mathbf{Var}(\hat{\theta}_m)$. The estimation θ of is provided by.

$$\hat{\theta}_M = \mathbf{M}^{-1} \sum_{m=1}^{M} \hat{\theta}_m \qquad (1)$$

While the variance is estimated by,

$$\mathbf{Var}(\hat{\theta}_m) = \left(1 + \frac{1}{\mathbf{M}}\right) \delta^2_{btw} + \delta^2_{wtn} \qquad (2)$$

Where

$$\delta^2_{wtn} = \frac{1}{\mathbf{M}} \sum_{m=1}^{M} \mathbf{Var}(\hat{\theta}_m) \qquad (3)$$

$$\delta^2_{btw} = \frac{1}{M\text{-}1}\sum_{m=1}^{M} \mathbf{Var}(\hat{\theta}_m - \hat{\theta}_M)^2 \qquad (4)$$

The variance estimator is asymptotically unbiased for finite $\mathbf{M}$. Because of this, when the variance estimator, which is used to estimate the repeated sampling variance $\hat{\theta}_M$ of, is not properly calibrated, it can be biased either downward or upward, depending on the situation. When each dataset is imputed using the maximum likelihood estimate from a parametric imputation model and the imputations are analyzed using a non, semi, or fully parametric model. The between imputation covariance can be used to calculate variance estimates $\mathbf{Var}(\hat{\theta}_m) = \left(\frac{1}{M}\sum_m (\hat{\theta}_m - \hat{\theta}_M)(\hat{\theta}_m - \hat{\theta}_M)'\right)$ and the average within imputation covariance $\mathbf{W} = \sum_m \mathbf{Cov}(\hat{\theta}^*_b)$:

$$\mathbf{Cov}(\hat{\theta}_M) = \mathbf{W} + \frac{M+1}{M}\hat{V} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{Cov}(\hat{\theta}_m) + \frac{M+1}{M(M\text{-}1)}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_{MI})(\hat{\theta}_m - \hat{\theta}_{MI})' \qquad (5)$$

In the scalar case, a $\mathbf{t}_R$ -distribution with approximately can be assumed to $\mathbf{Var}(\hat{\theta}_M)^{-\frac{1}{2}}(\hat{\theta}_M - \theta)$ construct confidence intervals for $\hat{\theta}_M$. There exist degrees of freedom, even if there are other approximations $\mathbf{R} = \mathbf{(M\text{-}1)[1+\{MW/(M+1)\}]^2}$, especially for small samples.

Multiple Imputation with Bootstrapping

Let us consider a situation in which there is no analytic or ideal solution to estimate $\mathbf{Cov}(\hat{\theta}_m)$ the outcome. The following bootstrap percentile confidence intervals can be used to order a solution if there are no missing data points: based on $\mathbf{B}$ the results of bootstrap samples $\mathbf{D}^*_b$ for example $\mathbf{b} = 1, 2, .., \mathbf{B}$. We are able to obtain B point estimates $\hat{\theta}^*_b$. Consider the ordered set of estimates $\Theta^*_B = \{\hat{\theta}^*_b; \mathbf{b=1,2,...,B}\}$ where $\hat{\theta}^*_{(1)} < \hat{\theta}^*_{(2)} < \mathbf{...} < \hat{\theta}^*_{(B)}$ and the bootstrap confidence interval for θ is.

$$\left[\hat{\theta}^*_{(1ower)}\ ;\ \hat{\theta}^*_{(upper)}\right] = \left[\hat{\theta}^{*,\alpha}\ ;\ \hat{\theta}^{*,1-\alpha}\right] (6)$$

The α - percentile of the ordered bootstrap estimates Θ_B^* is represented by the symbol. For this data set $\mathbf{D} = \{\mathbf{D}^{obs}, \mathbf{D}^{mis}\}$, multiple imputation is used to fill in the gaps. There are two options for each of the $\mathbf{M}$ imputed data sets $\mathbf{D_m}$. Samples of Bootstrap code are drawn $\mathbf{M} \times \mathbf{B}$, resulting in data sets $\mathbf{D}^*_{\mathbf{m,b}}$; $\mathbf{b=1,2,...,B}$; $\mathbf{m=1,2,...,M}$ being produced. Each imputed data set is subjected to a set of bootstrap samples in order to estimate the standard error $\hat{\theta}^*_M$, in each imputed data set respectively, i.e., $\mathbf{Var}(\hat{\theta}_m)=\mathbf{(B-1)}^{-1}\sum_b(\hat{\theta}_{m,b}-\bar{\hat{\theta}}_{m,b})^2$ with $\bar{\hat{\theta}}_m=\mathbf{B}^{-1}\sum_b\hat{\theta}_{m,b}$. This gives $\mathbf{M}$ point estimates and $\mathbf{M}$ standard errors. In this approach, MI is used first, followed by bootstrapping on each imputed dataset. The following is the algorithm:

Algorithm: Multiple Imputation with Bootstrapping

1. Complete datasets are created by imputing missing values into the observed data and then completing the datasets. The analysis model should be fitted to each estimate.
2. Draw bootstrap samples for each imputed dataset.
3. For imputation $\mathbf{m}$,then calculate

$$\mathbf{Var}(\hat{\theta}_m)=\mathbf{(B-1)}^{-1}\sum_b(\hat{\theta}_{m,b}-\bar{\hat{\theta}}_{m,b})^2$$

where $\bar{\hat{\theta}}_m=\mathbf{B}^{-1}\sum_b\hat{\theta}_{m,b}$

4. After that, the Rubin's rule is applied, with $\hat{\theta}_m$ ($\mathbf{m=1,2,...,M}$) representing the point estimates and $\mathbf{Var}(\hat{\theta}_m)$ ($\mathbf{m=1,2,...,M}$) representing the complete data variance estimates.

Under congeniality variance estimates $\hat{\theta}_m$ are provided by MI boot Rubin, which is asymptotically unbiased. We can't anticipate this method to generate unbiased variance estimates in all circumstances of uncongeniality because it is based on Rubin's criteria.

XGBOOST

The XGBoost model, which primarily leverages the gradient boosting framework for optimal estimations, is one of the most effective algorithms in machine learning. XGBoost is a type of boosting technique that converts weak learners into stronger ones, thereby improving the random guessing process. Boosting is a prominent technology that involves developing trees using information from previously produced trees in a sequential process that follows one another. The approach also provides a high rate of regularization when boosting the model and effectively handles missing values. The gradient boosting decision

tree was used to create XGBoost, which was built on top of the CART tree. The following is the goal function:

$$Obj_n = \sum_{j=1}^{T}\left[\omega_j G_j + \frac{1}{2}\omega_j^2 (H_j + \lambda)\right] + \gamma T \qquad (7)$$

ω_j is the weight of leaf node j, $G_i = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$. g_i and h_i are respectively the first and second derivatives of $\hat{y}_i^{(t-1)}$ on the loss function $l(y_i^t, \hat{y}_i^{(t-1)}), I_j$ is the set of samples contained on the leaf with index j.The objective function takes the minimal value as well, resulting in the quadratic function having a derivative. A quadratic function is taken as a derivative w_j by each leaf j, and this derivative is zero for extreme values.

$$\omega_j^* = \frac{G_j}{(H_j + \lambda)} \qquad (8)$$

$$Obj_n = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{(H_j + \lambda)} + \gamma T \qquad (9)$$

When a leaf node is split, the gain after the split is calculated,

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{(H_L + \lambda)} + \frac{G_R^2}{(H_R + \lambda)} - \frac{(G_L + G_R)^2}{(H_L + H_R + \lambda)}\right] - \gamma \qquad (10)$$

Calculate the gain on each branch point of all features in order to discover the optimal tree structure, and then choose the node that yields the greatest gain, which is also the node with the fastest decline in the branching objective function, to discover the optimal tree structure. When the gain of this node falls below a certain threshold, the tree's growth is stopped.

Proposed Method: MI with Regularized Method

Consider the data set D_k which has missing values in k variables and other are fully observed denoted as $D_{obs,k} = (D_{1k}, D_{2k}, ..., D_{rk})^T$ and $D_{miss,k} = (D_{r+1k}, D_{r+2k}, ..., D_{nk})^T$ respectively. From the k variables first r components are missing values and other $(n-r)$ components are observed values. The complement data set $D_{k'} = (D_{obs}k', D_{miss}k')^T$. Then the observed

data are $\mathbf{D}_{obs,k} = (\mathbf{D}_{obs,k}, \mathbf{D}_{obs,k'}, ..., \mathbf{D}_{miss,k'})^T$ and the missing data are $\mathbf{D}_{miss}$; there are $\mathbf{r}$ complete cases and $(\mathbf{n}-\mathbf{r})$ n–r incomplete cases with $\mathbf{D}_k$ missing. The imputation model reduces to

$$f(\mathbf{D}_{miss,1} \mid \mathbf{D}_{obs,1}, \mathbf{D}_{-1}) = \int f(\mathbf{D}_{miss,1} \mid \mathbf{D}_{miss,1}, \theta) f(\theta \mid \mathbf{D}_{obs,1}, \mathbf{D}_{obs,-1}) \mathbf{d}\theta$$

To obtain the posterior distribution of θ, $\mathbf{f}(\mathbf{D}_{miss,1} \mid \mathbf{D}_{obs,1}, \mathbf{D}_{-1})$ we can posit and fit a regression model with $\mathbf{D}_{obs,k}$ as the outcome variable and $\mathbf{D}_{obs,-k}$ as the set of predictors. We will use a linear regression model as an example just for illustration.

$$\mathbf{D}_{obs,1} = \alpha_0 + \mathbf{D}_{obs,-1}\alpha + \in$$

Where $\in \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I}_r)$ **and** $\alpha = (\alpha_1, \alpha_2 ..., \alpha_{p-1})^T$

Specifically, we need to fit the imputation model using $\mathbf{r}$ complete cases. We denote by $\mathbf{S}$ the set of variables in $\mathbf{D}_{obs,-k}$ that are associated with $\mathbf{D}_{obs,k}$, also known as the true active set, and denote by $|\mathbf{S}| = \mathbf{q}$ its cardinality, that is, the number of important variables for imputing $\mathbf{D}_k$, and by $\mathbf{D}_{obs,S}$ the corresponding design matrix. Specifically, we define the active set as the subset of predictors that have been selected for impute $\mathbf{D}_1$, which is denoted by $\hat{\mathbf{S}}$, and the corresponding design matrix is denoted by $\mathbf{D}_{obs,\hat{S}}$. To achieve model trimming when fitting the imputation model, we propose combining regularization techniques such as Lasso or ALasso. Both model trimming and parameter estimation are accomplished using a regularization technique that incorporates a bootstrap step to simulate random draws from a distribution. $\mathbf{f}(\theta \mid \mathbf{D}_{obs,1}, \mathbf{D}_{obs,\hat{S}})$. . The algorithm for $\mathbf{m}^{th}$ imputation in the MIRXG approach is as follows:

Algorithm: Regularized Multiple Imputation with XGBoost (RMIXG)

Input : The Missing Dataset(D_{miss}) of size $n \times p$;

No.of.Imputations (M);

K = Sorted indices of Columns in D_{obs}
w.r.t. Missing Propotion.

for i=1 to M do

for j=1 to p do

var_list[j] $\leftarrow$ Regression
$(D_j^* \mid D_1^*, ..., D_{j-1}^*, D_{j+1}^*, ..., D_p^*)$

---

if L_1 then

Use LASSO Regression and Estimate

θ_L drawn from $f(\theta_L \mid D_{obs}, D_{obs,-1})$

$D_{imp} \leftarrow$ impute D_{mis} from $f(D_{mis} \mid D_{mis,-1}, \theta_L)$

if L_2 then

Use Ridge regression and estimate

θ_R drawn from $f(\theta_R \mid z_{obs}, Z_{obs,-1})$

$D_{imp} \leftarrow$ impute D_{mis} from $f(D_{mis} \mid D_{mis,-1}, \theta_R)$

else

Use Enet and estimate θ_E drawn

from $f(\theta_E \mid D_{obs}, D_{obs,-1})$.

$D_{imp} \leftarrow$ impute D_{mis} from $f(D_{mis} \mid D_{mis,-1}, \theta_E)$

end for

end for

$D_{imp} = (D_{imp}^{(1)}, D_{imp}^{(2)}, ..., D_{imp}^{(M)})$

$D^* \leftarrow$ bootstrap samples of D_{imp}

while not γ do

for K in k do

Separate the matrix D^* into : x, B, b

and A based on K^{th} column;

Train a XGBoost model : b ~ A;

Predict D^* by feeding B into the trained model;

D^*_{obs}new $\leftarrow$ update K^{th}column using predicted D^*;

end for

update γ;

end while

return the imputed matrix D^*_{obs}new

Simulations

The sample size in all simulations is set to n = 100., the outcome variable, and are the dependent variables $\mathbf{D=(d_1,d_2,...,d_p)}$ in each simulated data set. Total $\mathbf{d_k}$ number of variables contains missing data, k=2 with $\mathbf{p=200}$ and $\mathbf{p=1000}$. The data $\mathbf{(d_3,d_4,...,d_p)}$ generated using the distribution of type multivariate normal of considering mean= $\mathbf{(0,....,0)_{p-1}}$ and ρ is the autocorrelation of type $\mathbf{AR}$, varies from 0.2: 0.6. The final outcome variable y with respect to the D is described as:

$$\mathbf{Y=\beta_i=1+\beta_1 D_1+\beta_2 D_2+\beta_3 D_3+\beta_4 D_5+\varepsilon} \quad (11)$$

Where $\beta_i=1$, and random noise $\varepsilon \sim N(0.6)$.Also, it is observed that in study missing data with respect to the individual variable $\mathbf{d_1}$ is generated with combination of $\mathbf{p}$ and ρ using normal distribution with variance $\sigma^2_{d_1}=1$ and mean $\mu_{d_1}=\alpha_0+D_{S^\alpha}$. From the mean, the $\mathbf{S}$ denoted as the true active set (4, 20, 30, 40, 50, and 70). The resultant active set excluding with missing data is represented as $\mathbf{D_S=\{d_2,d_3,d_{50},d_{53}\}}$ $\mathbf{\{d_2,...,d_{11},d_{50},...,d_{59}\}}$. $\{d_2,...,d_{11},d_{50},...,d_{59},d_{70},...,d_{79},d_{90},...,d_{99},d_{110},...,d_{119}\}$ is shown respectively. The study consider the noise proportion =0.3 to specify missingness in the respective variables for example in $\mathbf{d_1}$. Missing values generated $\mathbf{d_k}$ from a logistic model approximately 30% of $\mathbf{d_1}$ missing.

Results

The work in this paper, compared proposed method (i.e., MI-REG-XGB) with the other two imputation techniques (i.e., MI+XGB and MI+REG). However, to test results used following measures bias, mean square error (MSE) and β coverage rate of the 95% confidence interval (CR). The simulation results are evaluated by fixing the coefficients values $\beta_1,\beta_2=1$. From the Tables 1–2 provides the results with respect to the correlation ρ with set of (0.2, 0.4 and 0.6). It is observed that from each data set with various imputation methods with auxiliary variable effect in terms of true active set (i.e., q= 20,30,40 or 70) is compared in the study. The data set is generated with two dimensions (i.e., p= 400 and 2000) and fixed size samples (i.e., n=300) with the variable correlation.

Table 1 Simulation results for estimating n= 100 andq =30, 20

ρ=0.2				
dimension Size	Methods	Bias	MSE	CR
p=400	MI	0.066	0.023	0.924
	MI-XGB	0.062	0.022	0.902
	MI-REG	0.050	0.017	0.948
	MI-REG-XGB	0.044	0.012	0.956
p=2000	MI	0.074	0.025	0.894
	MI+XGBOOST	0.066	0.024	0.862
	MI+REG	0.056	0.019	0.918
	MI+REG+ XGB	0.051	0.017	0.925
ρ=0.4				
Sample Size	Methods	Bias	MSE	CR
p=400	MI	0.068	0.017	0.946
	MI+XGBOOST	0.059	0.018	0.932
	MI+REG	0.053	0.014	0.956
	MI+REG+ XGB	0.042	0.013	0.960
p=2000	MI	0.083	0.021	0.927
	MI+XGBOOST	0.075	0.020	0.916
	MI+REG	0.069	0.016	0.932
	MI+REG+ XGB	0.058	0.012	0.944
ρ=0.6				
Sample Size	Methods	Bias	MSE	CR
p=400	MI	0.051	0.013	0.968
	MI+XGBOOST	0.056	0.014	0.946
	MI+REG	0.038	0.010	0.971
	MI+REG+ XGB	0.029	0.008	0.975
p=2000	MI	0.075	0.016	0.932
	MI+XGBOOST	0.062	0.018	0.916
	MI+REG	0.050	0.013	0.950
	MI+REG+ XGB	0.042	0.009	0.962

Table 2 Simulation results for estimating n= 100 and q = 70, 40

c=0.2				
Sample Size	Methods	Bias	MSE	CR
p=400	MI	0.017	0.014	0.924
	MI+XGBOOST	0.002	0.015	0.927
	MI+REG	0.008	0.012	0.934
	MI+REG+ XGB	0.002	0.007	0.942
p=2000	MI	0.030	0.017	0.894
	MI+XGBOOST	0.020	0.018	0.904
	MI+REG	0.006	0.013	0.912
	MI+REG+ XGB	0.004	0.012	0.934
c=0.4				
Sample Size	Methods	Bias	MSE	CR
p=400	MI	0.016	0.015	0.938
	MI+XGBOOST	0.014	0.014	0.930
	MI+REG	0.013	0.009	0.947
	MI+REG+ XGB	0.011	0.007	0.950
p=2000	MI	0.034	0.012	0.928
	MI+XGBOOST	0.027	0.011	0.918
	MI+REG	0.004	0.011	0.926
	MI+REG+ XGB	0.002	0.010	0.940
c=0.6				
Sample Size	Methods	Bias	MSE	CR
p=400	MI	0.013	0.010	0.944
	MI+XGBOOST	0.010	0.010	0.936
	MI+REG	0.007	0.008	0.948
	MI+REG+ XGB	0.002	0.006	0.956
p=2000	MI	0.019	0.013	0.936
	MI+XGBOOST	0.014	0.012	0.926
	MI+REG	0.011	0.007	0.934
	MI+REG+ XGB	0.008	0.004	0.947

Moreover, it is observed from the Table 1 and 2 produces least bias and close CR value results of proposed imputation methods with small true active set (i.e., q = 20 or 30) compared to the standard MI method. The performance of proposed MI+REG+XGB is promising compared to the MI+REG. However, the MI+XGBOOST method produces the mixed results in all possible cases. Also, it is observed that the MI + XGBOOST method produces smaller bias and MSE compared with the regularized MI. In all possible cases with effect of correlation c= 0.2, 0.4 or 0.6, proposed exhibits the better results in terms of bias and MSE.

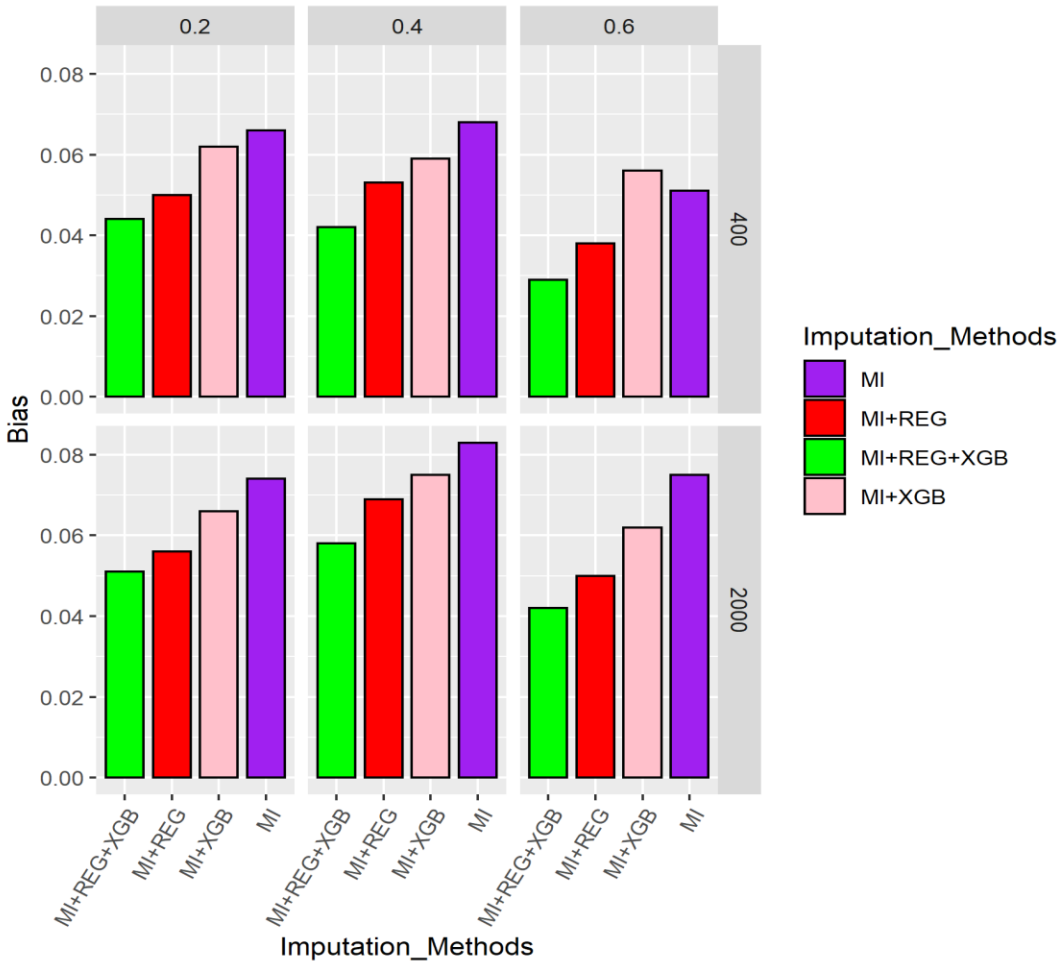

Fig. 3 Simulation results for Bias estimating n= 100 and q = 70, 40

From the results, it is observed that at high correlation cases, all regularized methods perform significantly good and produced negligible bias and small MSE by the proposed method. Tables 1 (q=70,40) and 2 (q=30,20) indicate similar patterns in comparisons amongst imputation approaches. MI+REG+ XGB has been found to outperform MI+REG The proposed method and MI with Regularized method exhibits promising results in terms of high correlation (i.e., c= 0.6) compared with medium and small correlation (i.e., c= 0.2and 0.4). However, in case of the normal MI method results were produced in mixed effect. The study, recommended that highly correlated variables fits imputation model better compared to the medium and small correlation. Also, proposed method summarizes that optimal selection of the true active set will retain better imputation results, compared to the performance of MI -REG and MI- XGB. Moreover, it is observed that, in case of data with different dimensions from p = 400 to 2000, existing method performance decreases respect to MSE, CRand bias.

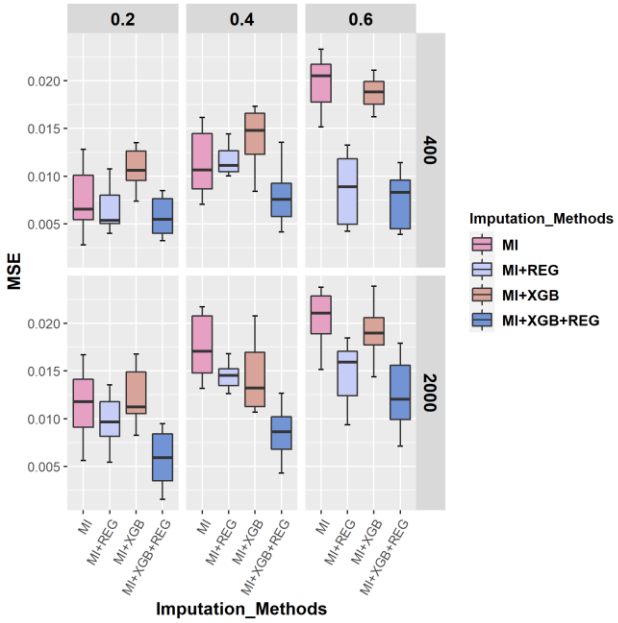

Fig. 3 Simulation results for MSE estimating n= 100 and q = 70, 40

However, proposed method deteriorates the performance very less compared to the standard imputation methods. From the Table 1 & 2, active set size (i.e., q=70, or 40), and active set size (i.e., q =30, or 20) exhibits related patterns on comparisons among the imputation methods.

The proposed method MI+REG+ XGB is produces the overall better performance when compared to the standard methods MI, MI+REG and MI+XGBOOST.

Also, it is shown that the case of q raises, imputation performance is with larger bias and MSE in case of MI+REG and MI+XGBOOST compared to the proposed method MI+REG+ XGB.

Real-World Datasets

The study also focused on the real datasets to compare the performance of imputation method. Among two data sets are high-dimensional and remain two data sets are small - dimension with reasonable number of the features <20. The data sets which includes:

- Delve Census
- Housing
- Mortgage
- Nitrogen
- Orange Juice

Delve Census: One of the high dimensional data sets is Delve census dataset with 104 features, and it is available from the University of Toronto1 with sample size of 2048. Next, one is Nitrogen dataset, includes total of features 141 with sample size of 1050various wavelengths. It is obtained from the Analytical Spectroscopy Research Group of the University of Kentucky.

Housing: The Housing dataset, with sample size of 506 and features of 13 and is available from the UCI Machine Learning repository. This data set work for prediction of the houses with desired number of the features.

Mortgage: The data set is Mortgage, composed of the number of features i.e., 16 and sample size 1049. The data set is available from the Federal Reserve Bank of Saint-Louis website.

Nitrogen: This data set all time work for nitrogen content prediction from a grass sample.

Orange Juice: The data set is Orange Juice, composed of the number of features i.e., 218 and sample size700. The data set fit for doing high dimensional data analysis.

Table 3 Real-World Datasets

Dataset Names	Sample Size	No. of. Features
Delve Census	2048	104
Housing	506	13
Mortgage	1049	16
Nitrogen	141	105
Orange Juice	218	700

Table 4 Methods Names and Description

ID	Methods Names	Methods Names
M1	MI	Multiple Imputation
M2	MI+XGB	Multiple Imputation with XGBoost
M3	MI+REG	Multiple Imputation with Regression
M4	MI+REG+XGB	Multiple Imputation with Regression and XGBoost

Table 5 MSE Performance of Imputation Methods on Real Datasets

Datasets	**True Active Set (q)=5**			
	M1	**M2**	**M3**	**M4**
Delve	2.3632	1.8987	1.8544	1.5656
Nitrogen	1.5478	0.9734	0.9866	0.7634
Housing	5.4314	4.5762	4.4816	3.6643
Mortgage	2.1533	1.3783	1.3487	0.9468
Orange Juice	4.0621	3.2864	3.1906	2.7273
Datasets	**True Active Set (q)=10**			
	M1	**M2**	**M3**	**M4**
Delve	3.7864	2.3683	1.9331	1.5820
Nitrogen	2.3751	1.0378	1.0109	0.8434
Housing	6.2495	4.5308	4.2262	4.0344
Mortgage	2.8293	1.5007	1.4647	1.0647
Orange Juice	4.5638	2.7481	2.6631	2.2519
Datasets	**True Active Set (q)=25**			
	M1	**M2**	**M3**	**M4**
Delve	4.1347	2.8151	2.5482	2.1309
Nitrogen	3.2765	2.1049	1.0573	0.9629
Housing	5.7952	4.4594	3.8974	2.5932
Mortgage	2.3213	1.5599	1.4385	1.2199
Orange Juice	4.2864	3.7163	4.7836	3.3207

Table 6 Coverage of 95% CI performance of Imputation Methods on Real Data sets

Datasets	**True Active Set (q)=5**			
	M1	**M2**	**M3**	**M4**
Delve	0.7256	0.886	0.904	0.922
Nitrogen	0.8695	0.905	0.924	0.942
Housing	0.8842	0.917	0.936	0.954
Mortgage	0.7392	0.860	0.878	0.895
Orange Juice	0.8523	0.912	0.936	0.951
Datasets	**True Active Set (q)=10**			
	M1	**M2**	**M3**	**M4**
Delve	0.8542	0.906	0.915	0.924
Nitrogen	0.8984	0.913	0.922	0.931
Housing	0.9031	0.928	0.937	0.946
Mortgage	0.8274	0.895	0.904	0.913
Orange Juice	0.8760	0.929	0.948	0.966
Datasets	**True Active Set (q)=25**			
	M1	**M2**	**M3**	**M4**
Delve	0.8472	0.899	0.912	0.926
Nitrogen	0.9214	0.933	0.947	0.961
Housing	0.8951	0.905	0.919	0.932
Mortgage	0.9102	0.934	0.948	0.962
Orange Juice	0.9237	0.946	0.967	0.985

from Table 4 and 5, shows results of real data sets, also proposed method shows low bias and reasonable CR comparable with regular methods the MI+REG and MI+XGBOOST methods. Proposed tends to produce greater performance in terms of bias and MSE when compared to all genuine active sets, and this is especially noticeable when the coefficients of c are 0, 0.2, 0.4, and 0.6. As long as the correlation between the variables is high, all regularized approaches perform reasonably well, with all of the methods proposed demonstrating low bias and tiny MSE. Regularized and suggested algorithms perform significantly better when c= 0.6 than when q= 10 or 25, but MI produces slightly low values. Also, it is observed that variables with highly correlated produces, adequate information for imputation, resulting in enhanced performance of MI REG and MI+REG+ XGB.

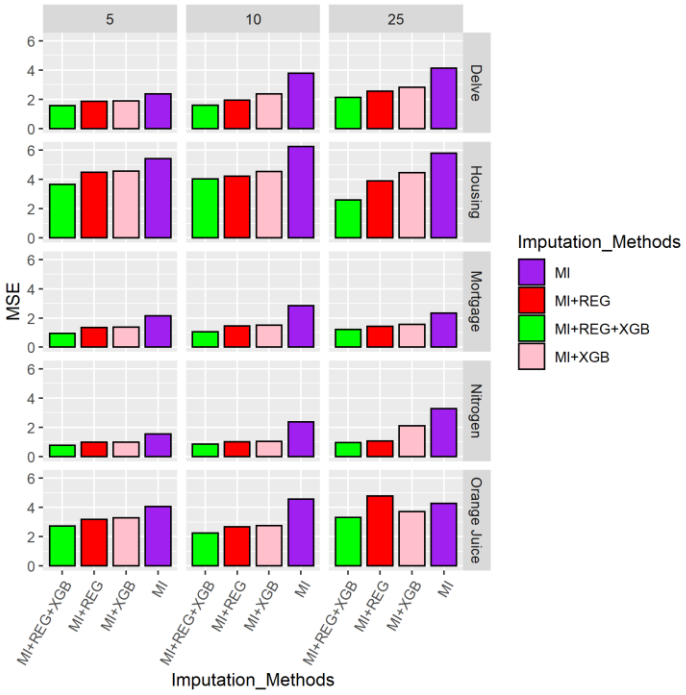

Fig. 4 Simulation results of MSE for real datasets MSE with True Active Set (q)=5,10,25

As **q** increases, each imputation method's performance degrades, resulting in increased biases and mean square errors (MSE), as illustrated in Figure. This degradation is significantly greater with MI+REG and MI+XGBOOST than it is with MI+REG+ XGB.

Conclusion

The work in this paper extends multiple imputations in terms of regularization and boosting for general missing data. The result of proposed method shows better performance with respect to measures of bias, RMSE and CR. In case of basic method MI+REG, shows large bias and MSE. However, the method MI+XGBOOST is

composition of MI with boosting raises performance with reasonable bias and smaller MSE. Moreover, it is observed that As long as the correlation between the variables is high, all regularized approaches perform reasonably well, with all of the methods proposed demonstrating low bias and tiny MSE. Also study focused on retrieving active subset for imputation strategy optimized true set leads to better results. The simulation and real study applied on all imputation strategies. Moreover, the proposed method dominates all existing methods with negligible bias and MSE.

References

Hariri, R.H., Fredericks, E.M., & Bowers, K.M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1), 1-16.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Gupta, M.K., & Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology*, *12*(4), 1243-1257.

Xingyi, L. (2009). Filling missing value algorithm based on Mahalanobis distance and gray analysis [J]. *Journal of Computer Applications*, *9*, 2502-2506.

Lavanya, K., Reddy, L.S.S., & Reddy, B.E. (2018). Modeling of missing data imputation using additive LASSO regression model in microsoft azure 2018. *J. Eng. App. Sci*, *13*, 6324-6334.

Schafer, J.L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica neerlandica*, *57*(1), 19-35.

Xingyi, L., Yao, T., & Chunhua, Z. (2010). Filling missing data method based on the Mahalanobis distance [J]. *Microcomputer Information*, *9*, 225-226.

De Silva, H., & Perera, A.S. (2016). Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data. *In Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 141-146.

Sengupta, S., & Das, A.K. (2012). Dimension reduction using clustering algorithm and rough set theory. *In International Conference on Swarm, Evolutionary, and Memetic Computing*, *Springer*, 705-712.

Pacheco, F., Cerrada, M., Sanchez, R.V., Cabrera, D., Li, C., & De Oliveira, J.V. (2016). Clustering algorithm using rough set theory for unsupervised feature selection. *In international joint conference on neural networks (ijcnn)*, 3493-3499.

Zhang, L., Lu, W., Liu, X., Pedrycz, W., & Zhong, C. (2016). Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values. *Knowledge-Based Systems*, *99*, 51-70.

Harel, O., & Zhou, X.H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, *26*(16), 3057-3077.

Patil, B.M., Joshi, R.C., & Toshniwal, D. (2010). Missing value imputation based on k-mean clustering with weighted distance. *In International Conference on Contemporary Computing*, Springer, 600-609.

Wiharto, W., & Suryani, E. (2020). The comparison of clustering algorithms k-means and fuzzy c-means for segmentation retinal blood vessels. *Acta Informatica Medica*, *28*(1), 42.

Huang, C.C., & Lee, H.M. (2004). A grey-based nearest neighbor approach for missing attribute value prediction. *Applied Intelligence*, *20*(3), 239-252.

Qin, Y., Zhang, S., Zhu, X., Zhang, J., & Zhang, C. (2009). POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert systems with applications*, *36*(2), 2794-2804.

Little, R.J., & Rubin, D.B. (2002). *Statistical analysis with missing data.* 2nd ed. New York: Wiley, 2002.

Rubin, D.B. (1987). *Multiple imputations for nonresponse in surveys.* New York: Wiley.