Supplementary material for "Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison"

Zhenhua Wang, Olanrewaju Akande, Jason Poulos and Fan Li *

Abstract

This supplementary material provides pseudocode for the MICE algorithm, and includes the data dictionary for the variables used in the simulation studies in the main text. It also includes the nonresponse models used to create the MAR scenario on the main text. Finally, it includes an evaluation based on five well-studied "benchmark" datasets.

*Zhenhua Wang is Research Associate in the Department of Statistical Science, Duke University, Durham, NC 27708 (E-mail: zhenhua.wang@duke.edu); Olanrewaju Akande is Assistant Professor of the Practice in the Social Science Research Institute, Box 90989, Duke University, Durham, NC 27708 (E-mail: olanrewaju.akande@duke.edu); Jason Poulos is Postdoctoral Associate in the Department of Health Care Policy, Harvard Medical School, Boston, MA (E-mail: poulos@hcp.med.harvard.edu); and Fan Li is Associate Professor of Statistical Science, Box 90251, Duke University, Durham, NC 27708 (E-mail: fl35@duke.edu).

1. MICE algorithm

Algorithm 1 MICE

input: incomplete data $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$
output: completed data $\mathbf{Y}^{(l)} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}^{(T)}_{\text{mis}})$
Specify order $(\mathbf{Y}_{(1)}, \ldots, \mathbf{Y}_{(p)})$ to iterate through sequence of conditional models
for $l = 1$ to L **do**
  initialize each $\mathbf{Y}_{\text{mis},(j)}$ by sampling from $\mathbf{Y}_{\text{obs},(j)}$
  for $t = 1$ to T **do**
    for $j = 1$ to p **do**
      fit conditional model
      $(\mathbf{Y}_{(j)}|\mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}^{(t)}_{(k)} : k < j\}, \{\mathbf{Y}^{(t-1)}_{(k)} : k > j\})$
      replace $\mathbf{Y}^{(t)}_{\text{mis},(j)}$ with draws from implied model
      $(\mathbf{Y}^{(t)}_{\text{mis},(j)}|\mathbf{Y}_{\text{obs},(j)}, \{\mathbf{Y}^{(t)}_{(k)} : k < j\}, \{\mathbf{Y}^{(t-1)}_{(k)} : k > j\})$
    end for
  end for
end for

2. Data dictionary for the ACS application

First, we present the data dictionary for the variables used in the simulation scenarios in Section 4 of the main paper. After pre-processing the 2018 American Community Survey (ACS) data as discussed in the main paper, we have 1,257,501 units, with 18 binary variables, 20 categorical variables with 3 to 9 levels, and 8 continuous variables. We treat this processed version as our final population from which we repeatedly sample from. Table 2.1 describes variables in this final population data.

Table 2.1. Variables from the 2018 ACS used to construct our population data.

Variable	Description	Type
ACR	Lot size	Ordinal (3 levels)
AGEP	Age	Numeric
BDSP	Number of bedrooms	Numeric
BLD	Units in structure	Nominal (4 levels)
COW	Class of worker	Nominal (4 levels)
DIS	Disability	Binary
HHL	Household language	Nominal (3 levels)
HHT	HH/family type	Nominal (4 levels)
HINS1	Insurance through a current or former employer or union	Binary
HINS2	Insurance purchased directly from an insurance company	Binary
HINS3	Medicare for people >= 65 years or with certain disabilities	Binary
HUGCL	Household with grandparent living with grandchildren	Binary
HUPAC	HH presence and age of children	Nominal (4 levels)
HUPAOC	HH presence and age of own children	Nominal (4 levels)
HUPARC	HH presence and age of related children	Nominal (4 levels)
JWMNP	Travel time to work	Numeric
JWRIP	Vehicle occupancy	Nominal (3 levels)
JWTR	Means of transportation to work	Nominal (3 levels)
LAPTOP	Laptop or desktop	Binary
LNGI	Limited English speaking household	Binary
MAR	Marital status	Nominal (4 levels)
MARHT	Number of times married	Ordinal (4 levels)
MULTG	Multigenerational household	Binary
MV	When moved into this house or apartment	Ordinal (7 levels)
NP	Number of persons in this household	Numeric
NR	Presence of nonrelative in household	Binary
PAOC	Presence and age of own children	Nominal (3 levels)
PARTNER	Unmarried partner household	Binary
PRIVCOV	Private health insurance coverage recode	Binary
PSF	Presence of subfamilies in household	Binary
PUBCOV	Public health coverage recode	Binary
PWGTP	Person's weight	Numeric
R18	Presence of persons under 18 years in household	Binary
R65	Presence of persons 65 years and over in household	Ordinal (3 levels)
RAC1P	Recoded detailed race code	Binary
RMSP	Number of rooms	Numeric
SCHL	Educational attainment	Ordinal (8 levels)
SEX	Sex	Binary
SRNT	Specified rental unit	Binary
SVAL	Specified owner unit	Binary
TEN	Tenure	Nominal (3 levels)
VEH	Vehicles (1 ton or less) available	Ordinal (4 levels)
WAGP	Wages or salary income past 12 months	Numeric
WIF	Workers in family during the past 12 months	Ordinal (4 levels)
WKHP	Usual hours worked per week past 12 months	Numeric
YBL	When structure first built	Ordinal (9 levels)

3. Nonresponse models for MAR scenario

Next, we describe the MAR scenario in the main paper in more detail. We begin by setting six variables — age, sex, marital status, race, educational attainment and class of worker — to be fully observed. We then randomly split the remaining 40 variables into three groups, consisting of 10, 15, and 15 variables. We create one logistic model per group, and condition on two of the fully observed six variables within each model. We use the logistic models to generate the item nonresponse indicators for each variable in that group.

Specifically, let A_i, S_i, M_i, R_i, E_i and C_i represent the age, sex, marital status, race, educational attainment and class of worker, of the $i = 1, \ldots, n$ individuals in the data. For each variable in the first group of 10 variables, we sample the item nonresponse indicator for each individual i using probability p_{i1}. For the second and third groups of 15 variables each, we use p_{i2} and p_{i3} respectively. We define p_{i1}, p_{i2}, and p_{i2}, as follows.

$$\begin{aligned} \text{logit}(p_{i1}) = &-1.3 + 0.4 \cdot \mathbb{1}[C_i = 0] - 1.3 \cdot \mathbb{1}[C_i = 1] \\ &+ 1.2 \cdot \mathbb{1}[C_i = 2] - 0.4 \cdot \mathbb{1}[C_i = 3] + 0.02 \cdot A_i; \end{aligned} \tag{3.1}$$

$$\begin{aligned} \text{logit}(p_{i2}) = &-1.2 - 1.5 \cdot \mathbb{1}[E_i \in \{0, 1\}] + 0.5 \cdot \mathbb{1}[E_i \in \{2, 3, 4\}] \\ &+ \mathbb{1}[E_i \in \{5, 6, 7\}] - 0.3 \cdot \mathbb{1}[M_i = 0] + 1.2 \cdot \mathbb{1}[M_i = 1] \\ &- 1.3 \cdot \mathbb{1}[M_i = 2] + 0.4 \cdot \mathbb{1}[M_i = 3]; \end{aligned} \tag{3.2}$$

$$\begin{aligned}\text{logit}(p_{i3}) = &-0.5 - 0.7 \cdot \mathbb{1}[S_i = 0] + 0.9 \cdot \mathbb{1}[S_i = 1] \\ &- 1.5 \cdot \mathbb{1}[R_i = 0] + 0.6 \cdot \mathbb{1}[R_i = 1].\end{aligned} \tag{3.3}$$

Here, $\mathbb{1}[\cdot] = 1$ when its argument is true and $\mathbb{1}[\cdot] = 0$ otherwise. Across all three groups, we set the corresponding entries for each variables as missing when the sampled nonresponse indicator is equal to one. This process results in approximately 30% missing rate for each of the 40 variables.

4. Evaluation based on "benchmark" datasets

Finally, we check the reproducibility and verify the claims of the deep learning methods, by comparing the methods based on our metrics using five "benchmark" datasets commonly used in the machine learning literature. The benchmark datasets are from the UCI Machine Learning Repository (Dua & Graff 2017), and are frequently used in the machine learning literature for evaluating missing data imputation methods. These datasets, which we describe in Table 4.1, vary vastly in size and structure. For example, the Breast Cancer dataset has only 569 sample units and no categorical variables; the Spam Detection dataset contains 4,601 observations but only continuous variables, whereas the Letter Recognition dataset contains 20,000 observations but only categorical variables. These extremes are very different from what one would expect from a survey data, for example, the ACS data used in our main evaluations.

Table 4.1. Characteristics of selected UCI benchmark datasets.

Data	# Samples	# Cont. variables	# Cat. variables	Avg. Corr*
Breast Cancer	569	30	0	0.394
Credit Card Default	30,000	14	9	0.163
Letter Recognition	20,000	0	16	0.182
Spam Detection	4,601	57	0	0.060
News Shares	39,644	44	14	0.068

* is the average absolute correlations among variables, as reported in Yoon, Zame, and van der Schaar (2018).

For each dataset, we follow Yoon, Zame, and van der Schaar (2018) and create an MCAR scenario by randomly setting 20% of the values of each variable to be missing independently. We use MICE-CART, GAIN, and MIDA to create $L = 10$ completed datasets for each benchmark dataset, and compute the metrics on the completed datasets. We again omit MICE-RF because the results in the main text already showed MICE-RF to be consistently inferior to MICE-CART in terms of performance and computation. We only generate 10 samples to remain consistent with the other scenarios in the main paper.

Also, for consistency, we prefer to evaluate all five datasets in the same manner. However, the sample sizes are not large enough to consider them population data from which we can repeatedly sample from without replacement. For example, the Breast Cancer dataset only contains 569 observations and the Spam Detection dataset only contains 4,601. Thus, we are unable to evaluate them in a meaningful way using absolute normalized bias, relative MSE or coverage. We therefore primarily evaluate them in a similar manner to the n=100,000 and 30% MCAR scenario in the main text. That is, we evaluate them using the weighted absolute bias metric.

We do report estimates of overall RMSE on the continuous variables and overall accuracy on the categorical variables, to once again provide evidence on how examining only the overall RMSE and accuracy metrics may be misleading for evaluating imputation methods. Here, we repeatedly create the same 20% MCAR process on each dataset, 10 times, and estimate the overall RMSE and accuracy on all 10 copies of the data. We report the average overall RMSE and accuracy across the 10 copies, and report the standard deviation of the values as the corresponding standard errors. While we once again prefer to draw repeated samples from a population as a way to properly account for the sampling mechanism, we follow this approach of repeatedly creating missing values of a single dataset to replicate the results from Yoon et al. (2018) for these two metrics on all the datasets, as much as possible.

Table 4.2 displays the median values of the estimated weighted absolute bias of the marginal probabilities of the categorical and binned continuous variables. MICE-CART significantly outperforms the other two methods. Specifically, MICE-CART results in the smallest weighted absolute bias in both categorical and continuous variables, across all the datasets. The difference is more pronounced with continuous variables, particularly with the Spam Detection dataset, which has 57 continuous variables, the highest number of all the datasets. This result is consistent with our findings in the main paper.

Table 4.3 displays the overall RMSE on continuous variables and overall accuracy on categorical variables for the datasets. MICE-CART achieves the highest

Table 4.2. Median values of the weighted absolute bias (×100), of the marginal probabilities of the categorical and binned continuous variables.

Data	MICE-CART	GAIN	MIDA
	Categorical		
Credit Card Default	0.07	1.76	0.56
News Shares	0.09	4.24	1.14
Letter Recognition	0.08	2.31	0.39
	Binned Continuous		
Breast Cancer	0.50	1.30	0.96
Credit Card Default	0.06	3.19	1.79
News Shares	0.07	3.69	3.16
Spam Detection	0.18	14.86	15.08

overall accuracy in the Credit Card Default and News Shares datasets, while MIDA achieves the highest accuracy in the Letter Recognition dataset. There is no consistent pattern in comparing all three methods using overall RMSE. MICE-CART achieves the lowest overall RMSE in the News Shares dataset, MIDA achieves the lowest overall RMSE in the Credit Card Default dataset, GAIN achieves the lowest overall RMSE in the Breast Cancer dataset, and both GAIN and MIDA achieve similar RMSE, lower than MICE-CART, in the Spam Detection dataset. MIDA is capable of preserving feature correlations (Gondara & Wang 2018) and performs comparatively well against the other methods on the datasets with highly correlated features. These patterns clearly differ from those reported earlier based on marginal and bivariate probabilities and weighted absolute bias. Therefore, as discussed in the main paper, we again warn against using the overall RMSE and accuracy as the only metrics for comparing imputation methods.

Finally, we note that we are unable to perfectly replicate the results from Yoon

Table 4.3. Overall RMSE on continuous variables and overall accuracy on categorical variables for each dataset, with estimated standard errors.

Data	MICE-CART	GAIN	MIDA
	RMSE		
Breast Cancer	0.107 ± 0.002	0.078 ± 0.003	0.100 ± 0.003
Credit Card Default	0.073 ± 0.003	0.075 ± 0.006	0.067 ± 0.003
News Shares	0.121 ± 0.021	0.181 ± 0.014	0.173 ± 0.015
Spam Detection	0.064 ± 0.001	0.054 ± 0.001	0.055 ± 0.001
	Accuracy		
Credit Card Default	0.740 ± 0.001	0.642 ± 0.031	0.678 ± 0.002
News Shares	0.930 ± 0.001	0.796 ± 0.053	0.876 ± 0.001
Letter Recognition	0.395 ± 0.001	0.332 ± 0.015	0.443 ± 0.001

et al. (2018) on all the datasets. First, for datasets containing both categorical and continuous variables, that is the Credit Card Default and News Shares datasets, the authors do not report a separate overall RMSE for the continuous variables as we do here. The authors also primarily report overall RMSE for the Letter Recognition dataset, while we report overall accuracy since the dataset only contains categorical variables. We do replicate the results for the Spam Detection dataset, and while we are unable to perfectly replicate the results for the Breast Cancer dataset, the overall trends in Yoon et al. (2018) are consistent with our results here. We note that the Spam Detection and Breast Cancer datasets contain only continuous variables. Second, the publicly available version of the author's code for GAIN is more updated than what was used in the Yoon et al. (2018). We also did not find any publicly available version of the author's code for MIDA used in Lu, Perrone, and Unpingco (2020); we simply reproduced the architecture. Thus, another possible reason for the discrepancies is the differences in parameter tuning. As we mentioned in the main

paper, this once again highlights how much the performance of machine learning methods can be highly dependent on parameter tuning.

References

Dua, D., & Graff, C., (2017), UCI Machine Learning Repository, http://archive.ics.uci.edu/ml

Gondara, L., & Wang, K., (2018), MIDA: Multiple imputation using denoising autoencoders, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 260–272.

Lu, H., Perrone, G., & Unpingco, J., (2020), Multiple imputation with denoising autoencoder using metamorphic truth and imputation feedback.

Yoon, J., Zame, W. R., & van der Schaar, M., (2018), Estimating missing data in temporal data streams using multi-directional recurrent neural networks, *IEEE Transactions on Biomedical Engineering*, *66*(5), 1477–1490.