Automatic Delta-Adjustment Method applied to Missing Not At Random Imputation

Ricardo Cardoso Pereira[1], Pedro Pereira Rodrigues[2], Mário A. T. Figueiredo[3,4], and Pedro Henriques Abreu[1]

[1] Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
{rdpereira,pha}@dei.uc.pt
[2] Center for Health Technology and Services Research, Faculty of Medicine (MEDCIDS), University of Porto, 4200-319 Porto, Portugal
pprodrigues@med.up.pt
[3] Instituto Superior Técnico, University of Lisbon, 1049-001 Lisbon, Portugal
[4] Instituto de Telecomunicações, 1049-001 Lisbon, Portugal
mario.figueiredo@tecnico.ulisboa.pt

Abstract. Missing data can be described by the absence of values in a dataset, which can be a critical issue in domains such as healthcare. A common solution for this problem is imputation, where the missing values are replaced by estimations. Most imputation methods are suitable for the Missing Completely At Random (MCAR) and Missing At Random (MAR) mechanisms but produce biased results for Missing Not At Random (MNAR) values. An effective approach to mitigate this bias effect is to use the delta-adjustment method. This method assumes the imputation is performed for the MAR mechanism and adjusts the imputed values to become valid under MNAR assumptions by applying a correction factor. Such adjustment is usually defined manually by a domain expert, which often makes this method unfeasible. In this work, we propose an automatic procedure to find an approximate delta adjustment value for every feature of the dataset, which we call Automatic Delta-Adjustment Method. The proposed procedure is validated in an experimental setup comprising 10 datasets of the healthcare domain injected with MNAR values. The results from seven state-of-the-art imputation methods are compared with and without the adjustment, and applying the correction provides a significantly lower imputation error for all methods.

Keywords: Missing Data · Imputation · Delta-Adjustment · Missing Not At Random.

1 Introduction

Missing data is a common problem in real-world data. It can be described by the absence of values in one or more features of a dataset. However, the missing values can assume different characteristics that are directly related with the missingness causes. There are three different mechanisms that categorize these causes [19, 6]:

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

- Missing Completely At Random (MCAR), which describes missing values that are the result of a purely random event, meaning the missingness causes are unrelated to any data. An example would be if someone randomly misses a question in a form;
- Missing At Random (MAR), which states that the missing values are related to part of the observed data (e.g., a specific range of a feature). For example, in a medical report, female people would have the results for prostate-related exams missing;
- Missing Not At Random (MNAR), which refers to missing values that are related to themselves or to other unobserved data, and, therefore, the missingness causes are unknown. For example, in a life insurance questionnaire, people that smoke a lot may not want to disclose how many cigarettes they smoke per day.

The existence of missing values impacts negatively any task performed with the data, such as predictive and statistical inference. Therefore, this issue is usually handled in a pre-processing stage to avoid being propagated. The most common approach to address missing data is imputation, which provides plausible estimates to replace the missing values. Such estimations can be performed through simple statistical strategies or even complex machine learning models [11]. However, in general terms, the imputation approaches tend to only be suitable for the MCAR and MAR mechanisms. Since the models base their estimations on the available data, the imputation tends to be biased when performed under MNAR assumptions because the existent data is not enough to properly model its missingness causes [10]. Nevertheless, there are a few approaches to reduce this bias, one being the delta-adjustment method [20]. Assuming that the imputation is performed with a model designed for the MAR mechanism, the missing values are likely to be either underestimated or overestimated. In other words, they are probably shifted towards a lower or higher domain. The delta-adjustment method provides a simple and transparent solution to correct this shift: add or multiply the imputed values by a correction factor. However, this factor must be manually defined by domain experts, which makes the delta-adjustment method often unfeasible to be applied considering the complexity and cost of consulting these experts. Moreover, the lack of scientific rigor in this process may also compromise the generalization of the results.

In this work we introduce an automatic procedure to estimate the approximate delta adjustment values for every feature of the dataset, which we call Automatic Delta-Adjustment Method (ADAM). The procedure explores the distance between the biased imputation and other estimations that are sampled from Gaussian distributions with extreme means that stretch the range of considered values. We conducted an experimental setup with seven state-of-the-art imputation methods, comprising 10 datasets of the healthcare domain that were injected with missing values under MNAR. We compared the results of the imputation with and without the adjustments provided by ADAM trough the Mean Absolute Error (MAE), and validated the results trough the Wilcoxon signed-rank

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

test with a significance level of 5%. The use of ADAM's adjustment provided significant improvements for all imputation methods and missing rates.

The remainder of the paper is organized in the following way: Section 2 presents the related work regarding how to address the MNAR mechanism; Section 3 describes in detail the proposed ADAM method; Section 4 presents the design of the experimental setup; Section 5 displays the analysis of the obtained results; and Section 6 states the final conclusions and future directions.

2 Related Work

Any imputation method will likely produce biased results for the MNAR mechanism since the missing values depend on unknown data. Incidentally, most features suffering from MNAR values have their distribution shifted when only the available data is considered. This shift will be propagated to the imputation model, which leads to the biased estimations. The impact of such bias can be measured by performing sensitivity analysis, where the parameters of the imputation models are varied in order to understand the magnitude of the bias and if the estimations can be admitted as valid [2]. Such strategy is not a solution for the problem, but allows for a more informed decision about whether the imputation models are usable or not. Nevertheless, performing a correct sensitivity analysis often requires previous knowledge about the features' distribution and domain, information that is usually obtained from domain experts. Consulting such experts is a complex and expensive task, which leads to this type of study often being unfeasible.

For these reasons, the more common way to address MNAR is to use imputation methods suitable for MCAR and MAR, and apply procedures that make the estimations more resilient to the MNAR assumptions. A common strategy for this purpose is multiple imputation, where the missing values are imputed multiple times while varying specific parameters (usually in an automated way) to mitigate the uncertainty. The multiple estimations are then analyzed and aggregated into a final result. A state-of-the-art method which is based on this strategy is the Multiple Imputation by Chained Equations (MICE) [26]. Another useful approach to achieve estimations valid under MNAR is the delta-adjustment method. As previously described, this method provides a simple and understandable way to correct the distribution shift, which is to add or multiply the imputed values by a correction factor [20]. This approach, although very effective, relies once again on domain experts since they need to define the correction factors manually for each feature. Consequently, the method is often unfeasible to be used for the same reasons reported for the sensitivity analysis. More recently, neural network-based models have also been successfully used to perform imputation under MNAR, in particular denoising and variational autoencoders [16]. Such models are very resilient to noise and can accommodate the MNAR characteristics better than other imputation methods [7, 8, 15]. Nevertheless, there are imputation models designed for MCAR and MAR that are also often included in experimental baselines of MNAR studies, particularly

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

the k-Nearest Neighbors (kNN), matrix completion methods, and the common mean/mode imputation [3, 17]. However, such methods tend to produce poor results considering the existent bias.

Focusing on the delta-adjustment method, the following works address imputation under MNAR assumptions using it. Carreras et al. [5] conducted a sensitivity analysis study to understand the impact of assuming and treating missing data as MAR or MNAR in end-of-life care studies. The MICE method was used for both mechanisms, but for MNAR it was integrated with the delta-adjustment method. Four different adjustment values were considered, which were defined as the equispaced values between zero and half of the interquartile range of the features. The experiments were conducted with data from the ACTION study, a randomized controlled trial testing advance care planning in patients with advanced lung or colorectal cancer. The authors concluded that the imputation assuming MAR reflected that the missing values were related to poorer health conditions. These correlations changed when the MNAR mechanism was assumed, which shows that the obtained conclusions are sensible to the violation of the MAR assumptions.

Leacy et al. [9] performed a similar sensitivity analysis study to understand how the departure from MAR to MNAR influenced the tasks of estimating the prevalence of a partially observed outcome and performing parametric causal mediation analyses with a partially observed mediator. The study used data from a tuberculosis (TB) and human immunodeficiency virus (HIV) prevalence survey that was conducted as part of the Zambia–South Africa TB and AIDS Reduction Study, between 2006 and 2010. The shift from MAR to MNAR was once again performed with the delta-adjustment method integrated in a multiple imputation procedure. Three adjustment values were manually chosen based on the experts opinion and on data from 3 consecutive annual rounds of HIV counseling and testing in the Karonga District of Malawi (2007 to 2010). Each of these values represented different magnitudes of departure from the MAR assumptions. The authors concluded that the estimation of the overall HIV prevalence was considerably different when assuming the MAR or MNAR mechanisms, particularly for strong departures between them.

Rezvan et al. [18] also conducted a sensitivity analysis study where the missing values imputed under MAR with multiple imputation were shift to MNAR with the delta-adjustment method. The data used in the experiments was from the Longitudinal Study of Australian Children, and the goal was to estimate the association between exposure to maternal emotional distress at the age of four/five years and total difficulties at the age of eight/nine years. The adjustment values were defined with the help of domain experts through an elicitation process that allows for the formulation of the expert's feedback into a probability distribution. The authors concluded that there are significant increases in the magnitude of the association between maternal distress and total difficulties when the MNAR assumptions are assumed with a large departure from MAR.

Tan et al. [23] proposed a review study about the use of controlled multiple imputation in randomized controlled trials where missing data exists. The anal-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

ysis considered the trials in phases II, III and IV published in The Lancet and New England Journal of Medicine between January 2014 and December 2019, covering primary and sensitivity analysis studies. The findings show that 56% of the controlled multiple imputation was performed with the delta-adjustment method. Nevertheless, most of the works report the used delta values but do not provide justifications to why the experts decided towards those values.

In conclusion, the delta-adjustment method has been widely used to shift missing values under MAR assumptions to the MNAR mechanism, reducing greatly the departure between the two. Its use has been particularly relevant in the healthcare domain, where the MNAR mechanism appears abundantly. However, the definition of the delta values is always performed by domain experts, which is the problem that we are addressing here. To the best of our knowledge, the automatic estimation of the approximate delta values is being introduced for the first time in our work.

3 Automatic Delta-Adjustment Method

In this work an automatic procedure that is capable of estimating approximate delta adjustment values is proposed. Considering the MNAR characteristics, it is impossible to find an optimal delta value since it would depend on the missing values themselves. Therefore, we rely on statistics to find an approximate value that will bring the estimates made under MAR assumptions valid under MNAR. We called our approach Automatic Delta-Adjustment Method (ADAM).

ADAM's goal is to find a factor comprised within $[0, 1]$ for each feature of the dataset. This factor will be multiplied by the imputed values in order to adjust them. The proposed procedure to estimate the factor for a specific feature X_i is presented in Algorithm 1. The procedure assumes that the missing values correspond to the smaller values of the feature (the opposite scenario is addressed later), and comprises these main steps:

1. The mean of the available values within the first quartile is calculated (μ_{Q1}). This value is representative of the lower tail of the feature, and it is used instead of the minimum because it better represents the group of smaller values and it is resilient to extreme factors. Additionally, the standard deviation of all the available values in the feature is also calculated (σ_{all});
2. The missing values are imputed three independent times by sampling from a Gaussian distribution where the standard deviation is one and the mean varies μ_{Q1} minus σ_{all} according to the empirical rule: $\mu_{Q1} - \sigma_{all}$, $\mu_{Q1} - 2\sigma_{all}$, and $\mu_{Q1} - 3\sigma_{all}$. Such variation is used to define a reasonable range for the missing values, since the imputed values with MAR models are likely to be shifted towards a higher domain. Moreover, when later calculating the adjustment factor, these three imputations are weighted so that the more extreme values (i.e., further away from μ_{Q1}) have a decreased impact on the calculation.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

Algorithm 1 Pseudocode for the ADAM procedure. The missing values are on the lower tail of the feature X_i. The algorithm receives as input the feature data containing missing values ($X_i\_missing$) and already imputed ($X_i\_imputed$), and it returns the adjusted data for the feature ($X_i\_adjusted$).

Input: $X_i\_missing$, $X_i\_imputed$
Output: $X_i\_adjusted$
μ_{Q1} = mean($Q1$ values from $X_i\_missing$)
σ_{all} = standard_deviation($X_i\_missing$)
for each j in $\{1, 2, 3\}$ **do**
    $gaussian\_imputed_j$ = missing data $\sim \mathcal{N}(\mu_{Q1} - j * \sigma_{all}, 1)$
end for
$scalar_0$ = PCA (1 comp.) applied to $X_i\_imputed$
for each j in $\{1, 2, 3\}$ **do**
    $scalar_j$ = PCA (1 comp.) applied to $gaussian\_imputed_j$
    $dist_j$ = euclidean_distance($scalar_0$, $scalar_j$)
end for
Normalize $dist_j, \forall j \in \{1, 2, 3\}$ so that $\sum_1^3 dist_j = 1$
$factor_i = (\, 3dist_3 + 2dist_2 + dist_1 \,) \,/\, 6$
$X_i\_adjusted = X_i\_imputed - factor_i * X_i\_imputed$
return $X_i\_adjusted$

3. The Principal Component Analysis (PCA)[5] method is used to condense the data from the features into a single representative numeric value. This transformation is applied individually to the dataset imputed with the MAR model and the datasets imputed in the previous step, which leads to four different scalars by feature. To achieve these results the datasets must be transposed before the transformation since we are condensing the data from the features and not the features themselves. The obtained values for each feature are then used to estimate how far away are the imputations from the previous step when compared to the original imputation with the MAR model. For this purpose, we calculate the euclidean distance between the value representing the imputation with the MAR model and each one of the remaining scalars;
4. To calculate the factor, the distances from the previous step are normalized so that their sum is equal to one (which is necessary to achieve a factor within $[0, 1]$), and a weighted mean is calculated so that the imputations performed according to the empirical rule have an impact in the factor proportional to their means.

After performing the previous step, we have an independent factor ($factor_i$) within $[0, 1]$ for each feature (X_i) of the dataset. To adjust the values imputed

[5] PCA is a feature extraction technique often used for dimensionality reduction. It computes the principal components of the data and returns the first n, which is a user-defined parameter [1].

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

with the MAR model, the following equation is applied to each of the D features:

$$X_i = X_i - factor_i * X_i, \forall i \in \{1, ..., D\} \quad (1)$$

As previously stated, the described procedure assumes that the missing values are the smaller values of the feature. However, the opposite scenario where the missing data corresponds to the larger values is also valid. To address this scenario, the procedure suffers two specific changes:

- In step 1), the last quartile is instead calculated (μ_{Q4}) since it now represents the higher tail of the feature.
- In step 2), the mean is varied on the opposite direction ($\mu_{Q4}+\sigma_{all}$, $\mu_{Q4}+2\sigma_{all}$, and $\mu_{Q4}+3\sigma_{all}$), since the imputed values with MAR models are now shifted towards a lower domain.

Finally, the adjustment operation also changes since the imputed values are now being shifted upwards:

$$X_i = X_i + factor_i * X_i, \forall i \in \{1, ..., D\} \quad (2)$$

The identification of the feature's missing tail (i.e., if the missing values correspond to the smaller or larger values of the feature) should be made manually through exploratory data analysis or even by using domain knowledge.

4 Experimental Setup

To evaluate if ADAM is effective in adjusting the imputed values, an experiment was conducted to compare the imputation results before and after applying it. The experimental setup comprised the following seven state-of-the-art imputation methods:

- Mean imputation, where the mean of the available values of each feature are used to impute;
- Multiple Imputation by Chained Equations (MICE), which is a multiple imputation-based approach where several Bayesian ridge regressions are fitted in a round-robin procedure with 100 iterations. Each regression defines as the dependent variable one of the features containing missing values, and uses the remaining as the independent variables [4];
- k-Nearest Neighbors (kNN) imputation with $k = 5$, which selects the k nearest neighbors of the instance being imputed by calculating the Euclidean distance between the available values, and uses the mean of these k neighbors to impute the features with missing data [21];
- SoftImpute, which is a matrix completion iterative method based on nuclear-norm regularization, that estimates the missing values through soft-threshold singular value decomposition [12]. A maximum of 100 iterations was considered;

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

- Denoising Autoencoder (DAE), which is an autoencoder trained with data containing additional noise, which in this context is the existence of missing values. Incidentally, an autoencoder is a special type of neural network that tries to reproduce the input data at the output layer, usually by learning a compressed representation of that data [25, 16]. The architecture of the networks was defined with the following hyperparameters: a single hidden layer with a number units equal to half of the input dimension; ReLU as the activation function; batches of 64 instances; a maximum of 200 epochs; Adam as the optimization algorithm; Mean Squared Error as the loss function; a learning rate of 0.001; a dropout layer with a rate of 25% for regularization; early stop if the validation loss has no improvements over 100 epochs; a reduction of the learning rate by 80% if there are no improvements over 100 epochs; and Sigmoid as the activation function for the output layer, so that the data is normalized within $[0, 1]$;
- Variational Autoencoder (VAE), which is a generative variant of the autoenconder that learns the multidimensional parameters of a Gaussian distribution (i.e., mean and standard deviation), and by sampling from these is able to generate new data with similar characteristics [13, 15, 16]. The architecture of the networks was defined with the same hyperparameters as the DAE, adding the two layers needed to represent the Gaussian parameters.
- Generative Adversarial Imputation Nets (GAIN), which is a direct application of the well-known generative adversarial networks to the problem of missing data [28]. The generator network performs the imputation, while the discriminator tries to distinguish between original and imputed data. The networks were parametrized with the hyperparameters reported by the authors of the method.

The architecture and hyperparameters used for the seven imputation methods were defined through a grid search procedure. This is a common strategy that aims to obtain optimal hyperparameters that conform to common use cases. Regarding the implementation of these algorithms, the autoencoders were implemented with the Keras library, the SoftImpute was coded from scratch, the GAIN code was obtained from the original authors[6], and the remaining methods were directly used from the Scikit-learn library. Furthermore, the implementation of ADAM is available on GitHub[7].

The experiment considered 10 public datasets from the healthcare domain, covering clinical research based on routinely collected data for different pathologies. This domain was chosen since it frequently suffers from missing data under the MNAR mechanism [14]. All datasets are available at the UC Irvine repository[8], and they cover different ranges of instances and features, as Table 1 shows.

In order to have a controlled experiment, the datasets were all complete (i.e., without missing data) and the missing values were artificially generated

[6] `https://github.com/jsyoon0823/GAIN`
[7] `https://github.com/ricardodcpereira/ADAM`
[8] `https://archive.ics.uci.edu/ml`

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

Table 1. Datasets used in the experimental setup.

Dataset	*# Instances*	*# Features*	
		Continuous	*Categorical*
wisconsin	569	31	0
ctg	2126	21	2
pima	768	9	0
liver	583	10	1
diabetic-retinopathy	1151	16	4
parkinsons	195	23	0
bc-coimbra	116	10	0
thoracic-surgery	470	14	3
spine	310	13	0
mammographic-masses	830	2	4

according to the MNAR mechanism. The used strategy removed the smaller values [24] or larger values [27] of the orderable features upon a certain missing rate (which excludes nominal categorical features since they are non-orderable). Such strategy was applied in a multivariate fashion where the missing rate is defined for the entire dataset and several features are injected with missing values simultaneously [22]. Consequently, each feature has a different number of missing values, which grouped together sum up to the desired global missing rate. The imputation is performed for all features at once, and the results are assessed through the Mean Absolute Error (MAE) calculated between the ground truth (i.e., original data) and the imputed values.

All datasets were normalized within $[0, 1]$ and split into train and test sets with 70% and 30% of the instances, respectively. The normalizer uses the minimum and maximum values from the train set and it is then applied to both sets. This strategy keeps the test data isolated from the training data, preventing bias in the test set. However, when dealing with high missing rates (usually above 50%), it is possible for the test set to contain unseen values, which makes its normalization boundaries go slightly beyond the aimed $[0, 1]$ domain. For the neural network-based methods, 20% of the train set was used for validation. The non-orderable features (e.g., categorical nominal) were transformed through one-hot encoding (i.e., dummy coding). The missing values were injected independently in each of the described sets in order to ensure that all of them had equal missing rates and MNAR assumptions. Five different missing rates were considered (5%, 10%, 20%, 40%, and 60%) in order to cover different levels of missingness. For the neural network-based methods the missing values were pre-imputed with the mean imputation.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

To mitigate bias and stochastic behaviors the experiment was executed 30 independent times, with the data being randomly split into the train and test sets in each run. The results here presented are the mean of these 30 runs. Each run was executed in a computer with the following specifications: Windows 11, CPU AMD Ryzen 5600X, 16GB RAM, and GPU NVIDIA GeForce GTX 1060 6GB. The time complexity of applying ADAM for each of the imputation methods was not directly measured, but the impact was not significant.

5 Results

The obtained results for the adjustment of imputed values provided by ADAM are presented in Tables 2 and 3. For Table 2 the left tail was missing (i.e., smaller values of the features), and for Table 3 the right tail was removed (i.e., larger values of the features).

In an overall analysis, all methods achieved smaller imputation errors after the imputed values were adjusted through ADAM. This behavior is consistent for all missing rates and for both MNAR strategies with the smaller and larger values being removed, with the global error improvement being 11%. The enhancement also appears to be stable among the different levels of missingness, peaking at 18% for the 5% missing rate, which shows ADAM's resilience to variations on this factor.

To understand if the obtained results were statistically significant we applied the Wilcoxon signed-rank test with a significance level of 5%. This test was chosen because the normality assumptions were not met, and we have paired MAE values for each imputation method (before and after applying ADAM). The test was applied independently for each missing rate, and the one-sided alternative was used since we were only interested in evaluating if the MAE values after applying ADAM are significantly lower. The obtained p-values showed that the results are statistically significant with $p < 0.001$ for all settings, which corroborates the good performance obtained by ADAM.

6 Conclusions

In this work we proposed a procedure called Automatic Delta-Adjustment Method (ADAM) to automatically estimate the delta-adjustment values. We compared the results obtained by seven state-of-the-art imputation methods with and without the adjustments provided by ADAM. The experimental setup comprised 10 datasets from the healthcare context that were injected with missing values under MNAR in a multivariate fashion. We concluded that the adjustment performed by ADAM led to error improvements in all imputations methods and missing rates, achieving a global enhancement of 11%.

Motivated by the results achieved with ADAM, future work will be focused on integrating it with an auxiliary procedure to automatically identify the features' missing tails. A possible direction is to model this task as a binary classification

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

Table 2. Mean Absolute Error results of the imputation methods with and without the adjustment provided by ADAM. The left tail of the features was missing (i.e., smaller values).

Imp.	*ADAM*	*Missing Rate*				
		5%	*10%*	*20%*	*40%*	*60%*
AE	No	0.236 ± 0.09	0.238 ± 0.09	0.256 ± 0.10	0.322 ± 0.15	0.515 ± 0.34
	Yes	**0.192 ± 0.07**	**0.195 ± 0.06**	**0.212 ± 0.07**	**0.274 ± 0.12**	**0.464 ± 0.31**
GAIN	No	0.254 ± 0.08	0.252 ± 0.08	0.266 ± 0.10	0.324 ± 0.15	0.519 ± 0.34
	Yes	**0.206 ± 0.06**	**0.208 ± 0.06**	**0.220 ± 0.07**	**0.277 ± 0.12**	**0.469 ± 0.31**
MICE	No	0.184 ± 0.09	0.186 ± 0.09	0.204 ± 0.10	0.280 ± 0.14	0.475 ± 0.34
	Yes	**0.154 ± 0.07**	**0.157 ± 0.07**	**0.174 ± 0.07**	**0.245 ± 0.12**	**0.436 ± 0.31**
Mean	No	0.269 ± 0.08	0.264 ± 0.08	0.278 ± 0.09	0.334 ± 0.14	0.523 ± 0.33
	Yes	**0.217 ± 0.05**	**0.216 ± 0.06**	**0.229 ± 0.07**	**0.284 ± 0.11**	**0.472 ± 0.30**
SoftImp	No	0.178 ± 0.07	0.184 ± 0.07	0.197 ± 0.08	0.255 ± 0.11	0.429 ± 0.29
	Yes	**0.152 ± 0.06**	**0.160 ± 0.06**	**0.174 ± 0.06**	**0.233 ± 0.10**	**0.412 ± 0.28**
VAE	No	0.285 ± 0.11	0.279 ± 0.10	0.294 ± 0.12	0.348 ± 0.16	0.533 ± 0.35
	Yes	**0.227 ± 0.07**	**0.224 ± 0.07**	**0.238 ± 0.08**	**0.292 ± 0.12**	**0.477 ± 0.32**
kNN	No	0.185 ± 0.09	0.194 ± 0.09	0.219 ± 0.10	0.292 ± 0.14	0.487 ± 0.34
	Yes	**0.153 ± 0.07**	**0.162 ± 0.07**	**0.184 ± 0.07**	**0.254 ± 0.11**	**0.446 ± 0.31**

problem and use machine learning to solve it. Furthermore, we want to incorporate information from other datasets in the adjustments calculations, so that external data can be used to help reduce bias in MNAR settings. Finally, we also want to compare ADAM results to imputed data that was manually adjusted by domain experts.

Acknowledgements This work is supported in part by the FCT - Foundation for Science and Technology, I.P., Research Grant SFRH/BD/149018/2019. This work is also funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics **2**(4), 433–459 (2010)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

Table 3. Mean Absolute Error results of the imputation methods with and without the adjustment provided by ADAM. The right tail of the features was missing (i.e., larger values).

Imp.	*ADAM*	*Missing Rate*				
		5%	*10%*	*20%*	*40%*	*60%*
AE	No	0.710 ± 0.27	0.847 ± 0.52	1.417 ± 1.91	2.770 ± 3.37	4.011 ± 4.81
	Yes	**0.610 ± 0.26**	**0.744 ± 0.51**	**1.306 ± 1.91**	**2.655 ± 3.36**	**3.894 ± 4.80**
GAIN	No	0.725 ± 0.27	0.849 ± 0.52	1.410 ± 1.91	2.769 ± 3.37	4.036 ± 4.81
	Yes	**0.628 ± 0.26**	**0.747 ± 0.51**	**1.301 ± 1.91**	**2.654 ± 3.36**	**3.924 ± 4.80**
MICE	No	0.579 ± 0.23	0.716 ± 0.47	1.291 ± 1.87	2.678 ± 3.35	3.943 ± 4.80
	Yes	**0.496 ± 0.22**	**0.624 ± 0.46**	**1.183 ± 1.85**	**2.557 ± 3.34**	**3.813 ± 4.79**
Mean	No	0.774 ± 0.28	0.896 ± 0.53	1.448 ± 1.92	2.784 ± 3.36	4.019 ± 4.81
	Yes	**0.689 ± 0.27**	**0.804 ± 0.52**	**1.346 ± 1.91**	**2.671 ± 3.35**	**3.902 ± 4.80**
SoftImp	No	0.676 ± 0.27	0.826 ± 0.51	1.422 ± 1.90	2.834 ± 3.39	4.187 ± 4.84
	Yes	**0.579 ± 0.26**	**0.727 ± 0.50**	**1.319 ± 1.90**	**2.738 ± 3.38**	**4.119 ± 4.84**
VAE	No	0.758 ± 0.29	0.884 ± 0.54	1.441 ± 1.92	2.781 ± 3.37	4.018 ± 4.81
	Yes	**0.671 ± 0.28**	**0.790 ± 0.53**	**1.338 ± 1.92**	**2.669 ± 3.36**	**3.903 ± 4.80**
kNN	No	0.618 ± 0.24	0.761 ± 0.49	1.342 ± 1.89	2.722 ± 3.36	3.976 ± 4.81
	Yes	**0.509 ± 0.23**	**0.650 ± 0.48**	**1.224 ± 1.88**	**2.600 ± 3.35**	**3.852 ± 4.80**

2. Austin, P.C., White, I.R., Lee, D.S., van Buuren, S.: Missing data in clinical research: A tutorial on multiple imputation. Canadian Journal of Cardiology (2020)
3. Beaulieu-Jones, B.K., Lavage, D.R., Snyder, J.W., Moore, J.H., Pendergrass, S.A., Bauer, C.R.: Characterizing and managing missing structured data in electronic health records: data analysis. JMIR Medical Informatics **6**(1), e11 (2018)
4. Buuren, S.v., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of Statistical Software pp. 1–68 (2010)
5. Carreras, G., Miccinesi, G., Wilcock, A., Preston, N., Nieboer, D., Deliens, L., Groenvold, M., Lunder, U., van der Heide, A., Baccini, M.: Missing not at random in end of life care studies: multiple imputation and sensitivity analysis on data from the action study. BMC Medical Research Methodology **21**(1), 1–12 (2021)
6. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Computing and Applications **19**(2), 263–282 (2010)
7. Gondara, L., Wang, K.: Recovering loss to followup information using denoising autoencoders. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 1936–1945 (2017)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

8. Gondara, L., Wang, K.: Mida: Multiple imputation using denoising autoencoders. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 260–272 (2018)
9. Leacy, F.P., Floyd, S., Yates, T.A., White, I.R.: Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: application to a tuberculosis/hiv prevalence survey with incomplete hiv-status data. American Journal of Epidemiology **185**(4), 304–315 (2017)
10. Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R., Carpenter, J.R.: Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. PharmacoEconomics **36**(8), 889–901 (2018)
11. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, vol. 793. John Wiley & Sons (2019)
12. Mazumder, R., Hastie, T., Tibshirani, R.: Spectral regularization algorithms for learning large incomplete matrices. The Journal of Machine Learning Research **11**, 2287–2322 (2010)
13. McCoy, J.T., Kroon, S., Auret, L.: Variational autoencoders for missing data imputation with application to a simulated milling circuit. IFAC-PapersOnLine **51**(21), 141–146 (2018)
14. Peek, N., Rodrigues, P.P.: Three controversies in health data science. International Journal of Data Science and Analytics **6**(3), 261–269 (2018)
15. Pereira, R.C., Abreu, P.H., Rodrigues, P.P.: Partial multiple imputation with variational autoencoders: Tackling not at randomness in healthcare data. IEEE Journal of Biomedical and Health Informatics (2022)
16. Pereira, R.C., Santos, M.S., Rodrigues, P.P., Abreu, P.H.: Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. Journal of Artificial Intelligence Research **69**, 1255–1285 (2020)
17. Qiu, Y.L., Zheng, H., Gevaert, O.: Genomic data imputation with variational autoencoders. GigaScience **9**(8) (2020)
18. Rezvan, P.H., Lee, K.J., Simpson, J.A.: Sensitivity analysis within multiple imputation framework using delta-adjustment: Application to longitudinal study of australian children. Longitudinal and Life Course Studies **9**(3), 259–278 (2018)
19. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
20. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys, vol. 81. John Wiley & Sons (2004)
21. Santos, M.S., Abreu, P.H., García-Laencina, P.J., Simão, A., Carvalho, A.: A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of Biomedical Informatics **58**, 49–59 (2015)
22. Santos, M.S., Pereira, R.C., Costa, A.F., Soares, J.P., Santos, J., Abreu, P.H.: Generating synthetic missing data: A review by missing mechanism. IEEE Access **7**, 11651–11667 (2019)
23. Tan, P.T., Cro, S., Van Vogt, E., Szigeti, M., Cornelius, V.R.: A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data. BMC Medical Research Methodology **21**(1), 1–17 (2021)
24. Twala, B.: An empirical comparison of techniques for handling incomplete data using decision trees. Applied Artificial Intelligence **23**(5), 373–405 (2009)
25. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine learning. pp. 1096–1103 (2008)
26. White, I.R., Royston, P., Wood, A.M.: Multiple imputation using chained equations: issues and guidance for practice. Statistics in Medicine **30**(4), 377–399 (2011)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34

27. Xia, J., Zhang, S., Cai, G., Li, L., Pan, Q., Yan, J., Ning, G.: Adjusted weight voting algorithm for random forests in handling missing values. Pattern Recognition **69**, 52–60 (2017)
28. Yoon, J., Jordon, J., Schaar, M.: Gain: Missing data imputation using generative adversarial nets. In: International Conference on Machine Learning. pp. 5689–5698. PMLR (2018)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-35995-8_34