MULTIVARIATE IMPUTATION BY MAHALANOBIS DISTANCE OPTIMIZATION (MIMDO)

Geovert John D. Labita[1], Altea S. Labita[2], Bernadette F. Tubo[3]

•

[1,2]University of Science and Technology of Southern Philippines
[3]Mindanao State University - Iligan Institute of Technology
[1]geovertjohn.labita@ustp.edu.ph, [2]altea.labita@ustp.edu.ph, [3]bernadette.tubo@g.msuiit.edu.ph

Abstract

This paper introduces a new method for missing data imputation based on an optimization approach and is now available as an R package called "mimdo". This method deals with imputing the missing values by computing the values that minimize the Mahalanobis distance between an observation and the overall mean. The effectivity of mimdo was demonstrated in both classification and regression tasks using popular benchmark datasets. From all experiments, it was found out that using mimdo for imputing the missing values in the dataset, on the average, the classification rate is more than 80% and an R-squared of more than 50%. Furthermore, the consistency of the results were validated through simulation studies.

Keywords: missing data, cluster analysis, regression analysis, optimization, mahalanobis

1. Introduction

Real-world datasets always are accompanied by missing data which is one major factor affecting data quality [9]. As a result, missing data pose a challenge to most data analysis techniques because a substantial proportion of the data may be missing and predictions must be made for cases with missing inputs. Thus, missing data must be handled properly in order to obtain quality knowledge and also because many statistical models and machine learning algorithms rely on complete datasets.

The risk of missing data is evident in many applications. For example, in cluster analysis, missing values can complicate the application of clustering algorithms. Specifically, in astronomy, imaging sensors have limited sensitivity and may fail to detect light intensities below a minimal threshold frustrating the clustering of celestial bodies [10]. Similarly, according to [2], meteorological or weather stations are riddled with missing climatic data. Thus, the regional grouping of these climatic elements are greatly affected.

In regression analysis, as the amount of data that is missing increases, there can be a substantial reduction of sample size and a resulting loss of power which means that there is a potential for biases in the regression estimates and their standard errors. For example, the problem of non-response to one or more questions in budgetary studies may be very troublesome when the data are to be used in regression analysis [6].

A common practice for dealing with missing values in the context of clustering is to first impute the missing values, and then apply the clustering algorithm on the completed data [5]. However, some of the data imputation methods, in some way, have their drawbacks. For example, the basic data imputation technique like the mean imputation is fundamentally changing the structure of the underlying data. As a consequence, the final clusters obtained are to some extent, a consequence of the decision to replace the missing values by the means rather than the data itself

[4]. Moreover, highly correlated variables is also an issue in data imputation with multivariate imputation by chained equations (mice) as an example. In most cases, mice algorithm will leave these variables out of the imputation process [7].

Inspired from the work of [3], this paper introduces a new methodology for data imputation, which imputes the missing data by computing the values that will minimize the Mahalanobis distance between an observation and the overall mean. By utilizing the Mahalanobis distance, this novel method is also suitable for datasets with high correlation since Mahalanobis distance can handle non-spherical datasets which are commonly caused from highly correlated variables.

This paper is arranged as follows. Methodology is introduced and discussed in section 2. The model solution is presented and derived in section 3 while section 4 illustrates the some applications of *mimdo*. Section 5 presents the concluding remarks.

2. Methods

This section presents the derivation of the optimization model with schematic diagram and *R* installation of *mimdo*. It begins by providing a brief introduction to optimization before proceeding to give a detailed discussion of the methodology.

Definition 1. [1] Let $\boldsymbol{v} \in \mathcal{X} \subset \mathbb{R}^d$ be a vector of decision variables v_i, $i = 1, 2, \ldots, d$ where $\mathcal{X}$ is the ground set of vectors of decision variables. Also, let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$ be the objective function and $g_j : \mathbb{R}^d \to \mathbb{R}$ be the constraint function defining restriction on $\boldsymbol{v}$. Then the optimization problem is to

$$\begin{array}{rll} \min & f(\boldsymbol{v}) & \\ \text{subject to} & g_j(\boldsymbol{v}) \geq 0, & j \in \mathcal{I} \text{ (inequality constraints)} \\ & g_j(\boldsymbol{v}) = 0, & j \in \mathcal{E} \text{ (equality constraints).} \end{array} \tag{1}$$

For a maximization problem, the sign of f is changed. In this paper, f can be interpreted as the distance between two observations. In an unconstrained optimization, $\mathcal{I} \cup \mathcal{E} = \emptyset$ and $\mathcal{X} = \mathbb{R}^d$. The set of vectors $S \subset \mathbb{R}^d$ satisfying the constraints in model (1) is called the set of *feasible solutions* to the problem.

To derive the desired optimization model, we first specify the missing and known values for a given dataset $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ with p variables, that is,

$$\mathcal{M} = \left\{(i,q) : x_{iq} \text{ is missing, } 1 \leq q \leq p\right\},$$

$$\mathcal{N} = \left\{(i,q) : x_{iq} \text{ is known, } 1 \leq q \leq p\right\}.$$

Also, let J be the set of indices of all incomplete observations given by

$$J = \{i \ : \boldsymbol{x}_i \text{ has at least 1 missing coordinate}\}.$$

Now, let $\mathbf{W} \in \mathbb{R}^{n \times p}$ be the matrix with imputed values, where w_{jq} is the imputed value for entry x_{jq} for $(j,q) \in \mathcal{M}$. The full imputation for observation $\boldsymbol{x}_j$ is referred to as $\boldsymbol{w}_j$. The idea is to consider the missing data problem as an optimization problem so that the key decision variables are the missing values $\left\{w_{jq} : (j,q) \in \mathcal{M}\right\}$.

Let $\boldsymbol{\mu} = \left\{\mu_1, \ldots, \mu_p\right\}$ be the mean vector and $\boldsymbol{\Sigma}$ be the covariance matrix of a complete dataset, then the squared Mahalanobis distance of the observation $\boldsymbol{w}_i$ from the mean $\boldsymbol{\mu}$ is given by

$$M_i(\boldsymbol{w}_i, \boldsymbol{\mu}) = \begin{bmatrix} w_{i1} - \mu_1 & \cdots & w_{ip} - \mu_p \end{bmatrix} \boldsymbol{\Sigma}^{-1} \begin{bmatrix} w_{i1} - \mu_1 \\ \vdots \\ w_{ip} - \mu_p \end{bmatrix}$$

assuming that the inverse covariance matrix exists. In the presence of missing data, our goal is to solve for the values of $\left\{w_{jq} : (j,q) \in \mathcal{M}\right\}$ that would minimize M_j for each $j \in J$. Thus, if we let D_j to be the set of indices of variables containing missing values for each $j \in J$, we have the following optimization problem,

$$\min f\left(w_{jd}\right) = M_j\left(\boldsymbol{w}_j, \boldsymbol{\mu}\right) \tag{2}$$

with respect to w_{jd} where $(j,d) \in \mathcal{M}$, that is, $d \in D_j$ for $|D_j| \leq p$ and $w_{jq} = x_{jq}$ for all $(j,\ q) \in \mathcal{N}$.

Model (2) will be run for a number of iterations with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ being updated from each iteration. For the first iteration, the initial values of the decision variables $\{w_{jq} : (j,q) \in \mathcal{M}\}$ are the mean values from all observed data in each variable so that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be computed from this initial imputed dataset. For the second iteration onwards, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be updated from the solutions of model (2). As an illustration, the schematic diagram of the procedure is given in Figure 1.

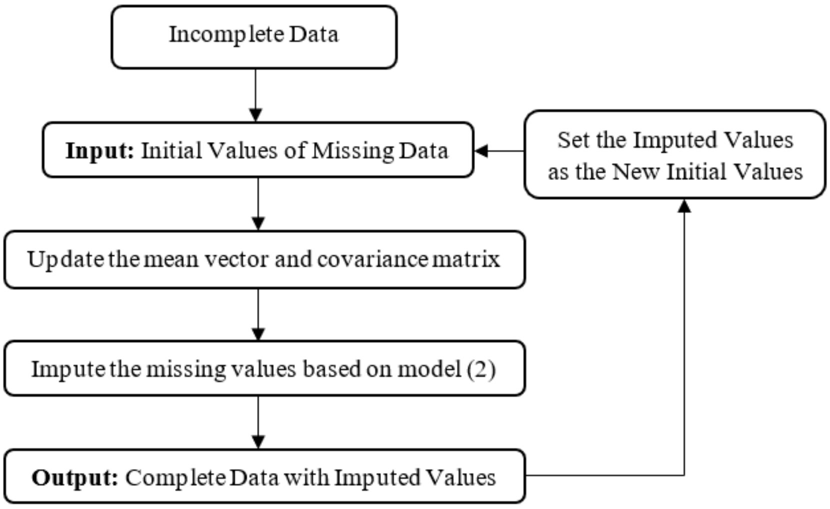
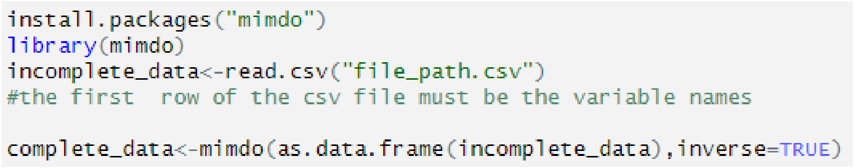

Incomplete Data
Input: Initial Values of Missing Data
Set the Imputed Values as the New Initial Values
Update the mean vector and covariance matrix
Impute the missing values based on model (2)
Output: Complete Data with Imputed Values

Figure 1: *Schematic Diagram of mimdo*

The default number of iterations is set to 30, but can be adjusted for large datasets to avoid long running time. After the desired number of iterations is performed, the final complete imputed dataset is then obtained based on the average imputed values. The proposed algorithm is now available on the Comprehensive *R* Archive Network (CRAN) as an *R* package called "mimdo: Multivariate Imputation by Mahalanobis Distance Optimization". The implementation is given as follows:



```
install.packages("mimdo")
library(mimdo)
incomplete_data<-read.csv("file_path.csv")
#the first  row of the csv file must be the variable names

complete_data<-mimdo(as.data.frame(incomplete_data),inverse=TRUE)
```

3. Results

This section presents the solution and properties of *mimdo*.

Proposition 1. For the case $|D_j| = 1$, let $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ be a given incomplete dataset with p variables. Also, let $\boldsymbol{W} \in \mathbb{R}^{n \times p}$ be the matrix with imputed values where w_{iq} is the imputed value for entry x_{iq}. If $(j,d) \in \mathcal{M}$, then the solution of the optimization problem (2) is given by

$$w_{jd} = \mu_d - \frac{1}{\sigma^*_{dd}} \sum_{a:a \neq d}^{p} \sigma^*_{da}\left(w_{ja} - \mu_a\right) \tag{3}$$

where $\mu_a, \mu_d, \sigma^*_{da} \in \mathbb{R}$ and $\sigma^*_{dd} > 0$.

Proof. Let $(j,d) \in \mathcal{M}$ and consider the optimization model (2). Now, we have to solve for w_{jd} that would minimize

$$M_j = \begin{bmatrix} w_{j1}-\mu_1 & \cdots & w_{jq}-\mu_q & \cdots & w_{jp}-\mu_p \end{bmatrix} \mathbf{\Sigma}^{-1} \begin{bmatrix} w_{j1}-\mu_1 \\ \vdots \\ w_{jq}-\mu_q \\ \vdots \\ w_{jp}-\mu_p \end{bmatrix}.$$

Suppose that

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \sigma^*_{11} & \cdots & \sigma^*_{1q} & \cdots & \sigma^*_{1p} \\ \vdots & & \vdots & & \vdots \\ \sigma^*_{q1} & \cdots & \sigma^*_{qq} & \cdots & \sigma^*_{qp} \\ \vdots & & \vdots & & \vdots \\ \sigma^*_{p1} & \cdots & \sigma^*_{pq} & \cdots & \sigma^*_{pp} \end{bmatrix}$$

where the diagonal elements are positive, then

$$M_j = \sum_{b=1}^{p}\sum_{a=1}^{p} \sigma^*_{ab}\left(w_{ja}-\mu_a\right)\left(w_{jb}-\mu_b\right).$$

To differentiate M_j, we have to separate the terms containing w_{jd} and using the symmetric property of the inverse covariance matrix, we have,

$$M_j = \sigma^*_{dd}\left(w_{jd}-\mu_d\right)^2 + 2\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{jd}-\mu_d\right)\left(w_{ja}-\mu_a\right) + \sum_{b:b\neq d}^{p}\sum_{a:a\neq d}^{p} \sigma^*_{ab}\left(w_{ja}-\mu_a\right)\left(w_{jb}-\mu_b\right)$$

$$\Rightarrow D_{w_{jd}}\left(M_j\right) = 2\sigma^*_{dd}\left(w_{jd}-\mu_d\right) + 2\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right).$$

Finally, equating the derivative to 0 will solve for the imputed value, that is,

$$2\sigma^*_{dd}\left(w_{jd}-\mu_d\right) + 2\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right) = 0$$

$$\Rightarrow \; w_{jd} = \mu_d - \frac{1}{\sigma^*_{dd}}\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right).$$

■

For the case when $|D_j| > 1$, we take the partial derivative with respect to w_{jd} for each $d \in D_j$, set to zero and obtain a system of equations of the form

$$\sigma^*_{d1}w_{j1} - \sigma^*_{d1}\mu_1 + \cdots + \sigma^*_{dd}w_{jd} - \sigma^*_{dd}\mu_d + \cdots + \sigma^*_{d,|D_j|}w_{j,|D_j|} - \sigma^*_{d,|D_j|}\mu_{|D_j|} + \sum_{q\notin D_j} \sigma^*_{dq}\left(w_{jq}-\mu_q\right) = 0$$

where we rewrite the subscripts as $\{1, 2, \ldots, |D_j|\}$ corresponding to the indices with missing values. Thus, in matrix form, we have,

$$\begin{bmatrix} \sigma^*_{11} & \sigma^*_{12} & \cdots & \sigma^*_{1,|D_j|} \\ \sigma^*_{21} & \sigma^*_{22} & & \sigma^*_{2,|D_j|} \\ \vdots & & \ddots & \vdots \\ \sigma^*_{|D_j|,1} & \sigma^*_{|D_j|,2} & \cdots & \sigma^*_{|D_j|,|D_j|} \end{bmatrix} \begin{bmatrix} w_{j1} \\ w_{j2} \\ \vdots \\ w_{j,|D_j|} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^*_{11}\mu_1 + \sum_{q\in D_j\backslash\{1\}} \sigma^*_{1q}\mu_q - \sum_{q\notin D_j} \sigma^*_{1q}\left(w_{jq}-\mu_q\right) \\ \sigma^*_{22}\mu_2 + \sum_{q\in D_j\backslash\{2\}} \sigma^*_{2q}\mu_q - \sum_{q\notin D_j} \sigma^*_{2q}\left(w_{jq}-\mu_q\right) \\ \vdots \\ \sigma^*_{|D_j|,|D_j|}\mu_{|D_j|} + \sum_{q\in D_j\backslash\{|D_j|\}} \sigma^*_{|D_j|,q}\mu_q - \sum_{q\notin D_j} \sigma^*_{|D_j|,q}\left(w_{jq}-\mu_q\right) \end{bmatrix}$$

in which the solution are the imputed values for w_{jd}, $d \in D_j$.

To show global optimality of w_{jd} in (3), we need to show that f is convex. This can easily be seen from the proof of Proposition 1 where

$$f = \sigma^*_{dd}\left(w_{jd}-\mu_d\right)^2 + 2\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{jd}-\mu_d\right)\left(w_{ja}-\mu_a\right) + \sum_{b:b\neq d}^{p}\sum_{a:a\neq d}^{p} \sigma^*_{ab}\left(w_{ja}-\mu_a\right)\left(w_{jb}-\mu_b\right)$$

which is a parabola. To verify this, the abscissa of the vertex of the parabola must contain the solution w_{jd}. Let $T = w_{jd}-\mu_d$, $\alpha = \sigma^*_{dd}$, $\beta = 2\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right)$, and $\gamma = \sum_{b:b\neq d}^{p}\sum_{a:a\neq d}^{p} \sigma^*_{ab}\left(w_{ja}-\mu_a\right)\left(w_{jb}-\mu_b\right)$, then

$$f = \alpha T^2 + \beta T + \gamma.$$

Now, the abscissa of the vertex of the parabola can be solved using the formula

$$T = -\frac{\beta}{2\alpha} = -\frac{2\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right)}{2\sigma^*_{dd}} = -\frac{1}{\sigma^*_{dd}}\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right)$$

$$\Rightarrow w_{jd}-\mu_d = -\frac{1}{\sigma^*_{dd}}\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right)$$

$$\Rightarrow \qquad w_{jd} = \mu_d - \frac{1}{\sigma^*_{dd}}\sum_{a:a\neq d}^{p} \sigma^*_{da}\left(w_{ja}-\mu_a\right)$$

which is the exact solution from Proposition 1. Thus, f is convex having w_{jd} as the global minimum. Figure 2 illustrates the model f with its global minimum in two-dimensional space assuming $\left|D_j\right| = 1$.

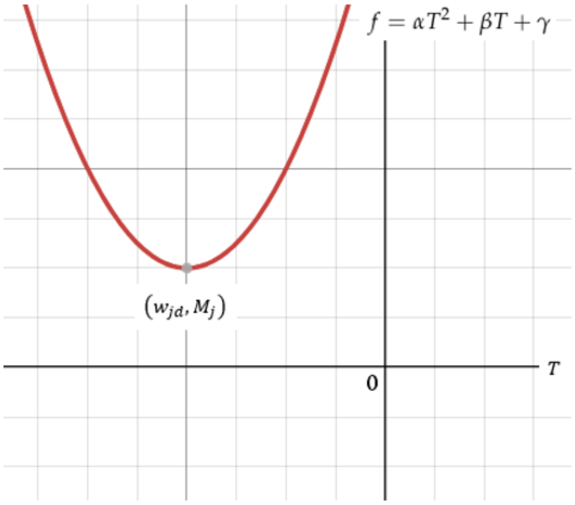


Figure 2: *Graph of Model (2) for the case* $\left|D_j\right| = 1$

As we can see from Figure 2, the vertex contains the global minimum which is the solution w_{jd} and the corresponding minimum Mahalanobis distance M_j. The next result shows the convergence of the solution of *mimdo*.

Conjecture 1. Let f be the objective function given in (2). If $f^{(t)}$ is the objective function value at iteration t, then

$$\lim_{t\to\infty}\left(f^{(t)} - f^{(t-1)}\right) = 0.$$

Conjecture 1 implies that as the number of iteration increases, the objective function in model (2) will likely to converge to a specific value. This means that for each $j \in J$, the imputed values $\left\{w_{jd}\right\}$, $d \in D_j$ converges to specific values. This conjecture will be shown in the application in the next section.

4. Application

For demonstration, *mimdo* is applied on popular datasets taken from the UCI (University of California Irvine) Machine Learning Repository which have been widely used by students, educators, and researchers all over the world. These datasets are the "Iris Flower" and "Yacht Hydrodynamics" downloaded at *https://archive.ics.uci.edu/datasets* which will be used for the empirical analysis of data imputation algorithms and is a good point of reference for assessment or comparison of different imputation methods.

In doing the experiment, incomplete datasets are generated (5%-30%) from complete datasets using MCAR (missing completely at random) and MAR (missing at random) missing data mechanisms from the *R* Package *missMethods*. For classification (Iris Flower), the percentage of missing data is generated per cluster while for regression (Yacht Hydrodynamics), the percentage of missing data is generated directly from the whole dataset. Specifically, 100 incomplete datasets will then be generated per missing rate. For assessment purposes, the missing elements among the data are then imputed using *mimdo*, *mice*, and *knn* (k-nearest neighbor) where *mice* is set to one single imputation with 30 maximum iterations (*mice* - *R* Package *mice*, *knn* - *R* Package *bnstruct*).

For classification task, the traditional K-means algorithm is applied into the imputed dataset and obtain the clustering accuracy (classification rate) based on the original cluster labels. For regression analysis, the imputed dataset is regressed with respect to the dependent variable and obtain the R-squared. The average result from 100 incomplete datasets is then reported. Figure 3 shows the average classification rate results using Iris dataset while Figure 4 shows the average R-squared results using Yacht dataset.

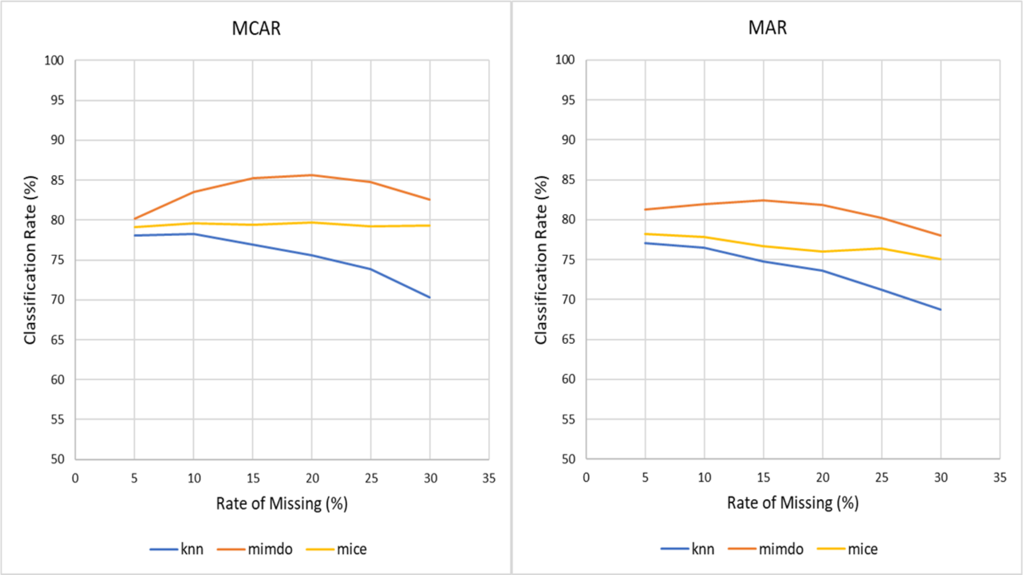


Figure 3: *Classification Rate Results from Iris Dataset*

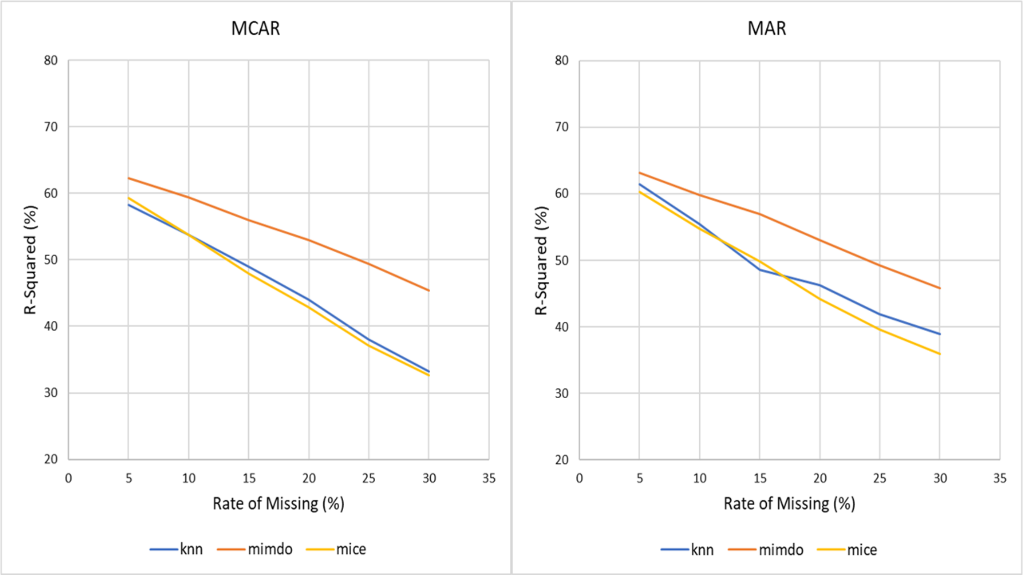


Figure 4: *Regression Results from Yacht Dataset*

It can be observed from Figures 3 & 4 that the results from *mimdo* are higher especially with MCAR assumption. Consistency of this result will be explored in section 4.2 using simulation studies.

4.1. Convergence of Solutions

This section shows the corresponding results of the convergence of the solutions of *mimdo*. This means that as the iteration increases, the objective function in model (2) will likely to converge to a specific value. For example, Figure 5 shows the average objective value differences between each consecutive iterations.

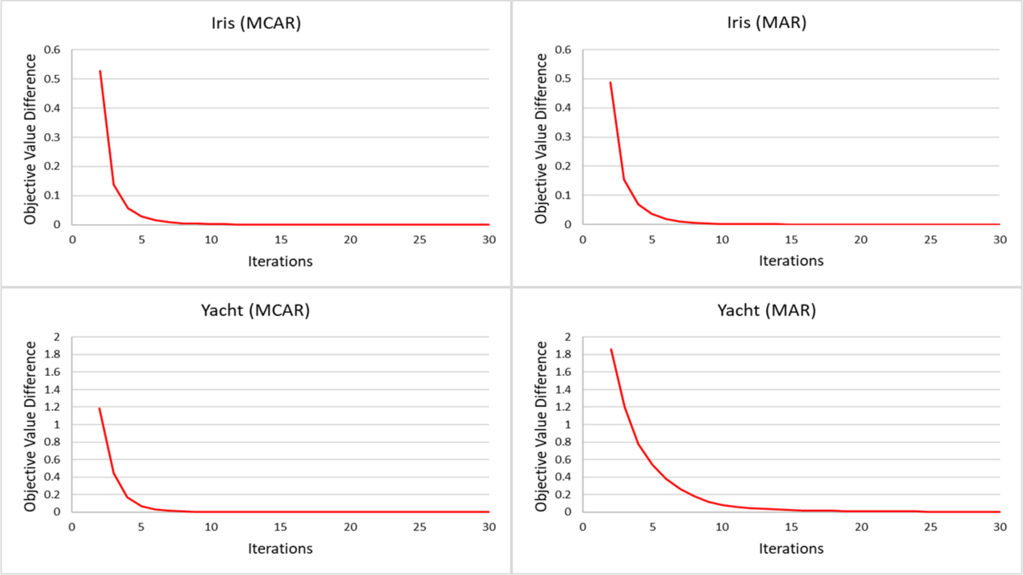


Figure 5: *Overall Average Variation*

It can be observed from Figure 5 that the graph decreases eventually to zero which indicates the consistency of the solution (imputed values) of *mimdo* as stated in Conjecture 1.

4.2. Simulated Datasets

In this section, we generate simulated datasets that have the same distributions and correlations as the existing datasets considered using the *simdf()* function from the *R* Package *faux*. Specifically, 100 simulated datasets were generated for each dataset considered and the average results were reported. Figures 6 & 7 show the average results using simulated Iris and simulated Yacht datasets respectively.

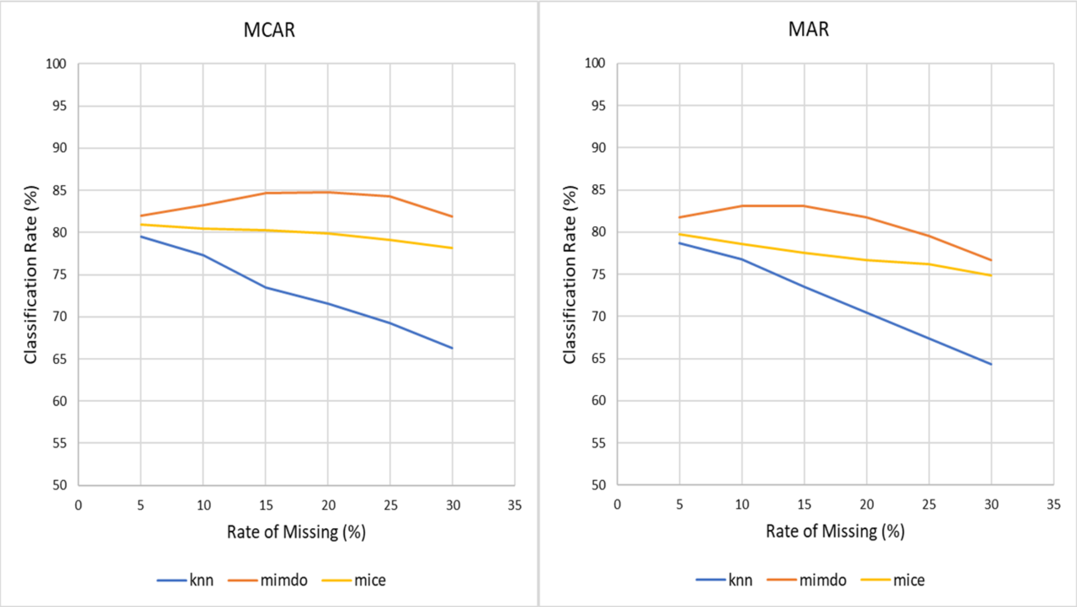


Figure 6: *Classification Rate Results from Simulated Iris Dataset*

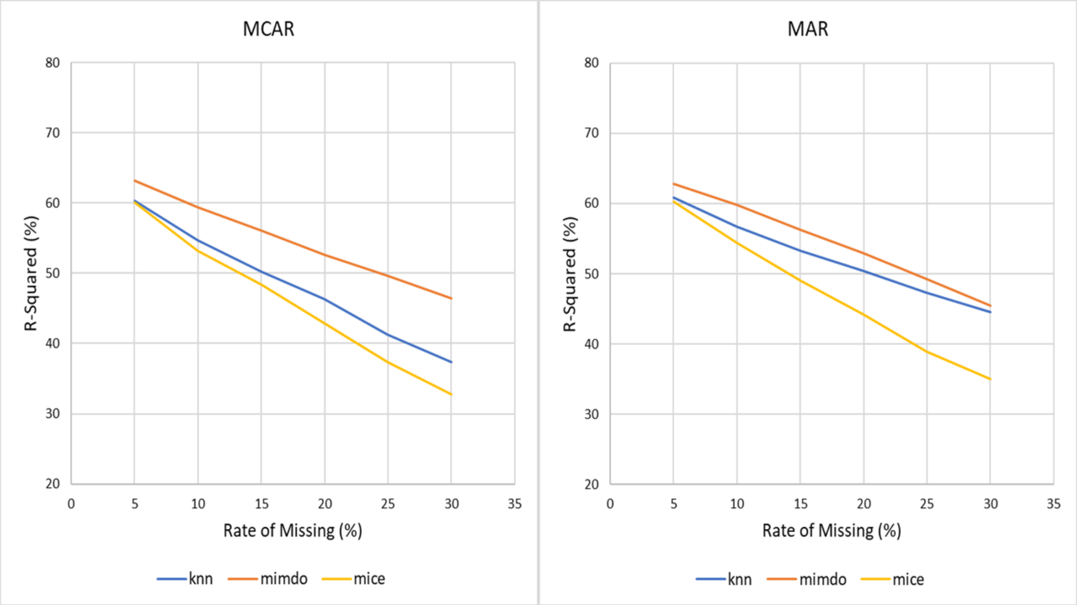
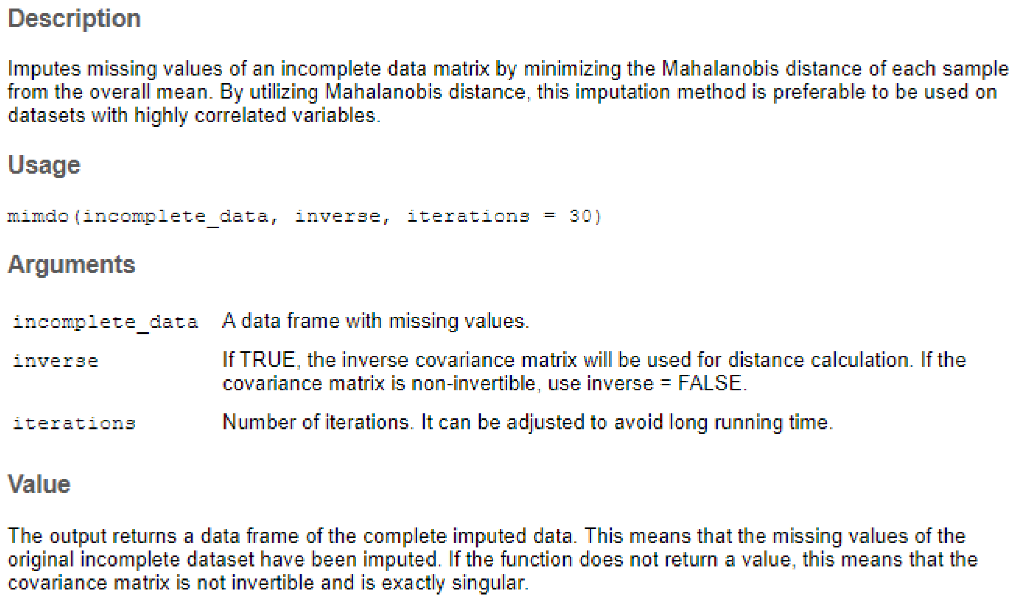


Figure 7: *Regression Results from Simulated Yacht Dataset*

It can be observed from Figures 6 & 7 the similarity of the results from the original dataset. This shows consistency of *mimdo* for this type of datasets.

4.3. *R* Package "mimdo"

This section presents the documentation of the *R* Package *mimdo*. It is built for the purpose of imputing highly correlated datasets up to the maximum tolerance of covariance matrix invertibility. However, if the covariance matrix is non-invertible, the option inverse=FALSE might not work very well. Thus, for this case, we recommend feature selection before doing *mimdo*.



Description

Imputes missing values of an incomplete data matrix by minimizing the Mahalanobis distance of each sample from the overall mean. By utilizing Mahalanobis distance, this imputation method is preferable to be used on datasets with highly correlated variables.

Usage

```
mimdo(incomplete_data, inverse, iterations = 30)
```

Arguments

`incomplete_data`	A data frame with missing values.
`inverse`	If TRUE, the inverse covariance matrix will be used for distance calculation. If the covariance matrix is non-invertible, use inverse = FALSE.
`iterations`	Number of iterations. It can be adjusted to avoid long running time.

Value

The output returns a data frame of the complete imputed data. This means that the missing values of the original incomplete dataset have been imputed. If the function does not return a value, this means that the covariance matrix is not invertible and is exactly singular.

5. Conclusions

The problem about missing data increases significantly in many fields especially if the dataset have highly correlated variables. For some cases, other data imputation algorithms will leave those variables out of the imputation process. The use of Mahalanobis distance in the imputation process provides an efficient estimation of missing data regardless of the correlation structure of the dataset.

In this paper, package *mimdo* accessible via R interface was applied to benchmark datasets and simulated datasets. On the average, using benchmark datasets, it achieved a clustering accuracy of 82.3% (classification task) and an R-squared of 54.4% (regression task) while using simulated datasets, it obtained 82.2% clustering accuracy and 54.5% R-squared. This means that the imputation method applied to this type of datasets is consistent.

The consistency of the solution (imputed values) of *mimdo* can also be observed from Figure 5 wherein the graphs decreased eventually to zero which means that the objective function in (2) will eventually converge.

References

[1] Andreasson, N., Evgrafov, A., & Patriksson, M. (2005). An introduction to optimization: Foundations and fundamental algorithms. *Chalmers University of Technology Press: Gothenburg, Sweden*, *1*, 1-205.

[2] Barrios, A., Trincado, G., & Garreaud, R. (2018). Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. *Forest Ecosystems*, *5*(1), 1-10.

[3] Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2017). From predictive methods to missing data imputation: an optimization approach. *J. Mach. Learn. Res.*, *18*(1), 7133-7171.

[4] Bock, T. (2013). 5 ways to deal with missing data in cluster analysis. *DISPLAYR Blog*.

[5] Boluki, S., Zamani Dadaneh, S., Qian, X., & Dougherty, E.R. (2018, August). Optimal clustering with missing values. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 593-594).

[6] Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(1), 67-82.

[7] Katitas, A. (2019). Getting Started with Multiple Imputation in R. *University of Virginia Library-Statlab Articles*, https://uvastatlab.github.io/2019/05/01/getting-started-with-multiple-imputation-in-r.

[8] Labita, GJ.D. (2024). Package 'mimdo'. *Comprehensive R Archive Network (CRAN)*. doi:10.32614/CRAN.package.mimdo

[9] Song, Q. & Shepperd, M. (2007). Missing data imputation techniques. *International journal of business intelligence and data mining*, *2*(3), 261-291.

[10] Wagstaff, K. & Laidler, V. (2005). Making the most of missing values: Object clustering with partial data in astronomy. *Astronomical Data Analysis Software and Systems XIV*, 347, 172-176.