

MISSING VALUE IMPUTATION FOR MULTIVARIATE DATA USING DATA DEPTH

Aman Bashir Sheikh

Enrollment No: 2022MSST003

M.Sc.Statistics



Under the guidance of

Dr. Mahesh Barale

Department of Statistics

Central University of Rajasthan, Ajmer

Plan to Talk

- 1 Abstract & Keywords
- 2 Introduction
- 3 Depth Function
- 4 Proposed Method
- 5 Performance Study
- 6 References

Abstract

The problem of missing data occurs in various practical situations and imputation of missing data becomes important task. This study addresses the prevalent issue of missing values in multivariate data. We are surrounded by missing data. Problems created by missing data in statistical analysis have long been swept under the carpet. These times are now slowly coming to an end. Despite the existence of numerous data imputation methods, this research introduces the innovative utilization of data depth to handle missing data in multivariate datasets. Data depth is gaining recognition in the field of multivariate analysis. By integrating data depth with established techniques, the aim is to increase the accuracy and robustness of imputation method, thereby enhancing overall data analysis outcomes.

Keywords

Data Depth, Missing Values, Multivariate Data, Normalized Root Mean Square Error.

Introduction

- The credibility of our findings always depend on the quality of the data.
- It is essential to address missing value challenges in data analysis.
- The existing methods are time-consuming, and lack desired accuracy.
- By incorporating data depth, we propose a method for imputation to enhance efficiency, and accuracy.
- By introducing this method, we aim to maximize analysis reliability via data depth.

Challenges Arising Due To Missing Values

Biased
Estimate

Loss of
Information

Misleading
Insights

Challenges
to Model
Building

Missing Data Mechanism

Corresponding to every observation X , there is a missing value indicator R , defined as:

$$R = \begin{cases} 1 & \text{if } X \text{ is observed} \\ 0 & \text{if } X \text{ is missing} \end{cases}$$

$$p(R|X_o, X_m)$$

- Missing Completely at Random (MCAR): pattern of missing value is totally random and does not depend on any variable.
- Missing at Random (MAR): the probability of missing data is dependent on observed information in the dataset.
- Missing Not at Random (MNAR): missingness is dependent on unobserved data rather than observed data.

Existing Methods

Numerous methods already exist in the literature

- Mean imputation
- Regression Imputation
- KNN Imputation
- Random Forest

Data Depth

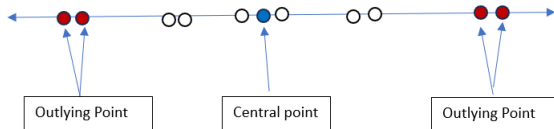
Depth Function

Depth Function

- [Tukey, 1975] firstly introduced concept of data depth.
- Depth function is extension to univariate notations of median, quantiles, ranks, and order statistics to the setting of multivariate data.
- Whereas probability density function measures local probability weight, a depth function measures centrality.

Data Depth

One Dimension



- Data depth of a point measure the relative position of that point in a given data cloud.
- Depth of point

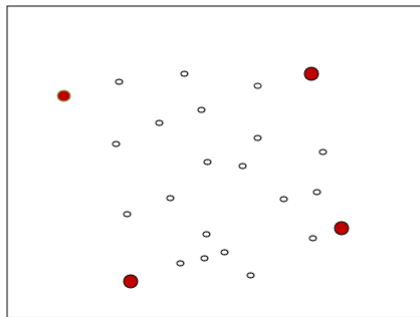
$$x = \min\{Fn(x), 1 - Fn(x)\} \quad (1)$$

, where $F_n(\cdot)$ is the proportion data point that are on the left of x .

- Maximum possible value of depth function is $1/2$.

Data Depth

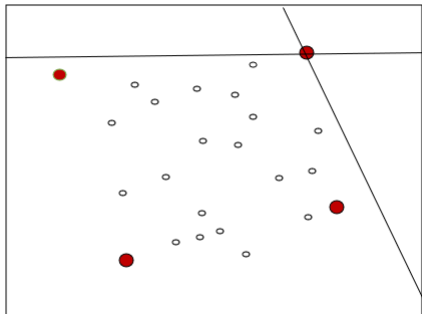
More Than one Dimension



- How to extend for more than one dimension R^p ?

Data Depth

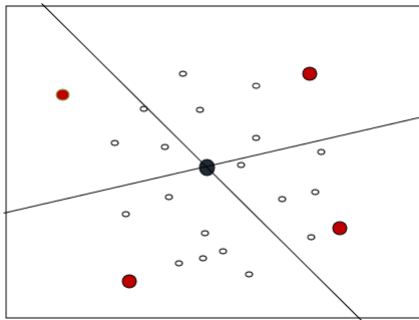
More Than one Dimension



- Consider line through x (hyperplane for $p \geq 3$).
- look at the proportion of the data points lying on the two sides of the line.

Data Depth

More Than one Dimension



- The line through the black point has almost equal proportion of data points in both half spaces.

Depth function

Different types of depth Function

- Mahalanobis Depth ([Mahalanobis, 2018])
- Tukeys half-space depth ([Tukey, 1975])
- Simplicial Depth ([Liu, 1990])
- Spatial depth ([Vardi and Zhang, 2000])
- Projection Depth ([Zuo and Serfling, 2000]), etc.

Depth Function

Mahalanobis Depth

The Mahalanobis Depth of a point $x \in S_n \subset R^p$ relative to a p-dimensional data set S_n is defined as :

$$M_h D(x; s_n) = \frac{1}{[1 + (x - \bar{x})^T S^{-1} (x - \bar{x})]} \quad (2)$$

Where \bar{x} and S are the mean vector and dispersion matrix of S_n ([Mahalanobis, 2018]).

Forward Thinking Approach

K-max method

- Let X be the *Matrix* of any dimension.
- Identify rows containing missing values, *i.e.* incomplete rows.
- Divide the data into two matrices, say Complete data (X_c) and incomplete data (X_m), based on incomplete rows.
- Construct the new matrix, named as X_{dummy} which has same dimensions as Mid (X_{mid}). by replicating i^{th} missing row.
- Find the depth of $D(X_{dummy}, X_{mid})$
- Extract K points which have *maximum* depth from complete data (X_{mid}).
- Take mean of those K points and impute at the place of missing value in the provided dataset X .

Forward Thinking Approach

Iterative Imputation

- Let X be the *Matrix* of any dimension.
- Identify rows containing missing values, *i.e.* incomplete rows.
- Divide the data into two matrices, say Complete data (X_c) and incomplete data (X_m), based on incomplete rows.
- Construct the new matrix, named as X_{dummy} which has same dimensions as Mid (X_{mid}). by replicating i^{th} missing row.
- Find the depth of $D(X_{dummy}, X_{mid})$
- Extract K points which have *maximum* depth from complete data (X_{mid}).
- Take mean of those K points and impute at the place of missing value in the provided dataset $X_{Imputed}$.
- Take the absolute difference of *i.e* $|X_{Imputed} - X_{Mid}|$.
- If the difference is less than ε *i.e* Specified error, then stop iteration otherwise repeat the above procedure.

Forward Thinking Approach

Class Based Imputation

- Let X be the matrix of any dimensions.
- Identify the categorical column and determine the number of categories it contains.
- Calculate the depth of data for each category $D(X_{cat[i]}, X_{Mid})$ and extract " K " points with the highest depth.
- Take a mean of that " K " points and impute at the place of missing value $X_{Imputed}$
- Take the absolute difference between $X_{Imputed}$ & X_{Mid} i.e $|X_{Imputed} - X_{Mid}|$
- If the Difference is less than ε ;i.e Specified Error then Stop the procedure O.W. repeat the above method

Performance Study

NRMSE

The performance of the proposed method will be evaluate using *Normalized Root Mean Square Error*

$$NRMSE = \sqrt{\frac{\text{mean}(X_{\text{original}} - X_{\text{imputed}})^2}{\text{var}(X_{\text{original}})}} \quad (3)$$

Where,

- X_{imputed} is imputed matrix
- X_{original} is Original Data (*without missing value*)

Performance Study

NRMSE Comparison

The proposed method is compare with the following methods:

- Mean imputation
- KNN imputation [Pujianto et al., 2019]
- Random Forest imputation [Stekhoven and Bühlmann, 2012]

Performance Study

NRMSE Comparison

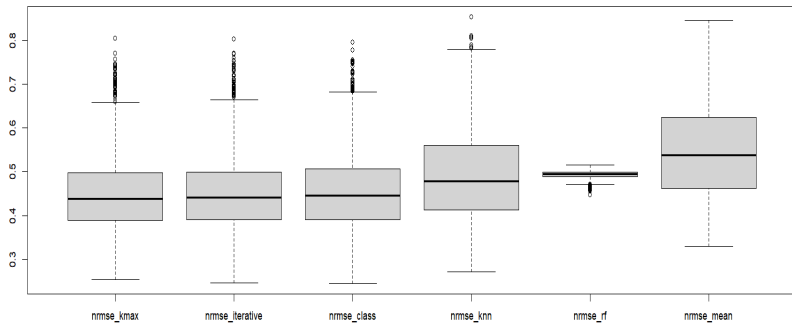
Dataset Consider

Name	Columns	Rows
Bupa	6	345
banknoten	7	200

- produced missing value by (*MCAR*)
- missing percentage 5, 10, and 15
- Depth Function : *Mahalanobis*

Performance Study

Bupa Dataset 5%

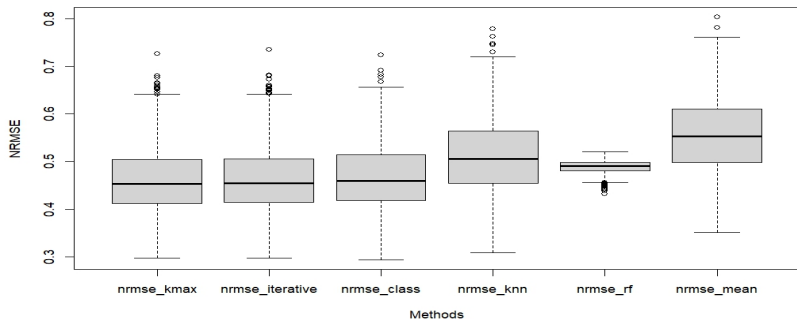


NRMSE comparison for Bupa data with 5% missing

The K-Max method has the smallest NRMSE compared to the other methods.

Performance Study

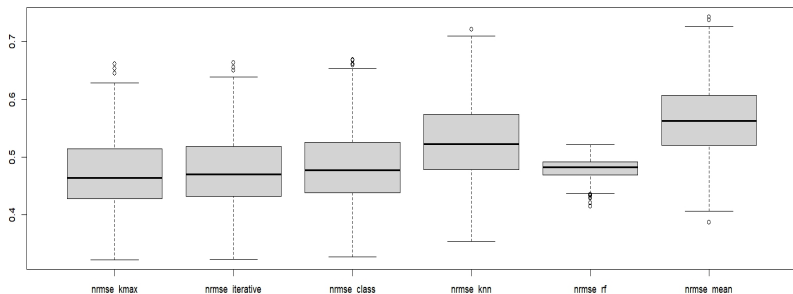
Bupa Dataset 10%



NRMSE comparison for Bupa data with 10% missing
The K-Max method has the smallest NRMSE compared to the other methods.

Performance Study

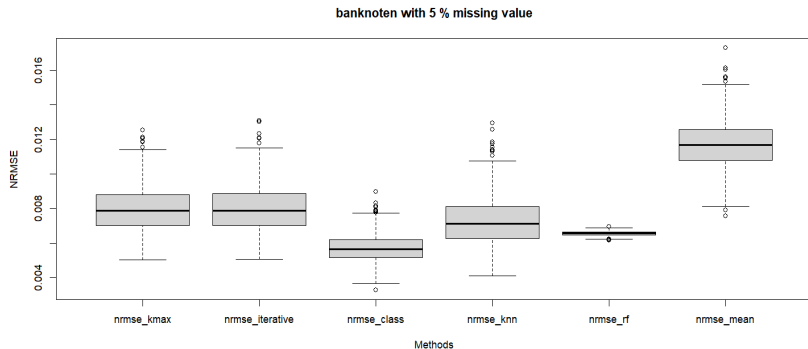
Bupa Dataset 15%



NRMSE comparison for Bupa data with 15% missing
The K-Max method has the smallest NRMSE compared to the other methods.

Performance Study

banknoten Dataset 5%

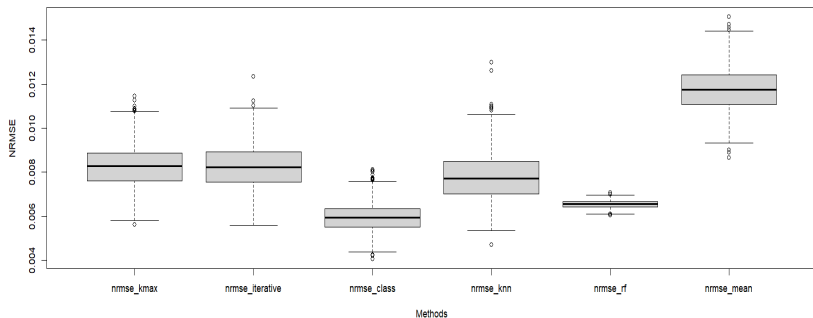


NRMSE comparison for banknoten data with 5% missing

The Class Based Imputation method has the smallest Average NRMSE compared to the other methods. i.e 0.005688

Performance Study

banknoten Dataset 10%

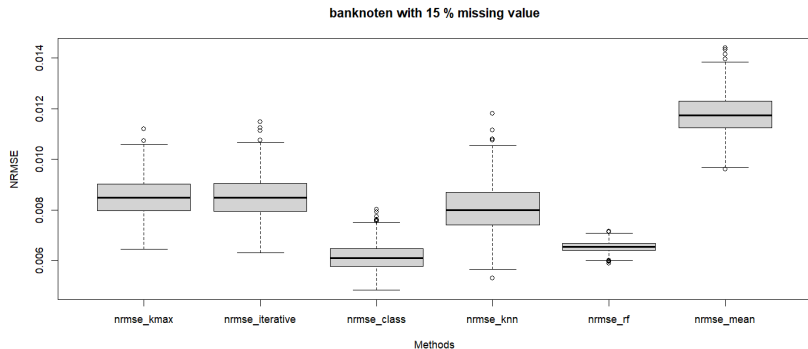


NRMSE comparison for banknoten data with 10% missing

The Class Based Imputation method has the smallest Average NRMSE compared to the other methods. i.e 0.005953

Performance Study

banknoten Dataset 15%

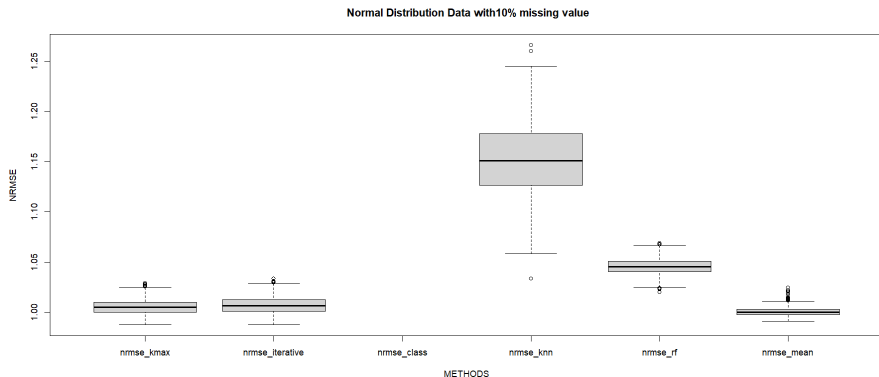


NRMSE comparison for banknoten data with 15% missing

The Class Based Imputation method has the smallest Average NRMSE compared to the other methods. i.e 0.006095

Performance Study

Normal Distribution Dataset 10%



NRMSE comparison for normal distribution data with 10% missing
K-max method has smallest NRMSE among all other methods i.e. 1.00663

Average NRMSE

Table: Average NRMSE Comparison

		Datasets	Bupa			Banknoten			Normal Distribution
		Method\Missing Percentage	5%	10%	15%	5%	10%	15%	10%
AVERAGE NRMSE	Proposed Method	K-Max	0.453497	0.464624	0.473875	0.007964	0.008281	0.008443	1.00663
		Iterative Imputation	0.455851	0.46668	0.477528	0.007971	0.008267	0.008419	1.006946
		Class-based Imputation	0.458311	0.47095	0.483791	0.005688	0.005953	0.006095	-
	Existing Method	KNN	0.495776	0.51278	0.526566	0.007226	0.007795	1.09112	1.15425
		Random Forest	0.493108	0.487748	0.480133	0.006557	0.006549	0.006536	1.04546
		Mean Imputation	0.548099	0.55629	0.56305	0.011695	0.01173	0.011749	1.00942

In the table above, we are comparing the average NRMSE of our proposed method with existing methods. Among the existing methods, K-Max and Class-Impute yield better results compared to the others.

Time Comparision

In the previous slides, we examined the average NRMSE comparison, revealing that K-Max and Class-based Imputation outperformed other methods. Now, we shift our focus to comparing the computational time required for imputing missing values.

Time Comparision

We have considered data from a normal distribution with parameter $\mu = (0,0,0,0,0)$ & $\Sigma = \text{diag}(1,1,1,1,1)$ having 10% missing values.

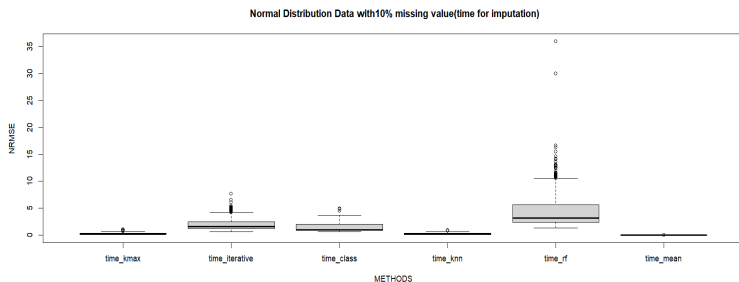


Table: AVERAGE Time for Imputation

METHOD	K-MAX	ITERATIVE IMPUTATION	IMPUTE CLASS	KNN	RANDOM FOREST	MEAN IMPUTATION
NRMSE	0.265664	1.97226	1.3320110	0.22018676	4.307334	0.0037

References I



Liu, R. Y. (1990).

On a notion of data depth based on random simplices.
The Annals of Statistics, pages 405–414.



Mahalanobis, P. C. (2018).

On the generalized distance in statistics.
Sankhyā: The Indian Journal of Statistics, Series A (2008-), 80:S1–S7.



Pujianto, U., Wibawa, A. P., Akbar, M. I., et al. (2019).

K-nearest neighbor (k-nn) based missing data imputation.
In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88. IEEE.



Stekhoven, D. J. and Bühlmann, P. (2012).

Missforest—non-parametric missing value imputation for mixed-type data.
Bioinformatics, 28(1):112–118.

References II



Tukey, J. W. (1975).

Mathematics and the picturing of data.

In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531.



Vardi, Y. and Zhang, C.-H. (2000).

The multivariate L_1 -median and associated data depth.

Proceedings of the National Academy of Sciences, 97(4):1423–1426.



Zuo, Y. and Serfling, R. (2000).

General notions of statistical depth function.

Annals of statistics, pages 461–482.

A large, irregular teal brushstroke graphic serves as a background for the text.

*Thank
You!*