Journal Paper Evaluation

TITLE: A genetic algorithm for multivariate missing data imputation

AUTHOR(S): Juan Carlos Figueroa-Garcia, Roman Neruda, German Hernandez-Perez

JOURNAL: Information Sciences.

DATE: Available online 17 November 2022

VOLUME: 619

PAGES: 21

INTRODUCTION:

1.DOES THE TITLE OF THE RESEARCH ARTICLE GIVE ANY INDCICATION OF THE TYPE OF STUDY BEING REPORTED: IE., DESCRIPTIVE, CORRELATIONAL, OR CASUAL-COMPARATIVE

The title of the research article "genetic algorithm for Multivariate missing data imputation" does not explicitly indicate the type of study being reported, such as descriptive, correlational, or causal-comparative. The title suggests that the focus of the research is on handling missing data in multivariate datasets using a genetic algorithm. To determine the specific type of study conducted in the research article, one would need to delve into the content of the paper itself and examine the methodology, data analysis techniques, and research objectives outlined within the document.

2.WERE THE INDEPENDENT AND DEPENDENT VARIABLES MENTIONED IN THE TITLE?

The independent and dependent variables were not explicitly mentioned in the multivariate missing data imputation. the title focuses on the topic of handling missing data in multivariate datasets using a genetic algorithm

3.IN WHAT PART OF THE ARTICLE DID YOU FIND WHAT KIND OF STATISTICAL TOOLS WERE BEING USED?

The statistical tools and concepts employed to address multivariate missing data imputation encompass several key methodologies. Firstly, the

Expectation-Maximization (EM) Algorithm is utilized for parameter estimation in models with latent variables, aiding in the retrieval of missing data. Multivariate Analysis plays a pivotal role, discerning means, covariance, correlations, and skewness while upholding the multivariate structure. Minkowski Distance serves to gauge disparities between available and completed data, crucial for assessing imputation accuracy. A Fitness Function is meticulously crafted to steer the optimization process towards imputations that faithfully mirror the original data's statistical properties. Dimensionality Reduction techniques are harnessed, facilitating simplified representations of multivariate data to streamline optimization. Additionally, Goodness of Fit Analysis is employed to select appropriate random variable generators, ensuring that generated values uphold the dataset's statistical characteristics. These tools collectively form a robust framework for tackling multivariate missing data imputation, ensuring the fidelity of the imputed data to the original dataset's statistical properties.

B. ANALYSING THE VARIABLES:

DEPENDENT VARIABLES:

In the research article "Multivariate missing data," the specific dependent variables are not explicitly outlined. However, in studies focusing on missing data handling using a genetic algorithm, dependent variables typically pertain to the accuracy and effectiveness of the imputation process. These may include metrics such as the accuracy of imputed values, measured through statistical indicators like RMSE or MAE. Additionally, preservation of matrix properties, such as means, covariances, correlations, and skewness, could serve as dependent variables, assessing how well the imputation method maintains the original dataset's characteristics. Moreover, the performance of the imputation algorithm itself, particularly the genetic algorithm or other statistical techniques employed, may be a key focus. To ascertain the precise dependent variables used in the study, referencing the full text of the research article, particularly the methodology, results, and discussion sections, is necessary, where researchers detail the variables of interest and the outcomes of the analysis**.**

INDEPENDENT VARIABLES:

In the research article "Multivariate missing data," specific independent variables are not explicitly mentioned. However, in studies focusing on missing data handling using a genetic algorithm, independent variables may encompass various factors. These could include manipulating the type or pattern of missing data, varying percentages of missing values, and selecting different imputation methods like the genetic algorithm. Additionally, characteristics intrinsic to the multivariate datasets, such as the number of variables, sample size, and distribution, could serve as independent variables. To ascertain the specific independent variables used in the study, one must refer to the full text of the research article, particularly the methodology section, where the experimental design and factors considered in missing data imputation are detailed.

HYPOTHESIS:

1. WERE THE HYPOTHESIS CLEAR AND UNDERSTANDABLE?

Yes, the hypothesis of the study is clear and understandable. The authors propose a genetic algorithm-based approach, termed Multiple Imputation Genetic Algorithm (MIGA), for imputing missing observations in multivariate data. They aim to address the limitations of classical estimation methods, which often perform poorly when dealing with multivariate data and multiple missing observations. That can be summarized by the help of Proposed solution, Hypothesis, Validation.Proposed approach leverages genetic algorithms to provide a flexible and efficient solution that preserves important statistical properties of the data.

2.WHAT WAS THE HYPOTHESIS (OR HYPOTHESES)? WHAT WAS THE NULL HYPOTHESIS (ES)? WAS IT APPROPRIATE FOR THE STUDY?

The hypothesis of the study was that the Multiple Imputation Genetic Algorithm (MIGA) can effectively impute missing observations in multivariate data by minimizing a new multi-objective fitness function, thus preserving the original statistical properties of the dataset even in the presence of missing data. The null hypothesis (H0) could be formulated as there is no significant difference in imputation accuracy between MIGA and other existing methods, such as the Expectation-Maximization (EM)

algorithm and auxiliary regressions, when applied to multivariate datasets with missing observations.

Hypothesis is appropriate to study the comparison between proposed method and established techniques is allowing for the evaluation of MIGA's effectiveness in addressing the challenges.

3.DID THE INTRODUCTION ADEQUATELY SET UP THE HYPOTHESES?

Yes, the introduction effectively establishes the hypotheses by providing context on the missing data problem in multivariate datasets and outlining the proposed solution using the Multiple Imputation Genetic Algorithm (MIGA). It articulates the motivation behind the study, highlighting the limitations of existing methods and the need for improved imputation techniques. The introduction clearly states the hypothesis that MIGA can effectively impute missing observations while preserving the original statistical properties of the dataset. Additionally, it implicitly presents the null hypothesis by emphasizing the comparison between MIGA and other established methods, suggesting that MIGA's performance may not significantly differ from existing techniques. Overall, the introduction adequately sets up the hypotheses by framing the research problem, proposing a solution, and hinting at the comparative analysis to be conducted.

Based on the implied hypotheses discussed in the excerpts from the research article "Multivariate missing data," we can attempt to state the null hypothesis for each alternative hypothesis as follows:

4.ATTEMPT TO STATE THE NULL HYPOTHESIS FOR EACH ALTERNATIVE HYPOTHESIS?

1. Alternative Hypothesis: Handling discrete data leads to NP-Hard computations, and the search space for solutions exponentially increases with the amount of missing values.

Null Hypothesis: Handling discrete data does not lead to NP-Hard computations, and the search space for solutions does not exponentially increase with the amount of missing values.

2. Alternative Hypothesis: Multivariate problems require preserving not only individual averages/variances but also covariances/correlations, posing challenges in maintaining dataset properties during imputation.

Null Hypothesis: Multivariate problems do not require preserving individual averages/variances and covariances/correlations during imputation.

3. Alternative Hypothesis: The complexity of finding an optimal solution for missing data imputation increases with the nature of variables (continuous/discrete/binary), matrix properties to be preserved, and the sample size with missing values.

Null Hypothesis: The complexity of finding an optimal solution for missing data imputation is not influenced by the nature of variables, matrix properties, or sample size with missing values.

It is important to note that the above statements are formulated based on the implied hypotheses and may not directly align with traditional null and alternative hypothesis formats. The hypotheses discussed in the research article focus on the challenges and complexities of handling multivariate missing data, and the null hypotheses presented here attempt to provide a structured representation of the opposite scenarios to the alternative hypotheses.

5.DID THE AUTHORS SPECIFY A SPECIFIC ALPHA RISK LEVEL FOR REJECTING THE NULL HYPOTHESIS? IF SO, WHAT WAS IT? IF THEY DID NOT SPECIFY THE ALPHA RISK LEVEL, WHAT DO YOU THINK IT MUST HAVE BEEN?

Based on the information provided in the excerpts from the research article "Multivariate missing data," the authors did specify a specific alpha risk level for rejecting the null hypothesis in the context of statistical tests conducted in the study. For example, in the section discussing the output analysis and multivariate tests on means and variances, the authors mention using a significance level of $\alpha = 0.05$ for certain tests such as the t-student test, f test, and v^2 distribution tests.

If the authors did not explicitly specify the alpha risk level for rejecting the null hypothesis in other parts of the study, it is common in statistical analyses to use a standard significance level of $\alpha = 0.05$. This level is widely accepted in research as a standard threshold for determining statistical significance. Therefore, if not explicitly stated, it is reasonable to assume that the alpha risk level in the study was likely set at $\alpha = 0.05$ for most statistical tests unless otherwise specified.

Setting the alpha risk level at 0.05 is a common practice in statistical hypothesis testing as it provides a balance between Type I and Type II errors and is widely accepted in many fields of research.

SAMPLE:

DO YOU BELIEVE THAT THE SAMPLE WAS LARGE ENOUGH?

The research article "Multivariate missing data" does not explicitly mention whether the sample size was considered adequate for the analyses conducted. Sample size adequacy depends on factors like research design, effect size, and desired statistical power. Without specific details, it's challenging to assess if the sample size was sufficient for multivariate analyses. Typically, larger samples are preferred for robust results in complex statistical tests. Researchers often conduct power analyses to determine adequate sample sizes based on effect sizes and desired power levels. Balance between practical constraints and statistical requirements guides sample size determination.

GIVEN THE SAMPLE SIZE COULD YOU COMPUTE THE STANDARD ERROR OF THE MEAN. TO ACCOMPLISH THIS YOU WOULD NEED THE VALUES FOR BOTH N AND THE STANDARD DEVIATION. DID THEY PROVIDE YOU WITH THIS DATA WHAT DO YOU BELIEVE THE CRITICAL REGION" FOR REJECTION OF THE NULL HYPOTHESIS SHOULD HAVE BEEN.

To compute the standard error of the mean, you would need the sample size (N) and the standard deviation. Unfortunately, the provided PDF file does not contain specific values for N and the standard deviation that are necessary for this calculation.

Regarding the critical region for rejection of the null hypothesis, it typically depends on the significance level chosen for the hypothesis test. Common significance levels include 0.05, 0.01, and 0.1. Without knowing the specific significance level used in the study described in the PDF, it is challenging to determine the exact critical region for rejection of the null hypothesis. Then may be able to calculate the standard error of the mean and determine the region based on significance level specified in analysis.

RESULTS AND CONCLUSION:

ARE APPROPRIATE STATISTICAL TOOLS USED?

Yes, appropriate statistical tools are used in the study on genetic algorithms for multivariate missing data imputation. The document mentions the use of statistical tests to evaluate the performance of different methods, including the genetic algorithm (MIGA), the Expectation Maximization (EM) algorithm, and Auxiliary Regressions (AR). These statistical tests assess various aspects such as means, individual variances, covariances, correlations, and skewness. It compares the results obtained from different method using the tools and evaluating the accuracy and effectiveness of the methods in handling the multivariate missing data.

WERE GRAPHIC CHARTS USED? IF SO, WERE THEY HELPFUL IN SHOWING THE RESULTS.

In this paper graphics charts were not used. If graphic charts were used in the study, they could have been helpful in visually presenting the results of the genetic algorithm for imputing missing data. Graphical representations can make complex data more accessible and understandable, allowing researchers to quickly identify patterns, outliers, and the effectiveness of the imputation method.

DOES THE INVESTIGATOR RELATE THE RESULTS TO THE HYPOTHESIS.

Results to the hypothesis in the context of the study on genetic algorithms for multivariate missing data imputation. However, in scientific research, it is common practice for investigators to discuss how the results align with the initial hypothesis or research questions. When interpreting the results of a study, researchers typically compare the findings to the original

hypothesis to determine whether the data support or refute the hypothesis. This process helps to draw conclusions about the research question and provides insight into the validity of the initial assumptions made. It may conclude the results were related to the hypothesis and the implications of the findings.

DOES THE INVESTIGATOR OVER-CONCLUDE, THAT IS, ARE THE CONCLUSIONS SUPPORTED BY THE DATA

The investigator does not appear to over-conclude, as the conclusions drawn are supported by the data presented in the study on genetic algorithms for multivariate missing data imputation. The document mentions that the obtained results from the genetic algorithm method passed all statistical tests, including tests on means, individual variances, covariances, correlations, and goodness of fit.

Furthermore, the paper discusses how the results of the genetic algorithm method align with the original populations and fit the distribution of the available complete dataset. The statistical evidence presented in the study supports the conclusion that the proposed genetic algorithm method is appropriate for imputing missing data in multivariate matrices.