Muğla Sıtkı Koçman University

Department of Computer Engineering

Senior Design Project I

**Multiple Imputation Genetic Algorithm (MIGA)**

Analysis & Design Report

Öykü GÖKCE, Murat Avzet ASAN

Supervisor: Zeynep Filiz EREN

January 17, 2024

**Contents**

1 Introduction 3

2 Motivation 3

3 Literature Review 3

4 Method 3

5 Future Work 3

6 References 3

Project Title

# Introduction

Lately, missing data has gained attention in data mining and machine learning. Handling missing values by discarding observations is common but undesirable due to significant data reduction and altered statistical properties. Existing methods mainly focus on single-feature continuous data with strong assumptions.

 This project explores Genetic Algorithms (GAs) to address missing data in multivariate datasets. GAs, known for efficiency, navigate integer/binary spaces, making them suitable for complex problems. The goal is to enhance the effectiveness of handling multiple missing observations in multi-feature data.

# Motivation

Our motivation for this study is rooted in recognizing the limitations and drawbacks of commonly used methods for handling missing data. The traditional approach of removing incomplete observations can lead to losing valuable information and compromise the integrity of subsequent analyses, especially in real-world applications where the impact of missing data is increasingly noticeable. This prompts our interest in exploring new methodologies that directly tackle this challenge.

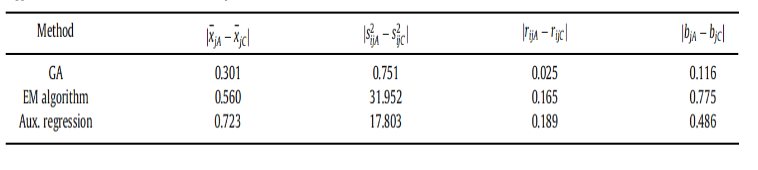
Additionally, existing statistical and artificial intelligence methods designed for single-feature data fall short when applied to the more complex challenges posed by missing data in datasets with multiple features. The complexities, including covariance structures, non-Gaussian variables, and simultaneous presence of multiple missing observations, require a more flexible approach. Hence, our motivation extends to investigating the effectiveness of Genetic Algorithms (GAs), known for their efficiency in navigating complex spaces, as a tailored solution for handling multiple missing observations while preserving essential statistical properties.

In summary, our work is motivated by the need to improve current methodologies for handling missing data in datasets with multiple features, emphasizing the preservation of data integrity to achieve more accurate and robust analyses.

# Literature Review

MIGA has obtained the best results i.e. the smallest differences for all measures and we point out that all methods obtained the biggest difference over covariances at the pair while the biggest difference over correlations were obtained at pairs respectively. Table 1 shows the biggest absolute differences on means, variances, correlations and skewness for all methods.

To summarize, MIGA passed all tests and shows better results (smallest differences) than the EM algorithm and auxiliary regressions with a proper imputation of discrete missing values.



# Method

This project analyzes the data from 73 different CSV files obtained using the "lost" package in R by filling in missing observations. In the initial phase, each file's content is examined, and missing observations are identified. Subsequently, the datasets from all CSV files are merged, resulting in a comprehensive dataset. Missing data analysis is conducted, and algorithms are developed to address the identified gaps. In the final stage, statistical analyses are performed on the extensive dataset obtained after imputation. These analyses aim to evaluate the impact of missing data imputation on the analysis results and enhance the reliability of the obtained outcomes.

# Future Work

Among the future plans for our project is to read the data frame in a different format and duplicate missing data to prevent miscalculations caused by missing values being in a single row. These significant adjustments aim to provide a more effective approach to the issue of missing data. In line with these modifications, updating the calculation functions is also part of our plans. Thus, the goal is to address missing data accurately and enhance the reliability of the analyses.

# References

[1] M. Abdella, T. Marwala, The use of genetic algorithms and neural networks to approximate missing data in database, in: IEEE (Ed.), IEEE 3rd

International Conference on Computational Cybernetics, 2005. ICCC 2005, IEEE. pp. 207–212.

[2] M.A. Aitkin, Some tests for correlation matrices, Biometrika 56 (1969) 443–446, URL:https://www.jstor.org/stable/2334438.

[3] T. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, 1984.

[4] G. Box, A general distribution theory for a class of likelihood criteria, Biometrika 36 (1949) 317–346.

[5] G. Box, Problems in the analysis of growth and wear curves, Biometrics 6 (1950) 362–389.

[6] G. Celeux, J. Diebolt, The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, Computational

Statistics Quarterly 2 (1993) 73–82.

[7] O. Cucconi, Un nuovo test non parametrico per il confront tra due gruppi campionar, Giornale Degli Econmisti Annali di Econmia 27 (1968) 225–248.

[8] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, Journal of Royal Statistical Society 39 (1977)

1–38.

[9] L. Devroye, Non-Uniform Random Variate Generation, Springer-Verlag, New York, 1986.

[10] F. Diebold, R. Mariano, Comparing predictive accuracy, Journal of Business and Economic Statistics 13 (1995) 253–263, https://doi.org/10.1198/

073500102753410444.

[11] J.C. Figueroa, D. Kalenatic, C.A. Lopéz, An evolutionary approach for imputing missing data in time series, Journal Of Circuits, Systems And Computers

19 (2010) 107–121.

[12] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, International Journal of Forecasting 13 (1997) 281–291,

https://doi.org/10.1016/S0169-2070(96)00719-4.

[13] H. Huang, H. Wang, M. Sun, Incomplete data classification with view-based decision tree, Applied Soft Computing 77 (2019) 356–365.

[14] O. Kazík, M. Pilát, R. Neruda, Meta learning in multi-agent systems for data mining, in: IEEE (Ed.), IEEE/WIC/ACM International Conferences on Web

Intelligence and Intelligent Agent Technology, IEEE. pp. 433–434.

[15] A. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione, G. Ist. Ital. Attuari 4 (1933) 83–91.

[16] P. Krishnaia, J. Lee, Likelihood ratio tests for mean vectors and covariance matrices, Handbook of Statistics 1 (1980) 513–570.

[17] S. Kullback, On testing correlation matrices, Applied Statistics 16 (1967) 80–85, URL:ttp://www.jstor.org/stable/2985240.

[18] X. Lai, X. Wua, L. Zhang, Autoencoder-based multi-task learning for imputation and classification of incomplete data, Applied Soft Computing 98

(2021), https://doi.org/10.1016/j.asoc.2020.106838 106838.

[19] A. Law, D. Kelton, Simulation System and Analysis, Mc Graw Hill International, 2000.

[20] L.A. Levine, G. Casella, Implementations of the Monte-Carlo EM algorithm, Journal of Computational Graphic Statistics 10 (2000) 422–439.

[21] T. Liu, H. Wei, K. Zhang, Wind power prediction with missing data using gaussian process regression and multiple imputation, Applied Soft Computing

71 (2018) 905–916.

[22] Q. Long, C. Wu, T. Huang, X. Wang, A genetic algorithm for unconstrained multi-objective optimization, Swarm and Evolutionary Computation 22

(2015) 1–14, https://doi.org/10.1016/j.swevo.2015.01.002.

[23] Mesquita, D.P., ao P.P. Gomes, J., Corona, F., Junior, A.H.S., Nobre, J.S., 2019. Gaussian kernels for incomplete data. Applied Soft Computing 77, 356–365.

[24] Mesquita, D.P., ao P.P. Gomes, J., Junior, A.H.S., Nobre, J.S., 2017. Euclidean distance estimation in incomplete datasets. Neurocomputing 248, 11–18.

doi: 10.1016/j.neucom.2016.12.081.

[25] S.F. Nielsen, The stochastic EM algorithm: Estimation and asymptotic results, Bernoulli 6 (2000) 457–489.

[26] T. Nishino, H. Murakami, The generalized cucconi test statistic for the two-sample problem, Journal of the Korean Statistical Society 48 (2019) 593–

612.

[27] M. Pilát, R. Neruda, Aggregate meta-models for evolutionary multiobjective and many-objective optimization, Neurocomputing 116 (2013) 392–402,

https://doi.org/10.1016/j.neucom.2012.06.043.

[28] R.D. Priya, S. Kuppuswami, A genetic algorithm based approach for imputing missing discrete attribute values in databases, WSEAS Transactions on

Information Science and Applications 9 (2012) 169–178.

[29] A.M. Sefidian, N. Daneshpour, Estimating missing data using novel correlation maximization based methods, Applied Soft Computing 91 (2020)

106249.

[30] N. Smirnov, Table for estimating the goodness of fit of empirical distributions, Annals of Mathematical Statistics 19 (1948) 279–281.

[31] D. Sovilj, E. Eirola, Y. Miche, K.M. Björk, R. Nian, A. Akusok, A. Lendasse, Extreme learning machine for missing data using multiple imputations,

Neurocomputing 174 (2016) 220–231, https://doi.org/10.1016/j.neucom.2015.03.108.

[32] N.H. Timm, Applied Multivariate Analysis, Springer-Verlag, 2002.

[33] C. Wang, H. Pan, Y. Su, A many-objective evolutionary algorithm with diversity-first based environmental selection, Swarm and Evolutionary

Computation 53 (2020), https://doi.org/10.1016/j.swevo.2019.100641 100641.

[34] A. Wójtowicz, P. Zywica, A. Stachowiak, K. Dyczkowski, Solving the problem of incomplete data in medical diagnosis via interval modeling, Applied Soft

Computing 47 (2016) 424–437.