**University of São Paulo
“Luiz de Queiroz” School of Agriculture**

**Archetypal analysis as an imputation method and multivariate data
augmentation**

Pórtia Piscitelli Cavalcanti

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

**Piracicaba
2021**

Pórtia Piscitelli Cavalcanti
BSc in Biotechnology

**Archetypal analysis as an imputation method and multivariate data
augmentation**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:

Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2021

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Cavalcanti, Pórtya Piscitelli

Análise de Arquétipos como método de imputação e aumento de dados multivariados / Pórtya Piscitelli Cavalcanti. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2021 .

49 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Método não supervisionado
2. Dados faltantes
3. Estudo de simulação
4. Estatística multivariada . I. Título.

ACKNOWLEDGEMENTS

I would like to thank everyone who somehow contributed to make this work happen.

I thank to my family who are always by my side supporting and loving me unconditionally. You are very special and I'm very lucky to have you!

To my partner for not letting me give up and for showing me my strength in the last moments. You are an inspiration to me. I am very happy by your side!

To all my friends who contributed to make this journey lighter and more joyful. I highlight the friends that I made in Piracicaba and in New Zealand, you were my second family! I am grateful to have met you!

To my Brazilian advisor for trusting in my potential and for encouraging me to do the exchange. It was a difficult choice and I'm happy that I did it!

To my New Zealand supervisors for the kind reception in the University of Otago, for all the learning and all the support in my internship.

To CAPES and CNPq for material and financial support.

CONTENTS

Resumo	5
Abstract	6
1 Introduction	7
References	8
2 Multivariate sample augmentation via archetypes in correlated normal distributions	11
Abstract	11
2.1 Introduction	11
2.2 Development	12
2.2.1 Archetypal Analysis	12
2.2.2 Algorithms for data augmentation via archetypes	13
2.2.3 Computational evaluation of sample augmentation	15
2.2.4 Results and Discussion	17
2.3 Conclusions	26
References	26
Annexes	28
3 Archetypal Analysis as an imputation method in a multivariate dataset with missing data	37
Abstract	37
3.1 Introduction	37
3.2 Development	38
3.2.1 Archetypal Analysis	38
3.2.2 Imputation methods	39
3.2.3 Simulation study	39
3.2.3.1 Single imputation via AA - SAA	41
3.2.3.2 Multiple imputation via AA - MAA	41
3.2.4 Results and discussion	42
3.3 Conclusions	47
References	47
4 Final Considerations	49

RESUMO

Análise de Arquétipos como método de imputação e aumento de dados multivariados

A estatística multivariada estuda a relação entre um conjunto de variáveis aleatórias e como analisá-las simultaneamente. Na estatística multivariada, os arquétipos são elementos extremos capazes de reescrever todas as observações de uma amostra, ou população, por meio de combinações lineares. Por meio da Análise de Arquétipos (AA), técnica multivariada que visa reduzir a dimensionalidade das observações, é possível encontrar e selecionar seus arquétipos, que são combinações convexas dos dados. A AA pode ser aplicada em várias áreas do conhecimento e com diferentes usos de arquétipos. Nesta tese, foram propostos dois usos diferentes da AA em contextos multivariados: como método de aumento amostral e como método de imputação. A primeira abordagem foi estudada em amostras provenientes de variáveis aleatórias normais bivariadas correlacionadas de diferentes estruturas de covariância, a partir das quais um estudo de simulação foi realizado para avaliar três algoritmos propostos e compará-los com métodos tradicionais. Observou-se que independentemente da estrutura de correlação entre as variáveis é possível aumentar até 20% do tamanho amostral. A segunda abordagem avaliou o uso de arquétipos para imputar valores por imputação Simples e Múltipla em um conjunto de dados multivariados, com dados faltantes simulados. Um estudo de simulação também foi conduzido para avaliar os métodos propostos e estes também foram comparados a métodos tradicionais. Os resultados foram promissores e os valores imputados foram muito semelhantes aos originais. Portanto, nas duas abordagens discutidas nesse trabalho, os resultados apontam para a capacidade dos arquétipos de representar o conjunto de dados e, assim, expressá-los como um novo dado ou preencher de forma satisfatória os possíveis valores ausentes.

Palavras-chave: Método não supervisionado, Dados faltantes, Estudo de simulação, Estatística multivariada

ABSTRACT

Archetypal analysis as an imputation method and multivariate data augmentation

Multivariate statistics studies the relation between a set of random variables and how to analyze them simultaneously. In Multivariate Statistics, archetypes are extreme elements capable of rewriting all observations of a sample, or population, by means of linear combinations. Through the Archetypal Analysis (AA), a multivariate technique that aims to reduce the dimensionality of observations, it is possible to find and select their archetypes, which are convex combinations of the data. AA can be applied in several areas of knowledge and with different uses of archetypes. On this thesis we proposed two different uses of the AA in multivariate contexts: as a sample augmentation method and as an imputation method. The first approach was addressed in samples from bivariate correlated normal random variables from different covariance structures and a simulation study was carried out to evaluate three proposed algorithms and compare them to traditional methods. It was observed that regardless of the correlation structure between the variables, it is possible to increase up to 20% of the sample size. The second approach have evaluated the use of archetypes to impute values by Single and Multiple imputation in a multivariate dataset, with simulated missing data. It was also conducted a simulation study to evaluate the proposed methods that were compared to traditional ones too. The results were promising and the imputed values were very similar to the originals. Therefore, in the two approaches discussed in this work the results points out to the ability of the archetypes representing the dataset and so expressing it as a new data or filling up possible missing values satisfactorily.

Keywords: Unsupervised method, Missing data, Simulation study, Multivariate statistics

1 INTRODUCTION

Multivariate statistics is the field of statistics that studies the simultaneous analysis of a collection of random variables (Manly and Alberto, 2016). Multivariate analysis incorporates the information about the structure of correlation between these variables, which provides more sophisticated analysis and optimized results than the separate univariate analyses of each variable (Ferreira, 2011). On the other hand, multivariate data can be hard to analyze and it requires appropriate tools for its modeling and understanding. According to James et al. (2013), the statistical learning is a framework that can provide such tools, which can be classified as supervised or unsupervised.

The supervised learning aims to predict an outcome measure based on a number of input variables, so it has the output variable to guide or supervise the process of learning (Hastie et al., 2009). In unsupervised learning, there is no output variable and we aim to explore particular characteristics of the input features, like estimating the joint probability density, locating outliers, searching out clusters or imputing missing data, exemplified by Izeman (2008).

Among the unsupervised learning methods, there are known multivariate techniques as Principal Components Analysis (PCA) and Cluster Analysis, for example. The first one aims to reduce the number of dimensions by linear combinations of the variables that results in uncorrelated transformed variables called principal components. Thus, we can explain the data variation by means of a few number of principal components (Manly and Alberto, 2016).

Cluster Analysis seeks to partition observations into subsets or clusters such that those within the same cluster are more closely associated to each other than those allocated to different clusters (Hastie et al., 2009). As stated by James et al. (2013), there exist a great number of clustering methods and one of the best-known approaches is the k-means. K-means clustering aims to group the observations into a pre-specified number of clusters and explains the data through centroids, which are the means of the defined clusters (Moliner and Epifanio, 2019).

Archetypal Analysis (AA), proposed by Cutler and Breiman (1994), is another unsupervised learning technique that lies somewhere in between PCA and k-means (Moliner and Epifanio, 2019). This multivariate method aims to reduce the space of observations of a data set by means of convex combinations of the observations themselves, which results in prototypes called archetypes. Therefore, AA approximates each observation by a convex combination of a collection of archetypes (Hastie et al., 2009).

The original solution to the AA problem presented by Cutler and Breiman (1994) is solved by an alternating constrained least squares algorithm, but Morup and Hansen (2012) demonstrated that a simple projected gradient method can also be used to efficiently compute AA. Besides, these authors showed how useful AA can be for a broad variety of machine learning problems. By summarizing information from all observations in a number of archetypes, AA facilitates the results interpretation (Martins Júnior et al., 2015). In addition, the interpretability of the archetypes is as easy as with clustering techniques (Moliner and Epifanio, 2019).

As the results of AA are convex combinations of the observed data, each archetype can be a mixture of observations or exactly a real extreme observation, also called pure archetype. In situations where artificial constructions makes no practical sense, the Archetypoid Analysis (ADA), introduced by Vinué et al. (2014), allows to select only pure types or archetypoids.

Due to the convex combination the archetypes are forced to lie on the convex hull of the data cloud (Hastie et al., 2009), which turns the results of AA and ADA sensitive to outliers (Cutler and Breiman, 1994). This problem was discussed in Eugster and Leisch (2011) and the authors proposed the Weighted AA and the Robust AA. By means of adaptations in the original algorithm, the Weighted AA enables to insert additional information about the observations - like correlation between them, and the

Robust AA seeks to find the archetypes of the majority of the data set rather than of its totality.

The completeness of the data is assumed in most of the statistical methods including AA and ADA. Notwithstanding, missing data or incomplete data are very common, specially in the multivariate real applications (Epifanio et al., 2020). So, Morup and Hansen (2012) and Epifanio et al. (2018) proposed modifications in the original AA algorithm to deal with missing values, however both approaches did not fulfill the AA theory. Thereafter, Epifanio et al. (2020) came up with modifications of these approaches and new procedures that adapt the AA algorithm by considering the missing values or skipping them such that the AA theory is met.

Another alternative showed by Epifanio et al. (2020) is that AA can be used as an imputation method. This idea meets the one addressed by Cavalcanti (2016) of using archetypes to generate new data. As the archetypes are able to rewrite the sample elements with a minimal error, they can generate elements not observed in the original sample or fill-in missing values by replacing it by an estimate.

This work aims to prospect two different applications of AA in multivariate data. In the second chapter of the thesis, sample augmentation through archetypes is addressed. A simulation study was carried out to evaluate and compare three algorithms for data augmentation via archetypes as to their efficiency. In Chapter 3, we present the use of AA as an imputation method for missing data. The archetypes were used to impute values in a data set with simulated missing data. And in Chapter 4, we make our final considerations.

References

- Cavalcanti, P. P. (2016). Proposta de algoritmos para aumento de dados via arquétipos. Mestrado em estatística aplicada e biometria, Universidade Federal de Alfenas - Unifal-MG, Alfenas, MG.
- Cutler, A. and Breiman, L. (1994). Archetypal Analysis. *Technometrics*, 36(4):338–347.
- Epifanio, I., Ibáñez, M. V., and Simó, A. (2018). Archetypal shapes based on landmarks and extension to handle missing data. *Advances in Data Analysis and Classification*, 12(3):705–735.
- Epifanio, I., Ibáñez, M. V., and Simó, A. (2020). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. *The American Statistician*, 74(2):169–183.
- Eugster, M. J. A. and Leisch, F. (2011). Weighted and robust archetypal analysis. *Computational Statistics & Data Analysis*, 55(3):1215–1225.
- Ferreira, D. F. (2011). *Estatística Multivariada*. Ed. UFLA, Lavras, 2 edition. 676 p.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.
- Izeman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning: with Applications in R*, volume 112. Springer.
- Manly, B. F. and Alberto, J. A. N. (2016). *Multivariate statistical methods: a primer*. Chapman & Hall/CRC press, 4 edition.
- Martins Júnior, J. M., Chagas, E. D. N., Nogueira, D. A., Ferreira, D. F., and Ferreira, E. B. (2015). A Análise de Arquétipos: uma revisão bibliográfica. *Revista Brasileira de Biometria*, 33(2):156–169.

- Moliner, J. and Epifanio, I. (2019). Robust multivariate and functional archetypal analysis with application to financial time series analysis. *Physica A: Statistical Mechanics and its Applications*, 519:195–208.
- Morup, M. and Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63.
- Vinué, G., Epifanio, I., and Alemany, S. (2014). Archetypoids: a new approach to define representative archetypal data. *Computational Statistics & Data Analysis*, 87:102–115.

2 MULTIVARIATE SAMPLE AUGMENTATION VIA ARCHETYPES IN CORRELATED NORMAL DISTRIBUTIONS

Abstract

The aim of this paper was to evaluate multivariate data augmentation through archetypes in samples from bivariate correlated normal random variables. A simulation study was carried out to evaluate the sample augmentation via archetypes in data from different covariance structures using three algorithms: Restricted Drawing of Coefficients (RDC), Unrestricted Standardized Drawing of Coefficients (USDC) and Unrestricted Non-Standardized Drawing of Coefficients (UNSDC). The three algorithms showed good results in the ability to preserve the original probability distribution of the data, the original vector of means and covariance matrices when augments of 10% of the initial sample size was performed, highlighting UNSDC, which allowed augments up to 20%.

Keywords: archetypal analysis; augmented data; multivariate statistics.

2.1 Introduction

An archetype can be defined as a “standard”, “model”, “prototype”, among others definitions; and its concept is used in different areas, such as in Psychology, Philosophy, Literature and even in Statistics. In Multivariate Statistics, a set of archetypes would be extreme elements that has the ability to rewrite the other observations of a sample or population.

Proposed by Cutler and Breiman (1994), the Archetypal Analysis (AA) is a multivariate technique that aims to reduce the observations space of a data set through its archetypes. These are obtained by linear combinations of the observations and are chosen by minimizing the errors committed in the reconstruction of each original data as linear combination of the archetypes.

There are applications of AA in various fields of knowledge such as health applications (Reeve et al., 2017) and environmental quality assessment (Drees et al., 2018). Among the several known applications of the archetypes, since they are able to rewrite the sample elements with a minimal error, it was proposed to use them to augment the sample data (Cavalcanti, 2016), generating elements not observed in the original sample.

When the statistical inference procedure presents problems due to missing data or because the data set contains fewer observations than desired, it may be interesting for the researcher to complete or just increase that sample. This can be done by means of the data augmentation technique, which is based on introducing unobserved data or latent variables through iterative methods or sampling algorithms (Tanner and Wong, 2010).

The data augmentation was introduced with the proposal of a deterministic algorithm, used to calculate estimates of maximum likelihood of incomplete data (Dempster et al., 1977). However, the term “data augmentation” had its origin with the implementation of a stochastic algorithm in a Bayesian context (Tanner and Wong, 1987). Thus, there are applications of augmented data for many purposes, both in frequentist and Bayesian inference.

The efficiency of the data augmentation via archetypes was evaluated by Cavalcanti (2016) for the specific case of independent and normally distributed random variables, so, one must evaluate how the increase of multivariate data happens when we have different correlations between the random variables. Considering the hypothesis that it is possible to increase multivariate data using archetypes regardless the covariance structure among the variables, this paper aims to evaluate how the multivariate sample augmentation using archetypes works when we have correlation structures between the variables.

2.2 Development

This section explains the AA in more details (2.2.1) and contains the description of three algorithms for sample augmentation using archetypes (2.2.2). Then, it also describes how a simulation study was carried out on the augmentation of samples from bivariate random variables correlated with different parameters and sample sizes (2.2.3).

2.2.1 Archetypal Analysis

Considering a collection of p random variables (or inputs) each one observed n times, the matrix ${}_n\mathbf{X}_p$ can represent a multivariate data set, such that $\mathbf{x}_i \in \mathbb{R}^p$ and $i = 1, \dots, n$ is a random vector $\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]$. According to Cutler and Breiman (1994), for a given ${}_n\mathbf{X}_p$, AA aims to find a set of random vectors $\mathbf{z}_k \in {}_K\mathbf{Z}_p$ called archetypes, where $k = 1, \dots, K$ and $K < n$, which are convex combinations of the observations:

$$\mathbf{z}_k = \sum_{i=1}^n \beta_{ik} \mathbf{x}_i \quad (2.1)$$

where $\beta_{ik} \geq 0$ and $\sum_{i=1}^n \beta_{ik} = 1$, characterizing the convex combination (Epifanio et al., 2020). Thus, for a given set of archetypes, AA minimizes

$$\|\mathbf{x}_i - \sum_{k=1}^K \alpha_{ik} \mathbf{z}_k\|^2 \quad (2.2)$$

which also occurs by convex combination, that is, $\alpha_{ik} \geq 0$ and $\sum_{k=1}^K \alpha_{ik} = 1$ (Epifanio et al., 2020). Therefore, in order to select a set of archetypes that best rewrite the original data, one must minimize the Residual Sum of Squares (RSS)

$$RSS = \sum_{i=1}^n \|\mathbf{x}_i - \sum_{k=1}^K \alpha_{ik} \mathbf{z}_k\|^2 = \sum_{i=1}^n \|\mathbf{x}_i - \sum_{k=1}^K \alpha_{ik} \sum_{l=1}^n \beta_{lk} \mathbf{x}_l\|^2. \quad (2.3)$$

The equation 2.3 can be solved by a least squares algorithm which uses alternating iterative procedures (Cutler and Breiman, 1994). The basic principles of this algorithm are based on two foundations: find the best coefficients α_{ik} for a given set of archetypes $\mathbf{z}_k$ and find the best archetypes $\mathbf{z}_k$ for one given set of coefficients α_{ik} (Epifanio et al., 2020). At each step, it is necessary to solve several problems of convex least squares by reducing the RSS successively (Cutler and Breiman, 1994).

Cutler and Breiman (1994) state that the selection of the archetypes occurs minimizing the RSS , that is, in the reconstruction of each observation of the original data as a convex combination of the archetypes, those that provide lower RSS will be chosen. It is important to emphasize that the higher the number of archetypes selected, the smaller will be the RSS , however the reduction of the data dimension will also be smaller. Thus, the researcher needs to choose the ideal number of archetypes (K) for each particular situation, provided that $1 \leq K \leq N$, where N is the number of elements at the boundary of the convex hull (Eugster and Leisch, 2009).

According to Eugster and Leisch (2009) there isn't an specific rule to choose the correct number of archetypes K . Therefore, there are visual techniques that can be used to define an appropriate number. These authors suggest doing a scree plot by running the AA algorithm for different numbers of archetypes and using the "elbow criterion" on the RSS where a "flattening" of the curve indicates the proper value of K .

The result of AA is a data set that contains K archetypes values for each p variable (matrix ${}_K\mathbf{Z}_p$), so it's easy to understand (Epifanio et al., 2020), especially when compared to the results of other

dimensionality reduction methods or grouping techniques. Another advantage of this analysis is to allow a simple classification or grouping of data by considering the coefficients α_{ik} of each observation $\mathbf{x}_i$ as conditional probability of an archetype $\mathbf{z}_k$, that is, $p(\mathbf{x}_i|\mathbf{z}_k)$ indicates which class (archetype $\mathbf{z}_k$) most likely to represent the point.

In recent years, lots of studies keep being carried out applying AA in different areas of knowledge as well as extensions of the technique have been proposed (Moliner and Epifanio, 2019; Epifanio et al., 2020). In this work we'll focus on the 2016 application that uses the archetypes in sample augmentation (Cavalcanti, 2016), as the archetypes are able to rewrite the sample elements with a minimal error, the proposed algorithms, explained in next section, allow us to generate elements not observed in the original sample.

2.2.2 Algorithms for data augmentation via archetypes

Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be a p -variate random sample and $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_a$ be a p -variate latent vectors computed from the archetypes in order to produce the augmented sample $\mathbf{X}_1, \dots, \mathbf{X}_n, \mathbf{W}_{n+1}, \dots, \mathbf{W}_{n+a}$, where a is the number of augmented data. Also let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ be the values of $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_a$ the computed values for $\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_a$. First, perform AA on the sample data to find a set of random vectors $\mathbf{z}_k$ containing the K better sample archetypes to represent it (equations 2.1 to 2.3). Using one of the algorithms explained below, select archetypes $\mathbf{z}_k$ and coefficients α_{*k} , where $* \in \{1, 2, \dots, n\}$. Then, compute the latent observations $\mathbf{w}_j$ ($j = 1, \dots, a$) as a combination of the drawn coefficients and archetypes

$$\mathbf{w}_j = \sum_{k=1}^K \alpha_{*k} \mathbf{z}_k. \quad (2.4)$$

Algorithms 1 and 2 draw coefficients considering the requirement of convex combination, and thus have different constraints to ensure that the sum of the coefficients is one. On the other hand, algorithm 3 does not meet this requirement, being a simpler and computationally lighter alternative. The algorithms are detailed as follows.

Algorithm 1 – Restricted Drawing of Coefficients (RDC)

The first algorithm consists of the following steps, illustrated by Figure 2.1:

- (i) Draw one of the archetypes $\mathbf{z}_k$ without replacement, where $k = 1, \dots, K$;
- (ii) Draw one coefficient α_{ik} related to the drawn archetype, say α_{*k} ;
- (iii) Compute the difference $d = 1 - \sum_{k=1}^K \alpha_{*k}$;
- (iv) Stop if you have drawn the last archetype or $d = 0$. In the latter case set to zero the other coefficients. Else, draw another archetype among the remaining;
- (v) Draw a coefficient related to the new archetype, since it is not greater than d ;
- (vi) Repeat items (iii) to (v) until stop decision (iv) is made.
- (vii) Multiply the coefficients drawn by the respective archetypes to find the unobserved data $\mathbf{w}_j$.

Figure 2.1: Illustration of the RDC steps.
Source: figure elaborated by the author.

Algorithm 2 – Unrestricted Standardized Drawing of Coefficients (USDC)

The second algorithm is based on the following steps, illustrated by Figure 2.2:

- (i) Draw a coefficient for each archetype in the respective set of coefficients: $\alpha_{*1}, \alpha_{*2}, \dots, \alpha_{*K}$;
- (ii) Sum the coefficients up $s = \sum_{k=1}^K \alpha_{*k}$;
- (iii) Standardize them dividing by s .
- (iv) Multiply the coefficients drawn by the respective archetypes to find the unobserved data w_j .

Figure 2.2: Illustration of the USDC steps.
Source: figure elaborated by the author.

Algorithm 3 – Unrestricted Non Standardized Drawing of Coefficients (UNSDC)

The third algorithm simply draw a coefficient for each archetype and multiply the coefficients drawn by the respective archetypes, that is, only steps (i) and (iv) of Algorithm 2, as can be seen in Figure 2.3.

Figure 2.3: Illustration of the UNSDC steps.

Source: figure elaborated by the author.

2.2.3 Computational evaluation of sample augmentation

In order to compare and evaluate the efficiency of each algorithm in the sample augmentation, the following simulation study was performed. Considering a correlated bivariate normal probability distribution with known parameters $X_1, X_2 \sim N_2(\mu, \Sigma)$. First the AA was applied in order to select a set of K archetypes z and then a new elements w were obtained with the three algorithms. In addition to the comparison of the algorithms each other, they were compared to a gold standard and a control procedure. These methods were used as references and consist of drawing sample elements as follows:

Gold Standard (GS): draws elements randomly by knowing the probability distribution of the random variable and the true value of its parameters;

Control (C): inspired by the non-parametric Bootstrap method, randomly draws only elements present in the sample with replacement.

At the end of the study, augmented data were obtained by each algorithm and reference methods, composed by the initial sample plus a certain number of increased data, proportional to the sample size. The initial samples had two different sample sizes ($n = 30$ and 100 observations) and the augmented proportions were $AP = 10\%$, 20% and 30% of n .

In order to represent different possibilities of correlation between the random variables we considered four structures of covariance matrix: variance component, compound symmetry, first order autoregressive and unstructured. As each structure contains specific variances and correlations, the scenarios of the simulation study were explained in more details below according to each covariance matrix structure.

Variance Component - VC:

$$\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

The VC structure is a diagonal matrix that represents independent observations and equal variances (σ^2), so it was considered two different values of variance ($\sigma^2 = 10$ and 200). Thus, the scenarios of the VC structure were composed by two sample sizes, three augment proportions and two values of variance ($2 \times 3 \times 2 = 12$), 12 scenarios in total.

First-order Autoregressive - AR(1):

$$\sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

In the AR(1) structure the covariance between two observations decreases as the time interval increases between them, but at all times the variances (σ^2) are equal. ρ is the autoregressive parameter and represents the correlation between the variables. We've considered six correlations ($\rho = 0.2; 0.5; 0.9; -0.2; -0.5$ and -0.9) and in all cases we've used $\sigma^2 = 1$. Therefore, the AR(1) structure scenarios were composed of two sample sizes, three proportions of increase and six correlations ($2 \times 3 \times 6 = 36$), giving a total of 36 scenarios.

Compound Symmetry - CS:

$$\begin{bmatrix} (\sigma^2 + \sigma_1^2) & \sigma_1^2 \\ \sigma_1^2 & (\sigma^2 + \sigma_1^2) \end{bmatrix}$$

This structure has equality of variances ($[\sigma^2 + \sigma_1^2]$) and constant covariance (σ_1^2) between any observations of the same unit due to independent errors. So, it requires information on the variance and on the correlation, in order to calculate the covariance. We've considered two values for the variances ($[\sigma^2 + \sigma_1^2] = 1$ and 200) and six values for the correlation ($\rho = 0.2; 0.5; 0.9; -0.2; -0.5$ and -0.9). Thus, the CS structure scenarios were composed of two sample sizes, three augment proportions, two variances and six correlations ($2 \times 3 \times 2 \times 6 = 72$), 72 scenarios in total.

Unstructured - UN:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

The UN covariance matrix allows unequal values between variances and covariance, so two values were considered for the variances ($\sigma_1^2 = 1$ and $\sigma_2^2 = 200$) and six values of correlations ($\rho = 0.2; 0.5; 0.9; -0.2; -0.5$ and -0.9). Thus, the scenarios of the UN structure were composed of two sample sizes, six correlations and three augment proportions ($2 \times 6 \times 3 = 36$), giving a total of 36 scenarios.

At the end we add up to 156 scenarios (12 VC + 36 AR(1) + 72 CS + 36 UN), each one simulated with 1000 Monte Carlo repetitions. After the sample augmentation using the algorithms, GS and C were used specific tests to check if the probability distribution of the random variable, its mean vector and covariance structure were the same in each scenario.

To assess multivariate normality, the Shapiro-Wilk test generalized by Royston (1983) was used, pointed out as one of the most widely used and as rigorous (Korkmaz et al., 2014); to evaluate the mean vector, the Hotelling's T^2 test (Kariya, 1981) was chosen, which is an extension of the univariate test and is also already consolidated in the literature (Hair Jr. et al., 2005). And to evaluate the covariance matrix, the covariance matrix equality test of Ledoit and Wolf (2002) was used, which proved to be robust even against high dimensionality. All the tests considered a level of 5% significance.

All algorithms have been programmed in R language and all analysis were executed in R software (R Core Team, 2021). Only the samples that presented the probability distribution and the established population parameters were used in the simulations, according to the tests performed at a 5% significance level. For AA, the *archetypes* package (Eugster and Leisch, 2009) was used and therefore executed through ready functions that weren't modified. The definition of the number of archetypes (na) used was done previously based on the scree plot for each initial sample and for all cases were used three archetypes ($na = 3$).

2.2.4 Results and Discussion

The results discussed in this section are related to the rejection rate (%) of the null hypothesis of each test performed in the augmented samples: Royston test (normality), Hotelling's T^2 test (mean vector) and matrix covariances equality test of Ledoit and Wolf. As the 5% level of significance was adopted, were considered acceptable the results below the upper limit of the exact binomial confidence interval $IC_{99\%}(\pi)$: [3.39%, 7.05%], where π is the population proportion parameter.

All the results obtained with GS, with the three algorithms (RDC, USDC and UNSDC) and with C were organized into tables and can be consulted in the Annexes section (Tables 2.1 to 2.9) organized according to the covariance matrix structure. In all of them GS presented very low rejection rates for the null hypotheses of the three tests, regardless the covariance matrix structure or the parameters evaluated, confirming its role in the simulation study.

Thus, we proceeded with the comparison on the results of the algorithms and C that were discussed according to the covariance matrix structure evaluated. Bar graphs were made to compare its rejection rate (%) of the null hypothesis of each test in the main scenarios and the sample augmentation was also graphically exemplified.

Variance Component

Starting with the VC structure, the sample augmentation in the scenario where $n = 30$ and $\sigma^2 = 10$ stood out. So its results were exploited in more details and can be seen in the Figure 2.4 below.

Figure 2.4: Rejection rate (%) of the null hypothesis of Royston's test (Normality), Hotelling's T^2 test (Means) and Ledoit and Wolf's test (Covariances), after sample augmentation of 10, 20 and 30% of n (augmented proportion) with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C) in scenarios where $n = 30$ and $\sigma^2 = 10$. Dashed lines represent the upper limit of the $IC_{99\%}(\pi)$, that is 7.05%.

Source: figure elaborated by the author.

Assessing the multivariate probability distribution, it was noticed that the three algorithms showed similar results and obtained better performance than C. USDC and UNSDC had a Royston test null hypothesis rejection ratio of less than 7.05% in all scenarios and C, on the other hand, allowed to increase 10% of n without interfering in the probability distribution only in scenarios where $n = 30$ (Figure 2.4).

In the analysis of the mean vectors, it was noticed that the three algorithms allowed to increase the initial sample by 10% of n without changing the value of the population mean vector and UNSDC still allowed to increase 20% of n (Figure 2.4). Figure 2.4 also shows that C allowed to increase up to 20% of n , but in the scenarios where $n = 100$ it's possible to augment just 10% of the sample size.

Finally, in assessing the equality of the covariance matrix of the augmented sample to the population matrix, a similar performance was observed for the three algorithms and C, allowing the sample to be increased by up to 20% of its initial size (Figure 2.4). However in scenarios where $n = 100$ the augment needs to be of 10% to obtain the same result.

In general, it can be said that the algorithms performed better when $n = 30$ and $\sigma^2 = 10$, making it possible to increase the initial sample by up to 20% with UNSDC without changing the distribution of probability, nor the estimates of its parameters, which was exemplified in Figures 2.5 and 2.6. With the other algorithms, as well as with the UNSDC, it is possible to increase the size of the original sample by 10% in all the scenarios evaluated, maintaining the population characteristics.

Figure 2.5: Illustration of augmenting the initial bivariate data X_1, X_2 with RDC, USDC and UNSDC in the scenario where $n = 30$, $\sigma^2 = 10$ and $AP = 20\%$ (six augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

In Figure 2.5, it is noted that the RDC and USDC algorithms generate new data only within the convex polygon (triangle) formed by their archetypes, due to the restriction on the convex combination, whereas UNSDC allows data to be generated outside the polygon.

Figure 2.6: Illustration of augmenting the initial bivariate data X_1, X_2 with GS and C in the scenario where $n = 30$, $\sigma^2 = 10$ and $AP = 20\%$ (six augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

In Figure 2.6 it can be seen that GS generates data totally independent of the archetypes and that C draws only elements present in the original sample - in this scenario, it selected elements within the triangle or in the surroundings.

First-order Autoregressive

For the AR(1) structure we will exemplify the sample augmentation with the scenario where $n = 100$ and $\rho = 0.9$ in order to show results of a bigger sample size and high correlation between

the random variables (Figure 2.7), but in advance we can say that the algorithms performed better in scenarios where $n = 30$.

Evaluating the probability distribution, it can be noted in Figure 2.7 that the USDC and UNSDC algorithms maintained the multivariate normal probability distribution after increases of 30% of the sample size, which happened in all scenarios, while RDC allowed increases of 20% and up to 30% in some scenarios where the correlations were low and/or negative. C, on the other hand, allowed increases of only 10% of n without interfering with the probability distribution in almost all scenarios.

Figure 2.7: Rejection rate (%) of the null hypothesis of Royston’s test (Normality), Hotelling’s T^2 test (Means) and Ledoit and Wolf’s test (Covariances), after sample augmentation of 10, 20 and 30% of n (augmented proportion) with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C) in scenarios where $n = 100$ and $\rho = 0.9$. Dashed lines represent the upper limit of the $IC_{99\%}(\pi)$, that is 7.05%.

Source: figure elaborated by the author.

When analyzing the mean vectors, we noted that the three algorithms allowed to augment the initial sample by 10% without changing the estimate of the population mean vector, regardless the sample size or the correlation between variables (Figure 2.7). Meanwhile, it was possible to increase the initial sample using C by up to 20% of its size in most of the scenarios and 10% in the other scenarios without a clear relationship with the correlations or sample size.

When verifying the results of the covariance matrix equality test, we noticed that the proportion of rejection of the null hypothesis that the covariance matrix was equal to the population matrix was less than 7.05% when augments of up to 20% of n were made. When $n = 30$, UNSDC allowed an increase of 20% in all scenarios, USDC and RDC allowed increases of 20% in scenarios where the correlations were low or intermediate and C allowed an increase of 20% in half of the scenarios. In the scenarios where $n = 100$ it was observed that it was possible to increase 10% of the sample size without changing the population covariance matrix with all the algorithms and with C as showed in Figure 2.7.

Therefore, the three algorithms performed better in scenarios where $n = 30$ and the correlations between the variables were lower. Even so, in all evaluated scenarios it was possible to augment the size of the initial sample by 10% with all the algorithms, without changing the characteristics of the population as we can see in Figures 2.8 and 2.9 below that are illustrating the data augmentation of the scenario with $n = 100$, $\rho = 0.9$ and $AP = 10\%$.

In Figure 2.8 it can be observed once again that the algorithms generate data according to the archetypes, with RDC and USDC only allowing data inside and at the boundary of the convex polygon formed by the archetypes and that UNSDC allows generating data outside (only one in this case).

Figure 2.8: Illustration of augmenting the initial bivariate data X_1, X_2 with RDC, USDC and UNSDC in the scenario where $n = 100$, $\rho = 0.9$ and $AP = 10\%$ (ten augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

And finally, in Figure 2.9 it becomes evident again that GS generates data regardless of archetypes and that C draw only elements present in the initial sample - in this case in a dispersed way.

Figure 2.9: Illustration of augmenting the initial bivariate data X_1, X_2 with GS and C in the scenario where $n = 100$, $\rho = 0.9$ and $AP = 10\%$ (ten augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

Compound Symmetry

Assessing the sample augmentation where the data covariance matrix has the CS structure, the scenario where $n = 30$, $(\sigma^2 + \sigma_1^2) = 200$ and $\rho = 0.2$ presented promising results shown in Figure 2.10 next page.

Starting with the probability distribution of the random variables, as we can see in Figure 2.10, USDC and UNSDC maintained the multivariate normal probability distribution after increases of 30% of the sample size, which happened in all scenarios, RDC allowed increases of 30% of n , which happened in almost all scenarios, specially where the correlation between the variables was low and/or negative and

C, on the other hand, allowed to augment 10% of n , which happened in almost all scenarios, specially where the correlation between the variables was high and/or positive.

Figure 2.10: Rejection rate (%) of the null hypothesis of Royston's test (Normality), Hotelling's T^2 test (Means) and Ledoit and Wolf's test (Covariances), after sample augmentation of 10, 20 and 30% of n (augmented proportion) with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C) in scenarios where $n = 30$, $(\sigma^2 + \sigma_1^2) = 200$ and $\rho = 0.2$. Dashed lines represent the upper limit of the $IC_{99\%}(\pi)$, that is 7.05%.

Source: figure elaborated by the author.

In the evaluation of the mean vectors, UNSDC, USDC and C presented rejection proportion of the null hypothesis of Hotelling's T^2 test below 7.05% when the augment proportion was 10% of n and sometimes when it was of 20%, whereas the RDC allowed to augment 10% of n in virtually all scenarios (Figure 2.10). The three algorithms obtained better performance when $0.2 \leq \rho \leq 0.5$.

In the verification of the covariance matrix test, it was observed that it was possible to make increases up to 20% of the sample size without changing the matrix of population covariance with the three algorithms and C when $n = 30$, as showed if Figure 2.10, and increases of 10% when $n = 100$. UNSDC was the only one that allowed these augments in all the scenarios.

Comparing all scenarios whose covariance matrix structure was the CS it can be said, in general, that the algorithms performed better when $n = 30$ and with lower and/or negative correlations between the variables. With all the algorithms, it was possible to increase the size of the original sample by 10% in all the scenarios evaluated, maintaining the population characteristics. UNSDC stands out for allowing to increase the sample without interfering in the multivariate probability distribution in all scenarios and for allowing a proportion of increase of up to 20% maintaining the estimates of the parameters μ and Σ of the initial sample.

In the next page we can see an illustration of the sample augmentation with the three algorithms (Figure 2.11) and with the reference methods (Figure 2.12). In this scenario, it was observed that the three algorithms generated data within the triangle formed by the three archetypes (Figure 2.11). In addition, it is important to note that one of the data generated by USDC was exactly one of the archetypes (the closest to the X_2 axis), so it is not visible in the graph.

Figure 2.11: Illustration of augmenting the initial bivariate data X_1, X_2 with RDC, USDC and UNSDC in the scenario where $n = 30$, $(\sigma^2 + \sigma_1^2) = 200$, $\rho = 0.2$ and $AP = 20\%$ (six augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

Figure 2.12 again shows that GS generates data totally independent of the archetypes and that C draws only elements present in the original sample - in this case, the majority drawn was found outside the polygon.

Figure 2.12: Illustration of augmenting the initial bivariate data X_1, X_2 with GS and C in the scenario where $n = 30$, $(\sigma^2 + \sigma_1^2) = 200$, $\rho = 0.2$ and $AP = 20\%$ (six augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

Unstructured

For the UN covariance matrix we will explore the scenario where $n = 30$ and $\rho = -0.5$. This scenario stood out and allows us to exemplify the sample augmentation where the correlation between variables is negative (Figure 2.13).

Assessing the multivariate probability distribution, in Figure 2.13 we can see that UNSDC and USDC maintained the multivariate normal probability distribution after increases of 30% of the sample size, which happened in virtually all scenarios, while RDC allowed increases of up to 30% in the scenarios

where correlations were low and/or negative. C, on the other hand, allowed increases of 10% of n without interfering with the probability distribution in almost all scenarios.

Figure 2.13: Rejection rate (%) of the null hypothesis of Royston's test (Normality), Hotelling's T^2 test (Means) and Ledoit and Wolf's test (Covariances), after sample augmentation of 10, 20 and 30% of n (augmented proportion) with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C) in scenarios where $n = 30$ and $\rho = -0.5$. Dashed lines represent the upper limit of the $IC_{99\%}(\pi)$, that is 7.05%.

Source: figure elaborated by the author.

When analyzing the mean vectors, we noted that the three algorithms allowed to increase the initial sample by 10% without changing the estimate of the population mean vector and UNSDC also allowed to increase 20% of n in the highlighted scenario (Figure 2.13). Meanwhile, it was possible to augment the initial sample by up to 20% of its size with C in the majority of the scenarios.

Evaluating the results of the covariance matrix equality test, one can see that the proportion of rejection of the null hypothesis that the covariance matrix was equal to the population matrix was less than 7.05% when augments of up to 20% of n were made in scenarios where $n = 30$ (Figure 2.13). UNSDC allowed an increase of 20% in all scenarios, USDC and RDC were limited to the scenarios where the correlations were low or intermediate and C allowed an increase of 20% in almost all scenarios. In the scenarios where $n = 100$, it was observed that it was possible to increase 10% of the sample size without changing the population covariance matrix with all the algorithms and with C.

Thus, again the algorithms had a better performance where $n = 30$ and with lower correlations between variables. Nevertheless, it was also possible to augment the sample size by 10% in all scenarios, without changing the characteristics of the population. In Figures 2.14 and 2.15 below one can see the illustration of the data augmentation in the scenario where $n = 30$, $\rho = -0.5$ and $AP = 20\%$.

Figure 2.14: Illustration of augmenting the initial bivariate data X_1, X_2 with RDC, USDC and UNSDC in the scenario where $n = 30$, $\rho = -0.5$ and $AP = 20\%$ (six augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

In Figure 2.14 it can be observed once again that the algorithms generate data according to the archetypes. RDC and USDC generated all data inside and at the boundary of the convex polygon formed by the archetypes while UNSDC generated one data outside the polygon. In addition, it is important to highlight that UNSDC generated two data very similar so they are overlapped on the graph (the black dot closest to the X_2 axis).

Figure 2.15: Illustration of augmenting the initial bivariate data X_1, X_2 with GS and C in the scenario where $n = 30$, $\rho = -0.5$ and $AP = 20\%$ (six augmented data). Empty dots represent the sample observations, black dots represent the augmented data and red dots joined by dashed lines represent the archetypes.

Source: figure elaborated by the author.

In Figure 2.15 it's clear one more time that GS generates data totally independent of the archetypes and that C draws only elements present in the original sample.

The results of the simulation study have indicated that the three proposed algorithms presented similar performance. Just as concluded by Cavalcanti (2016) for the standard multivariate normal distribution situation, we can also augment bivariate correlated data in 10% with RDC, USDC and UNSDC in all scenarios. UNSDC was slightly better allowing increases of 30% of the sample size without inter-

fering in the multivariate normal distribution and up to 20% of augment without changing the estimates of population parameters. In addition, it should be noted that the performance of the algorithms was better when the initial sample was smaller ($n = 30$) and the correlations between the random variables were lower and negative.

It was also observed that sometimes one of the data augmented by USDC was exactly an archetype, which happened regardless the parameters evaluated. As USDC is the algorithm that performs a standardized drawing of the coefficients that multiply the archetypes, it's understandable that this can often happen. This fact can be a problem if the data set has extreme observations or outliers, so we don't recommend using USDC in this cases. On the other hand, UNSDC proved to be competent in all scenarios and stands out for being the simplest and lightest algorithm as it performs an unrestricted and non standardized draw.

Comparing archetypes based sample augmentation to the reference methods, it confirms its superiority in the maintenance of the probability distribution when compared to C (which was inspired in nonparametric bootstrap) and has the advantage to require no strong assumption when compared to GS (which was inspired in parametric bootstrap or jackknife) (Cavalcanti, 2016).

Therefore, this study have indicated that it is possible to augment multivariate data using archetypes regardless the covariance structure among the variables. We recommend augmenting the initial sample size up to 10% with any of the algorithms or up to 20% with UNSDC, in situations where the initial sample are smaller ($n = 30$) and the correlations between the random variables are lower and negative, in order to keep unaltered simultaneously the probability distribution and its parameters.

2.3 Conclusions

Multivariate sample augmentation via archetypes is effective regardless the covariance structure among the variables, allowing to augment the original sample by up to 20% with no interference in the probability distribution or its parameters.

Future works should investigate the sample augmentation on more than two variables, non normal samples and mixtures of probability distributions.

References

- Cavalcanti, P. P. (2016). Proposta de algoritmos para aumento de dados via arquétipos. Mestrado em estatística aplicada e biometria, Universidade Federal de Alfenas - Unifal-MG, Alfenas, MG.
- Cutler, A. and Breiman, L. (1994). Archetypal Analysis. *Technometrics*, 36(4):338–347.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.
- Drees, L., Roscher, R., and Wenzel, S. (2018). Archetypal analysis for sparse representation-based hyperspectral sub-pixel quantification. *Photogrammetric Engineering & Remote Sensing*, 84(5):279–286.
- Epifanio, I., Ibáñez, M. V., and Simó, A. (2020). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. *The American Statistician*, 74(2):169–183.
- Eugster, M. J. A. and Leisch, F. (2009). From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software*, 30(8):1–23.
- Hair Jr., J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (2005). *Análise Multivariada de Dados*. Bookman, Porto Alegre, 5 edition.

- Kariya, T. (1981). A robustness property of Hotelling's T^2 -test. *The Annals of Statistics*, 9:211–214.
- Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014). Mvn: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162.
- Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, 30(4):1081–1102.
- Moliner, J. and Epifanio, I. (2019). Robust multivariate and functional archetypal analysis with application to financial time series analysis. *Physica A: Statistical Mechanics and its Applications*, 519:195–208.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Reeve, J., Böhmig, G. A., Eskandary, F., Einecke, G., Lefaucheur, C., Loupy, A., Halloran, P. F., study group, M.-K., et al. (2017). Assessing rejection-related disease in kidney transplant biopsies based on archetypal analysis of molecular phenotypes. *JCI insight*, 2(12).
- Royston, J. P. (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society*, 32(2):121–133.
- Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association*, 82(398):528–550.
- Tanner, M. A. and Wong, W. H. (2010). From EM to Data Augmentation: The Emergence of MCMC Bayesian Computation in the 1980s. *Statistical Science*, 25(4):506–516.

Variance Component

Table 2.1: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables.

n	σ^2	AP	Normality						Means						Covariances					
			GS	RDC	USDC	USDC	C	GS	RDC	USDC	USDC	C	GS	RDC	USDC	USDC	C			
30	10	10%	1.1	2.4	2.4	1.4	7.0	1.5	3.4	3.5	3.7	3.5	1.2	2.5	2.4	1.3	2.4			
		20%	1.6	3.9	3.4	1.1	14.1	2.0	7.5	7.3	6.9	5.8	2.3	6.4	4.9	4.9	6.2			
		30%	2.6	5.9	4.4	2.2	23.7	3.5	14.2	11.6	12.9	11.2	1.9	9.1	8.2	7.7	10.5			
	200	10%	1.4	1.9	1.6	0.8	5.2	2.0	3.9	4.1	3.2	3.5	1.9	2.7	2.5	2.1	3.2			
		20%	1.9	3.8	3.5	1.2	13.9	1.5	9.0	7.6	8.0	6.3	1.8	6.0	6.1	4.3	5.4			
		30%	3.0	6.3	3.7	2.1	25.6	3.1	15.6	13.7	14.6	11.9	2.5	8.0	7.8	7.7	10.2			
100	10	10%	0.7	1.9	1.6	0.8	7.7	1.1	5.1	4.3	3.3	4.9	1.7	4.1	3.6	2.9	4.4			
		20%	2.1	3.7	3.6	1.2	16.8	1.7	11.5	8.0	9.5	7.3	1.9	9.1	7.4	6.9	8.3			
		30%	2.3	7.4	4.6	1.9	26.0	2.4	20.7	16.0	16.2	8.3	2.4	16.3	14.3	16.7	10.6			
	200	10%	1.9	2.5	2.1	1.0	7.8	1.8	5.4	4.8	4.1	4.3	1.2	3.0	2.6	2.6	3.2			
		20%	2.8	4.4	3.7	1.6	16.1	2.1	12.7	9.7	10.3	7.5	2.4	9.7	7.7	9.2	7.7			
		30%	3.3	6.2	4.5	1.9	27.8	2.0	20.7	14.6	14.1	8.6	3.1	14.4	12.6	15.4	11.9			

n : number of observations, σ^2 : variance of the variables, AP: augment proportion.

Source: table elaborated by the author.

First-order Autoregressive

Table 2.2: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables and $n = 30$ observations.

ρ	AP	Normality						Means						Covariances					
		GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C			
- 0.9	10%	0.7	1.4	1.3	0.7	2.6	1.1	3.5	3.2	3.0	3.7	1.3	3.0	3.5	3.2	3.6			
	20%	0.7	2.3	2.3	1.9	7.7	2.4	10.9	8.6	8.1	6.7	1.4	7.5	7.9	5.9	6.1			
	30%	1.3	4.7	3.0	1.3	11.6	2.2	16.1	13.3	13.2	9.3	2.7	10.9	11.8	9.7	11.6			
- 0.5	10%	0.9	1.6	1.9	1.3	4.9	2.3	4.2	3.4	4.2	4.2	1.6	2.6	3.7	2.1	2.6			
	20%	2.2	4.6	3.3	2.1	13.9	2.5	8.1	7.6	7.1	7.0	1.6	5.8	6.0	4.4	7.1			
	30%	2.0	5.3	4.4	3.4	21.7	2.6	14.9	13.3	12.4	11.4	1.8	9.0	8.6	7.7	9.3			
- 0.2	10%	1.6	1.9	2.2	1.7	7.3	1.3	4.5	4.0	3.6	4.3	0.5	3.0	3.0	2.2	2.9			
	20%	2.6	4.0	3.2	1.5	15.7	2.6	9.0	7.8	8.8	7.0	2.1	4.8	6.4	5.5	8.0			
	30%	1.7	5.4	4.4	1.6	24.2	2.9	14.0	12.4	11.6	8.6	2.7	9.1	9.1	7.3	9.9			
0.2	10%	1.4	2.7	1.9	0.9	5.9	1.7	4.4	3.5	3.9	3.5	1.3	2.9	2.8	2.1	3.3			
	20%	1.6	3.3	2.4	1.2	13.1	1.9	9.0	7.1	8.9	6.4	1.8	5.1	4.6	3.6	6.3			
	30%	2.8	6.5	4.3	2.5	23.1	2.8	12.5	11.0	11.9	10.2	2.9	10.4	8.9	8.3	10.1			
0.5	10%	0.9	1.4	2.0	1.2	4.8	1.3	3.3	3.2	3.3	3.0	1.3	3.0	2.8	1.8	2.5			
	20%	2.6	4.0	3.2	1.7	13.9	1.4	7.9	7.1	8.0	5.9	1.5	6.0	7.0	5.6	5.2			
	30%	2.8	5.8	5.1	1.7	22.5	2.6	14.1	11.7	12.2	9.7	2.4	10.5	8.9	7.4	9.8			
0.9	10%	1.3	1.3	1.4	1.0	4.1	1.3	4.4	3.9	4.9	3.4	1.4	3.3	3.8	1.9	3.1			
	20%	1.5	4.6	3.9	1.6	11.0	3.1	9.9	9.8	11.4	7.3	2.1	7.1	6.6	5.4	7.3			
	30%	2.8	8.2	5.5	3.5	20.5	3.4	16.1	15.6	15.3	11.0	1.6	12.0	12.4	10.6	10.3			

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

Table 2.3: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables and $n = 100$ observations.

ρ	AP	Normality						Means						Covariances					
		GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C			
- 0.9	10%	0.9	1.5	1.3	0.9	3.3	1.7	5.0	5.0	4.8	3.9	1.8	4.6	4.8	3.7	3.6			
	20%	1.7	3.7	3.5	1.5	9.3	1.8	15.0	13.1	12.2	6.5	1.8	11.8	11.4	11.3	7.8			
	30%	1.1	3.8	3.9	1.9	14.5	2.1	27.8	22.6	22.9	8.2	2.2	19.9	19.2	19.8	11.4			
- 0.5	10%	1.1	2.2	1.8	0.5	6.1	1.6	5.9	6.0	4.9	4.8	0.8	3.9	3.3	2.6	3.2			
	20%	2.0	4.8	3.7	1.1	14.1	1.9	11.8	9.4	9.6	6.7	2.9	9.0	8.5	8.2	8.6			
	30%	1.4	8.6	4.7	1.5	24.1	3.1	20.6	15.8	16.6	9.7	2.7	16.5	13.7	15.3	12.7			
- 0.2	10%	1.8	2.2	1.9	0.6	6.3	1.2	5.1	4.6	4.7	3.2	1.3	3.3	2.7	2.3	2.5			
	20%	3.4	4.0	2.5	1.0	15.5	2.7	13.0	9.9	9.8	7.4	1.7	8.9	8.4	8.0	7.4			
	30%	3.3	7.5	5.7	1.2	26.0	2.2	19.7	16.5	16.3	9.1	2.6	16.0	15.1	16.9	10.3			
0.2	10%	1.5	2.7	1.9	1.0	7.6	1.6	4.6	4.7	4.8	4.3	1.0	4.2	3.4	2.6	3.6			
	20%	2.1	4.4	3.3	0.9	16.3	2.6	12.2	10.7	11.1	8.4	1.2	8.4	8.4	8.5	8.0			
	30%	3.0	6.8	3.9	1.6	27.5	2.6	21.6	16.2	15.6	11.2	2.1	17.2	14.4	16.5	12.1			
0.5	10%	2.3	2.8	2.5	1.2	7.1	1.5	5.1	4.1	3.9	3.2	1.2	4.5	2.8	2.7	3.7			
	20%	2.6	5.9	4.6	1.5	17.0	1.6	11.3	8.8	9.2	6.9	2.4	11.9	10.0	10.0	9.6			
	30%	3.2	9.8	6.2	1.9	27.5	2.5	23.4	17.7	16.9	8.6	2.9	16.2	14.2	18.3	11.5			
0.9	10%	1.2	3.1	2.3	1.3	6.4	1.1	6.6	6.1	5.5	4.7	3.1	6.5	6.0	4.4	5.2			
	20%	1.9	5.5	4.4	1.6	11.1	2.3	15.0	12.6	13.2	6.9	2.0	13.9	10.1	10.7	6.9			
	30%	2.1	8.9	5.5	2.3	19.9	2.5	25.4	21.0	20.9	7.8	2.1	19.0	17.7	17.0	12.4			

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

Compound Symmetry

Table 2.4: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables, $n = 30$ observations and variance ($\sigma^2 + \sigma_1^2$) = 1.

ρ	AP	Normality						Means						Covariances					
		GS	RDC	USDC	USDC	C	GS	RDC	USDC	USDC	C	GS	RDC	USDC	USDC	C			
- 0.9	10%	1.2	1.4	1.2	0.7	2.5	0.9	3.9	4.4	5.2	2.9	1.3	4.3	3.2	2.9	4.6			
	20%	0.9	3.0	2.2	0.8	7.9	2.4	7.9	7.7	7.5	6.9	2.0	6.0	6.4	5.4	6.2			
	30%	1.2	3.5	3.9	1.0	14.0	2.4	13.1	13.7	12.5	8.1	3.0	12.0	12.0	10.6	10.1			
- 0.5	10%	0.9	2.2	1.5	1.2	5.1	2.7	4.4	4.7	5.2	3.8	1.1	2.7	3.0	2.2	3.5			
	20%	1.6	4.1	3.1	1.6	13.0	2.3	9.8	9.6	9.7	6.8	1.9	5.7	4.6	4.0	4.9			
	30%	1.6	5.6	4.7	2.2	22.5	2.6	13.4	12.2	12.2	10.3	2.4	8.8	8.2	6.5	9.0			
- 0.2	10%	1.7	1.5	1.8	1.2	5.6	1.9	3.0	3.5	2.9	2.7	1.1	2.7	3.7	2.4	3.9			
	20%	2.7	5.1	2.7	1.2	14.7	1.6	8.5	8.0	7.5	7.3	2.4	5.7	5.8	4.9	5.8			
	30%	2.2	5.7	4.6	2.7	25.3	2.1	14.1	12.7	13.6	11.0	1.5	10.0	9.2	9.1	10.8			
0.2	10%	1.4	2.7	1.9	0.9	5.9	1.7	4.4	3.5	3.9	3.5	1.3	2.9	2.8	2.1	3.3			
	20%	1.7	4.9	3.6	1.6	16.2	3.1	7.6	7.2	7.2	7.7	1.9	5.9	6.1	5.3	6.5			
	30%	2.1	6.2	4.2	1.6	24.9	1.9	13.6	11.6	12.2	8.6	2.6	9.8	9.5	8.4	11.2			
0.5	10%	2.0	2.3	2.0	1.0	7.5	1.9	4.5	4.0	3.4	4.0	0.7	3.4	4.2	2.4	4.1			
	20%	2.9	4.8	3.6	1.7	14.1	2.0	8.8	7.3	7.8	6.0	1.4	6.7	7.4	5.1	7.5			
	30%	2.4	7.1	3.8	2.4	25.8	2.6	13.3	10.9	11.0	10.0	2.3	8.4	9.1	8.0	8.7			
0.9	10%	1.9	2.4	2.3	1.6	6.1	2.3	3.8	4.4	3.9	3.3	0.9	2.6	2.8	2.3	2.1			
	20%	2.6	5.3	3.1	2.0	9.8	2.4	10.7	8.6	7.8	6.9	2.1	6.5	6.2	6.3	5.6			
	30%	2.1	6.3	4.5	1.6	16.2	2.2	15.9	13.6	15.2	11.9	2.5	10.9	10.2	9.9	8.9			

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

Table 2.5: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables, $n = 30$ observations and variance $(\sigma^2 + \sigma_1^2) = 200$.

ρ	AP	Normality						Means						Covariances					
		GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C			
- 0.9	10%	0.9	1.1	1.1	0.4	1.9	1.6	4.7	4.1	4.0	4.4	1.3	3.6	4.0	3.1	3.0			
	20%	1.6	3.3	2.2	1.0	8.2	1.2	8.2	8.0	8.0	6.8	2.0	6.5	6.4	6.0	6.4			
	30%	1.3	4.4	2.5	1.5	11.0	1.9	14.9	13.5	14.1	8.9	2.2	10.2	10.3	9.9	9.6			
- 0.5	10%	1.5	1.9	2.4	1.3	4.9	1.3	4.1	4.0	4.0	3.6	1.7	3.3	2.6	2.7	3.2			
	20%	2.8	4.5	3.3	1.7	13.7	2.2	9.5	9.6	8.3	7.2	2.4	7.6	6.2	5.4	6.6			
	30%	2.5	5.9	5.2	1.6	22.0	2.6	15.3	12.6	12.2	8.9	2.3	9.0	8.8	7.0	9.1			
- 0.2	10%	2.5	2.6	2.0	1.4	7.2	1.1	3.5	3.4	4.2	3.0	1.0	2.6	2.5	1.5	3.2			
	20%	2.1	4.4	3.1	1.7	14.0	2.6	8.4	7.3	8.1	7.3	1.4	6.4	5.6	4.0	6.4			
	30%	2.4	5.6	5.6	2.3	24.4	2.8	13.0	11.8	12.1	9.8	2.6	10.4	10.2	9.0	11.2			
0.2	10%	1.8	2.0	1.7	1.5	6.3	1.6	4.6	4.1	3.1	4.2	1.7	2.7	2.7	1.9	4.2			
	20%	1.9	3.0	2.5	1.4	13.8	2.1	8.6	6.5	6.9	5.6	1.5	4.9	5.2	3.4	5.4			
	30%	3.3	6.6	4.7	1.9	23.5	2.6	14.0	9.8	11.4	8.2	2.2	10.1	9.6	8.6	9.7			
0.5	10%	1.7	2.4	3.0	1.8	7.3	1.5	4.5	3.4	4.0	3.7	1.3	2.8	3.1	2.3	4.0			
	20%	2.0	4.6	2.0	1.1	12.1	2.1	9.5	9.1	9.7	7.9	1.8	4.5	5.7	4.2	7.2			
	30%	2.7	6.7	4.1	1.4	22.6	3.1	14.8	13.7	12.8	11.2	2.1	9.8	9.4	8.2	10.0			
0.9	10%	1.1	2.0	1.7	1.4	3.8	2.1	4.3	3.7	3.0	3.0	1.4	3.6	3.0	3.0	3.1			
	20%	1.9	3.5	3.5	1.5	11.3	2.1	9.8	8.4	9.2	5.5	1.7	7.7	7.2	6.4	8.4			
	30%	2.3	8.5	4.8	3.1	19.4	2.5	15.0	14.2	13.9	9.6	1.7	11.5	10.6	9.5	11.1			

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

Table 2.6: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables, $n = 100$ observations and variance $(\sigma^2 + \sigma_1^2) = 1$.

	ρ	AP	Normality						Means						Covariances					
			GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C			
- 0.9		10%	1.1	2.0	1.7	1.1	3.4	1.1	7.9	7.2	5.9	3.5	0.7	5.6	5.2	3.0	3.8			
		20%	1.5	3.2	2.9	1.3	9.6	3.0	17.0	15.2	15.1	8.1	1.5	11.7	11.4	12.1	5.9			
		30%	1.4	5.4	3.3	1.5	13.7	3.0	28.2	23.8	26.8	12.0	3.0	19.3	18.2	19.6	11.7			
- 0.5		10%	1.8	2.9	2.4	0.8	7.3	1.6	4.9	5.3	4.4	4.1	1.5	5.5	5.0	4.5	4.5			
		20%	1.9	3.6	3.6	1.5	16.8	2.5	12.8	9.8	10.5	6.2	1.6	9.5	8.3	7.9	7.8			
		30%	2.5	6.5	4.6	1.4	24.4	1.5	23.3	17.7	18.8	10.8	2.6	14.2	14.2	16.4	10.9			
- 0.2		10%	1.4	2.4	2.3	0.3	6.9	1.7	4.3	4.4	4.6	3.1	2.7	5.2	4.3	3.8	5.7			
		20%	1.8	3.5	3.3	1.2	16.9	2.7	10.4	8.5	9.0	6.1	2.0	9.8	8.3	9.9	8.8			
		30%	3.1	5.5	4.8	1.6	28.7	3.5	24.0	18.0	16.5	10.1	2.6	16.4	16.5	17.6	10.6			
0.2		10%	1.1	1.9	1.1	0.8	6.9	1.1	4.3	3.5	4.2	2.9	1.0	4.0	4.0	3.4	4.6			
		20%	2.2	4.5	2.5	1.4	15.5	2.5	13.5	9.6	10.9	7.2	3.2	9.5	9.2	10.2	8.3			
		30%	2.2	7.9	4.8	1.5	28.2	2.5	22.3	16.6	17.1	9.8	1.9	13.7	12.7	16.3	10.9			
0.5		10%	1.8	2.0	2.3	0.9	6.7	1.6	3.9	3.6	3.5	3.6	1.3	4.9	3.6	3.3	5.3			
		20%	1.7	5.9	4.1	2.0	17.7	1.5	12.4	9.3	9.8	7.7	1.8	9.1	7.8	8.6	7.2			
		30%	1.7	7.1	6.2	1.9	23.1	2.1	20.5	17.0	17.9	11.1	2.3	16.6	13.7	17.4	11.1			
0.9		10%	1.7	2.9	3.0	1.4	6.0	1.3	6.6	5.9	5.5	3.1	1.2	6.0	5.1	4.6	5.6			
		20%	2.2	5.4	3.9	1.5	14.2	2.4	16.9	14.4	14.6	8.2	1.8	11.4	9.8	10.8	8.6			
		30%	2.3	8.3	5.5	3.0	19.9	2.7	31.3	25.0	26.6	11.7	3.4	21.9	18.1	17.6	12.1			

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

Table 2.7: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables, $n = 100$ observations and variance ($\sigma^2 + \sigma_1^2$) = 200.

ρ	AP	Normality						Means						Covariances					
		GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C			
- 0.9	10%	1.3	1.7	1.0	0.9	4.4	1.6	6.5	5.8	5.7	3.0	1.0	4.5	3.9	3.1	3.3			
	20%	0.9	3.9	3.4	1.0	9.4	1.8	15.7	14.6	14.0	6.6	2.2	10.2	9.8	8.4	6.7			
	30%	1.2	6.1	4.2	1.1	13.3	2.5	26.7	24.8	25.4	10.4	1.7	17.9	16.8	15.2	12.9			
- 0.5	10%	2.0	2.0	2.1	1.2	8.3	1.7	5.7	3.6	4.5	3.8	0.7	3.4	4.1	3.1	4.5			
	20%	1.6	4.1	3.5	1.6	15.4	1.6	11.6	9.6	8.7	7.2	2.7	9.5	8.6	8.4	8.3			
	30%	1.9	7.3	4.7	1.8	24.4	2.7	21.1	15.2	16.5	10.2	3.0	16.9	14.6	17.0	11.0			
- 0.2	10%	1.5	2.9	2.7	1.0	7.2	1.9	5.2	2.8	4.2	2.9	1.1	3.8	2.9	1.7	3.8			
	20%	2.5	4.3	3.7	1.4	15.2	2.7	14.8	11.3	11.0	7.8	1.8	10.5	8.6	9.2	8.7			
	30%	2.5	5.9	4.1	1.7	24.7	3.0	19.9	14.8	17.2	11.1	2.4	14.9	14.5	17.5	11.1			
0.2	10%	1.6	2.4	2.1	1.0	8.5	1.2	5.4	4.1	3.7	3.8	1.5	3.7	3.3	2.8	3.0			
	20%	1.1	4.8	3.8	0.8	17.1	2.1	10.4	7.7	8.7	6.5	2.1	9.6	9.8	9.8	8.1			
	30%	2.5	6.3	4.8	1.4	26.3	1.9	19.8	16.1	15.2	9.8	2.6	15.7	13.8	16.6	10.5			
0.5	10%	1.9	2.1	2.4	1.1	5.7	1.9	5.3	5.5	5.6	3.5	0.9	4.1	4.0	4.1	3.6			
	20%	2.7	6.4	4.1	1.9	15.0	2.5	12.2	8.9	9.5	5.7	1.1	9.8	8.4	8.8	7.9			
	30%	2.4	8.2	6.2	2.2	28.0	2.6	21.6	17.1	18.2	9.4	2.8	17.6	16.3	17.4	13.3			
0.9	10%	1.8	2.6	2.4	1.6	6.1	1.2	6.2	5.0	5.5	3.7	2.3	4.7	5.9	3.7	4.8			
	20%	1.8	5.9	4.2	1.7	12.7	2.3	17.9	15.7	15.9	7.6	1.8	11.4	12.1	10.1	8.1			
	30%	2.0	9.6	6.8	1.8	18.6	3.1	29.8	24.0	25.6	12.5	2.5	20.7	17.0	17.9	10.8			

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

Unstructured

Table 2.8: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables and $n = 30$ observations.

ρ	AP	Normality						Means						Covariances																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										
		GS			UNSDC			C			GS			UNSDC			C			GS			UNSDC			C																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																														
		RDC	USDC	UNSDC	RDC	USDC	UNSDC	RDC	USDC	UNSDC	RDC	USDC	UNSDC	RDC	USDC	UNSDC	RDC	USDC	UNSDC	RDC	USDC	UNSDC	RDC	USDC	UNSDC	RDC	USDC	UNSDC																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
- 0.9	10%	0.8	1.0	1.1	0.7	2.6	1.5	3.7	3.7	3.6	3.8	1.3	4.1	3.5	2.6	4.2	- 0.5	10%	1.0	1.4	1.3	1.1	4.4	1.1	3.5	3.4	2.7	3.0	1.4	3.4	1.5	1.3	3.4	- 0.2	10%	1.4	2.6	1.9	1.0	6.4	2.1	4.0	3.9	3.6	3.1	1.1	2.5	3.3	2.2	3.9	0.2	10%	0.7	3.1	1.9	1.3	6.0	1.2	3.5	3.7	3.7	4.2	1.0	3.0	3.3	1.8	2.8	0.5	10%	1.6	2.9	2.3	1.3	7.2	1.5	3.7	4.0	3.8	2.9	0.9	3.3	2.8	2.1	3.8	0.9	10%	2.4	2.8	2.1	1.1	4.9	2.2	5.2	3.9	3.8	3.9	1.7	4.5	5.3	3.1	3.6	30%	1.6	4.4	3.5	1.0	13.9	2.7	14.2	12.6	13.7	9.8	2.4	10.6	8.7	7.8	9.6	30%	2.5	7.1	3.9	3.0	25.4	3.6	15.5	12.8	12.8	9.8	2.3	10.4	9.2	9.5	10.9	30%	3.3	6.2	5.0	1.6	24.4	2.7	14.4	13.3	12.5	10.6	2.9	8.3	8.8	8.2	8.9	30%	2.4	6.3	5.2	2.1	23.4	2.6	13.2	11.7	11.5	8.8	2.0	8.6	8.3	6.7	9.5	30%	2.7	7.3	5.2	2.5	23.1	2.8	15.2	12.5	13.6	10.9	2.2	9.6	9.1	9.0	8.6	30%	2.9	8.6	5.8	2.8	18.6	2.1	14.5	14.6	14.0	10.5	2.2	10.7	12.3	9.9	10.0																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
	20%	0.9	2.4	1.9	1.7	6.9	2.5	9.3	9.4	8.3	6.2	2.0	8.7	9.4	7.4	6.5		6.9	20%	1.4	4.3	3.5	1.0	14.1	2.3	9.7	7.4	7.0	6.0	2.3	7.8	6.8	6.5		6.9	20%	2.2	5.2	3.3	2.7	15.8	2.4	9.2	7.2	7.9	5.9	1.4	6.9	6.5	4.6		6.3	20%	2.0	5.3	4.1	1.2	15.4	2.7	9.8	7.8	7.6	7.0	2.3	6.0	5.5	5.0		6.8	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2		6.0	20%	2.5	7.3	5.2	2.5	23.1	2.8	15.2	12.5	13.6	10.9	2.2	9.6	9.1	9.0		8.6	20%	2.7	7.3	5.2	2.5	23.1	2.8	15.2	12.5	13.6	10.9	2.2	9.6	9.1	9.0	8.6	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	20%	1.7	3.7	2.5	2.1	11.9	2.2	8.7	7.8	8.5	6.3	2.4	7.5	6.3	5.2	6.0	

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

Table 2.9: Rejection rate (%) of the null hypothesis that the random variable follows a normal distribution (Normality) with zero mean vector (Means) and covariance matrix equal to the populational one (Covariances), after sample augmentation with the gold standard (GS), the three algorithms (RDC, USDC and UNSDC) and the control (C), for $p = 2$ variables and $n = 100$ observations.

ρ	AP	Normality						Means						Covariances					
		GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C	GS	RDC	USDC	UNSDC	C			
- 0.9	10%	0.8	1.8	1.6	1.2	3.9	2.1	6.1	4.8	5.1	3.5	1.9	5.9	5.5	4.4	3.7			
	20%	1.5	2.6	2.6	0.9	6.5	2.4	18.0	14.2	13.9	8.0	2.0	12.5	9.6	10.0	8.7			
	30%	1.5	6.0	4.8	1.3	16.9	3.0	28.8	25.3	25.6	10.0	2.6	22.7	19.0	19.3	12.1			
- 0.5	10%	1.8	1.6	1.6	1.0	6.7	2.2	4.7	3.9	4.7	3.3	0.9	3.9	3.4	2.5	4.0			
	20%	2.2	4.0	3.1	1.7	16.4	1.9	13.4	9.6	11.6	6.9	2.0	10.3	8.3	8.5	6.6			
	30%	2.3	6.8	4.7	2.0	26.8	2.5	21.4	18.4	16.8	9.3	2.9	17.5	14.2	15.4	11.7			
- 0.2	10%	1.5	1.6	1.8	0.8	6.7	2.2	6.4	4.9	5.0	3.7	1.9	4.6	3.8	4.5	5.3			
	20%	2.3	4.0	2.3	1.3	16.2	2.5	13.2	10.9	11.3	7.2	2.0	10.7	9.6	8.3	7.8			
	30%	3.1	5.9	4.2	2.2	26.5	2.1	19.2	17.1	15.0	10.0	2.1	15.0	13.1	15.9	11.2			
0.2	10%	1.3	2.4	1.4	0.6	7.1	1.0	5.0	4.2	4.3	3.3	1.1	4.1	3.9	3.1	5.1			
	20%	1.5	4.4	2.0	1.4	15.4	1.7	12.4	11.4	10.8	7.8	2.4	8.8	8.4	8.0	7.1			
	30%	3.3	5.9	5.1	2.1	24.5	2.0	21.5	16.4	15.3	10.4	2.8	15.7	14.6	17.3	12.3			
0.5	10%	1.5	2.2	1.3	0.6	6.0	2.1	6.0	5.7	4.5	3.9	1.3	4.1	3.4	2.9	4.5			
	20%	2.3	5.0	4.1	1.8	17.4	1.5	10.5	7.2	7.8	6.2	2.6	8.6	8.5	8.5	7.4			
	30%	1.7	7.2	5.0	1.9	27.8	2.6	23.0	17.2	18.1	8.8	2.6	14.4	13.4	17.0	10.8			
0.9	10%	1.2	2.4	2.3	2.1	6.2	1.4	5.3	5.4	5.2	3.2	1.5	5.6	4.9	3.6	5.3			
	20%	2.3	5.3	4.8	2.0	12.3	1.5	16.2	13.6	14.7	5.5	1.1	12.0	10.6	9.3	6.8			
	30%	3.1	9.3	7.4	2.5	21.3	2.5	26.9	23.0	23.6	10.5	1.8	21.9	18.9	20.4	12.6			

ρ : correlation of the variables, AP: augment proportion.

Source: table elaborated by the author.

3 ARCHETYPAL ANALYSIS AS AN IMPUTATION METHOD IN A MULTIVARIATE DATASET WITH MISSING DATA

Abstract

This study aims to evaluate the use of archetypes to impute values in a multivariate dataset about the Hector dolphin, endemic to New Zealand, with simulated missing data. The dataset consists on six body measurements of 59 dolphin skeletons from the North and South Islands. Missing completely at random data (MCAR) was simulated as follows: nine observations were removed at random, from which response variables values were removed also randomly, considering that the number of values removed follows the zero-truncated Binomial distribution. Thus, the Archetypal Analysis (AA) was applied to the standardized complete data and, from the three archetypes found, the missing values were imputed via AA by a Single and a Multiple approach, SAA and MAA, respectively. The results were compared to two known methods of imputation: the mean imputation (MI) and the Predictive Mean Matching (PMM). SAA, MAA and PMM have showed promising results indicating a great match between the original values and the imputed values, highlighting SAA. So the use of AA as an imputation method in a MCAR situation have allowed to impute the dolphins measures properly.

Keywords: archetypes; imputed data; multivariate.

3.1 Introduction

In real applications, specially in multivariate dataset, missing data are common, which results in incomplete data. Three types of missing data can be cited: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR the causes of missing values are unrelated to the data; in MAR the missing values probability is related to the values for some other observed variables; and in MNAR the probability of missing an observation is related to its value (Epifanio et al., 2020).

Regardless the type of missing data, most statistical techniques assume the completeness of the data, and so the incomplete observations are often discarded. Thus, sometimes an entire row of observations is discarded due to a missing value in only one column. This way, one may loose valuable information (Van Buuren, 2018).

A common approach to deal with it is the imputation method, which means to replace the missing data with an estimate (Donders et al., 2006). Once the missing values are filled, the full dataset can be analyzed normally. Among the techniques to impute missing data, one can cite the Archetypal Analysis (AA).

AA is a multivariate technique, proposed by Cutler and Breiman (1994), whose objective is to reduce the space of observations of a dataset by means of convex combinations of the data. In this way, archetypes are convex combinations of data points and data points are explained as convex combinations of archetypes (Epifanio et al., 2020). In addition, a set of archetypes consist on extreme elements that lies on the boundary of the data convex hull (Cutler and Breiman, 1994).

Bauchhage et al. (2015) presented another approach to AA where they used sub-gradient algorithms for optimization over the simplex to determine archetypes and reconstruction coefficients. They also applied AA as an autoencoder which inspired the use of AA to deal with missing data. Therefore, the aim of this study is to evaluate the use of archetypes to impute values in a dataset with simulated missing data.

3.2 Development

3.2.1 Archetypal Analysis

AA consists on represent each observation in a dataset as a mixture of the archetypes (Cutler and Breiman, 1994). As explained by the authors, the archetypes themselves are mixtures of the observations and are selected by minimizing the squared error of each observation as a combination of the archetypes.

Considering n observations of a set of q random variables represented by a $n \times q$ data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ and also an integer $K \leq \min\{q, n\}$ representing the number of archetypes, AA consists on determine stochastic matrices $\mathbf{B}$ ($n \times K$) and $\mathbf{A}$ ($K \times n$), such that $\mathbf{X} \approx \mathbf{XBA} = \mathbf{ZA}$, where $\mathbf{Z}$ is the archetypes matrix (Bauckhage et al., 2015).

The first AA approach presented by Cutler and Breiman (1994) proposed an alternating least squares procedure where $\mathbf{B}$ is initialized randomly and solves

$$\min_{\mathbf{A}, \mathbf{B}} E = \|\mathbf{X} - \mathbf{XBA}\|^2 \quad (3.1)$$

for $\mathbf{A}$ and then, given $\mathbf{A}$, solves (3.1) for $\mathbf{B}$. This process is repeated until the convergence towards a local minimum (Bauckhage et al., 2015).

Later, Morup and Hansen (2012) suggested another approach that runs faster than the original one based on alternating projected gradient. They observed that

$$E = \|\mathbf{X} - \mathbf{XBA}\|^2 = \text{tr}(\mathbf{X}^T \mathbf{X} - 2\mathbf{X}^T \mathbf{XBA} + \mathbf{A}^T \mathbf{B}^T \mathbf{X}^T \mathbf{XBA}) \quad (3.2)$$

and hence $\nabla_{\mathbf{A}} E = 2[\mathbf{Z}^T \mathbf{ZA} - \mathbf{Z}^T \mathbf{X}]$ and $\nabla_{\mathbf{B}} E = 2[\mathbf{X}^T \mathbf{XBA} - \mathbf{X}^T \mathbf{XA}^T]$. Thus, the AA parameters can be computed by alternating updates $\mathbf{A} \leftarrow \mathbf{A} - \eta_{\mathbf{A}} \nabla_{\mathbf{A}} E$ and $\mathbf{B} \leftarrow \mathbf{B} - \eta_{\mathbf{B}} \nabla_{\mathbf{B}} E$ where $\eta_{\mathbf{A}}$ and $\eta_{\mathbf{B}}$ are step size parameters tuned by line-search.

In order to avoid costly quadratic optimization routines or costly projections of gradients onto a feasible set, Bauckhage et al. (2015) proposed the Rapid AA. Based on Frank and Wolfe (1956) procedure and on the work of Clarkson (2010), the idea of this technique is to use the efficiency of computing gradients instead of the elaborate quadratic programming (Bauckhage et al., 2015). It consists on two algorithms described below:

Algorithm 1

Considering the data matrix $\mathbf{X}$, the matrix of the archetypes $\mathbf{Z}$ and a parameter $t_{max} \in \mathbb{N}$, Algorithm 1 consists on Frank-Wolfe procedure to compute matrix $\mathbf{A}$, whose columns reside in the simplex Δ^{K-1} .

1. Initialize the $K \times n$ matrix $\mathbf{A} = [\mathbf{e}_1, \mathbf{e}_1, \dots, \mathbf{e}_1]$, where $\mathbf{e}_1 = [1, 0, \dots, 0]^T \in \mathbb{R}^K$;
2. Set $t = 0$;
3. Compute the gradient matrix $\mathbf{G} = \nabla_{\mathbf{A}} E = 2\mathbf{Z}^T \mathbf{ZA} - \mathbf{Z}^T \mathbf{X}$;
4. Update the columns $\mathbf{a}_i$ of $\mathbf{A}$ for each $i \in \{1, \dots, n\}$ as $\mathbf{a}_i = \mathbf{a}_i + \frac{2}{t+2}(\mathbf{e}_j - \mathbf{a}_i)$ where $j = \text{argmin}_l G_{il}$;
5. Set $t = t + 1$;
6. Repeat the steps (3) to (5) until updates “become small” or $t = t_{max}$.

Algorithm 2

Considering the data matrix $\mathbf{X}$, the matrix of coefficients $\mathbf{A}$ and a parameter $t_{max} \in \mathbb{N}$, Algorithm 2 consists on Frank-Wolfe procedure to compute matrix $\mathbf{B}$, whose columns reside in the simplex Δ^{n-1} .

1. Initialize the $n \times K$ matrix $\mathbf{B} = [\mathbf{e}_1, \mathbf{e}_1, \dots, \mathbf{e}_1]$, where $\mathbf{e}_1 = [1, 0, \dots, 0]^T \in \mathbb{R}^n$;
2. Set $t = 0$;
3. Compute the gradient matrix $\mathbf{G} = \nabla_{\mathbf{B}} E = 2[\mathbf{X}^T \mathbf{X} \mathbf{B} \mathbf{A} \mathbf{A}^T - \mathbf{X}^T \mathbf{X} \mathbf{A}^T]$;
4. Update the columns $\mathbf{b}_j$ of $\mathbf{B}$ for each $j \in \{1, \dots, K\}$ as $\mathbf{b}_j = \mathbf{b}_j + \frac{2}{t+2}(\mathbf{e}_i - \mathbf{b}_j)$ where $i = \text{argmin}_i G_{ji}$;
5. Set $t = t + 1$;
6. Repeat the steps (3) to (5) until updates “become small” or $t = t_{max}$.

Therefore, alternating these two algorithms and updating the sub-gradient along affine directions $\mathbf{e}_j - \mathbf{a}_i$ and $\mathbf{e}_i - \mathbf{b}_j$ within the simplices Δ^{K-1} and Δ^{n-1} , respectively, the results converge to a set of archetypes $\mathbf{Z} = \mathbf{X} \mathbf{B}$ (Bauckhage et al., 2015).

3.2.2 Imputation methods

Incomplete data or data with missing values are a common problem in researches of all areas of knowledge (Donders et al., 2006). Among the methods of handling this situation the listwise deletion or the complete-case analysis is the most used and consists of eliminating all cases that have one or more missing values (Van Buuren, 2018). However, as stated by Donders et al. (2006), this procedure may produce biased estimates of means and regression coefficients for example and, then, leads to inefficient analyses.

A quick fix for incomplete data is to replace the missing values by the mean or the mode in categorical data (Van Buuren, 2018). Despite of being a fast and simple fix, according to Van Buuren (2018), the mean imputation can disturb the relations between variables in multivariate datasets, underestimate the variance, bias other estimates than the mean and also bias the estimate of the mean itself when data are not MCAR.

One can achieve much better results using more sophisticated techniques like multiple imputation (Donders et al., 2006). The idea was developed by Rubin (2004) who observed that imputing just one value for the missing ones could not be correct in general. So he suggested creating multiple imputations that could reflect the uncertainty of the missing data. Then, the complete datasets resultant can be analyzed separately, and its results combined together (Noghrehchi et al., 2020).

Among the different methods of imputing missing values the archetypes can be used to perform single or multiple imputation. Based on the approach of Bauckhage et al. (2015) we conducted a simulation study to evaluate the use of AA as an imputation method, which will be explained in the next section.

3.2.3 Simulation study

This study was performed on a dataset about the endemic New Zealand Hector’s dolphin, seeking differences between the South Island and the North Island populations (Baker et al., 2002). North Island skulls were sufficiently different to be made into a new subspecies: Maui’s dolphin. The following six measurements (variables) were taken from dolphin skeletons in museums: condylobasal length (CBL),

rostrum width at midlength (RWM), rostrum length (RL), zygomatic width (ZW), mandible length (ML) and rostrum width at base (RWB). From a total of 59 specimens, 13 were from the North Island and 46 from the South Island.

First of all, we applied the Principal Components Analysis (PCA) on the original dataset to understand and visualize the differences between the South Island and the North Island dolphins' populations. Then we proceeded studying the possibilities of missing data in this dataset.

In this type of data, there may be missing values due to damaged or missing bones in the collection, which could occur completely at random. So, in order to simulate MCAR data, 15% of the data (nine observations) were randomly chosen from the original dataset, and then some variables values were randomly removed from each one. This procedure was done considering that the number of removing values r in each observation follows a zero-truncated Binomial distribution $ZTBin(6, p)$ with probability p :

$$P(r) = \binom{6}{r} \frac{p(1-p)^{6-r}}{1 - (1-p)^6} \quad (3.3)$$

Considering that the probability of removing one value is p and, thus, the probability of not removing one value is $1 - p$, the probability of removing the entire row (6 values) is $1 - (1 - p)^6$. As 15% of the observations were chosen (nine of 59), the probability of removing one value is

$$1 - (1 - p)^6 = \frac{9}{59}$$

$$p = 0.0272.$$

Hence, three values of the probability p were evaluated in different scenarios: $p = 0.02$, 0.03 and 0.04 (Figure 3.1).

Figure 3.1: Probabilty density of a zero-truncated Binomial distribution $ZTBin(6, p)$.

Source: figure elaborated by the author.

In each scenario r was randomly generated considering the $ZTBin(6, p)$ distribution. Then, r variables were randomly removed from each one of the nine observations previously selected. After simulating the missing data, the computational study was performed in order to evaluate AA as single and multiple imputation methods.

In both approaches it started by performing the Rapid AA on the scaled complete data (50 observations) and choosing three archetypes. This number of archetypes was determined by running the Rapid AA algorithm for different numbers of archetypes and using the “elbow criterion” on the Residual Sum of Squares (RSS), called E on equation (3.2), in a scree plot as recommended by Eugster and Leisch (2009). In addition, for the defined number of archetypes ($K = 3$) we also repeated the procedure four

times and chose the set of archetypes with the lowest RSS to ensure that they could best represent the dataset. Next steps of the imputation methods will be explained separately below.

3.2.3.1 Single imputation via AA - SAA

The single imputation of missing data by means of the archetypes consisted of estimate $\mathbf{A}^*$ just for the missing values, that were scaled in accordance with the complete data, and then multiply it by the archetypes, such that $\tilde{\mathbf{X}} = \mathbf{Z}\mathbf{A}^*$, where $\tilde{\mathbf{X}}$ is the matrix of imputed values. This method was compared to the traditional mean imputation (MI), where the missing values were replaced by the mean of the other values taking in account the specific variable and the island of the dolphin.

Both imputation methods SAA and MI were repeated 1000 times for each scenario evaluated, considering the three p of the $ZTBin(6, p)$ distribution used for simulating the datasets (three scenarios in total). Then, two metrics were calculated to compare the imputed values to the original data: the Root Mean Squared Error (RMSE) and the agreement index of Willmott. The RMSE was calculated by

$$RMSE = \sqrt{\frac{1}{ni} \sum_{l=1}^{ni} (x_l - \tilde{x}_l)^2} \quad (3.4)$$

where x_l is the original data, $\tilde{x}_l$ is the imputed value and ni is the number of imputed values, which varies according the number of missing values of each one of the nine observations selected from the original data, if just one value were removed from each observation in the simulation study, for example, then $ni = 9$.

The Willmott index (d) represents the ratio of the Mean Squared Error (MSE) and the potential error:

$$d = 1 - \frac{\sum_{l=1}^{ni} (x_l - \tilde{x}_l)^2}{\sum_{l=1}^{ni} (|\tilde{x}_l - \bar{x}| + |x_l - \bar{x}|)^2} \quad (3.5)$$

where $\bar{x}$ is the mean of the removed values and $0 \leq d \leq 1$ (Willmott, 1981). The closer the agreement value is from 1 the better is the match.

3.2.3.2 Multiple imputation via AA - MAA

The multiple imputation of missing data by means of the archetypes was executed in the same way of the single method, but repeating the procedure m times and calculating m data for each missing value. To do this, we selected m set of archetypes by updating the initial values of the archetypes matrix $\mathbf{Z}$ in Algorithm 1 of Rapid AA with the previous selected in order to achieve more accurate archetypes and hence results even closer to the real ones.

The results of this method was compared to the Predictive Mean Matching (PMM) which is the default method implemented in the *mice* R package (Van Buuren and Groothuis-Oudshoorn, 2011). PMM is a versatile and easy-to-use method in which linear regression and random value selection for imputation are combined (Van Buuren, 2018). Santos and Conde (2020) explain that a linear prediction is used considering that the variable of interest is the variable to be imputed and the other variables are the explanatory ones. Thus, the imputations are very realistic as they are based on values observed elsewhere and because imputations outside the data range will not occur (Van Buuren, 2018).

Both imputation methods MAA and PMM were repeated 1000 times for each scenario evaluated, considering the three p specified and three values of m : 3, 5 and 10 (nine scenarios in total). As in the

single approach, the RMSE and the Willmott index were calculated by comparing the mean of the m imputed values to the original data using the equations (3.4) and (3.5).

At the end of the simulation study, the PCA was performed in the scaled filled dataset in order to characterize and visualize the complete data, their archetypes and the imputed values. All computing was performed in R software (R Core Team, 2021).

3.2.4 Results and discussion

Exploring the original data on its totality, in Figure 3.2 one can see the first two principal components, which explain together 93.6% of the total variance and then are enough to characterize dolphins' measurement.

Figure 3.2: Biplot of the first two principal components PC1 and PC2 resulting of the PCA on the original dataset about the endemic New Zealand Hector's dolphin. Red dots represent the North Island dolphins and blue dots represent the South Island dolphins.

Source: figure elaborated by the author.

The North Island dolphins (red dots) are located in the first and fourth quadrants of the graph as well as the variables vectors, which means that these specimens presented higher values of the response variables. The majority of the South Island dolphins (blue dots) are located in the second and third quadrants, hence, in the opposite direction of the variables vector, this way, they presented lower values and can be considered smaller than the North Island dolphins (Figure 3.2).

Next, Table 3.1 shows the nine observations randomly selected from the original dataset that were used to simulate the missing values. It can be seen that there is only two data from North Island and seven data from the South Island, which was expected as the number of North Island's specimens was smaller. And it's also noticeable that the measurements of the dolphins of the North Island are in general higher as we could see in Figure 3.2.

Table 3.1: Observations randomly selected from the original dataset to simulate the missing values.

Island	CBL	RWM	RL	ZW	ML	RWB
North	300	61	146	153	246	85
North	284	57	143	148	232	83
South	285	46	140	141	223	75
South	285	52	138	147	228	83
South	277	52	131	143	220	83
South	284	49	134	136	222	76
South	287	46	143	139	230	76
South	264	47	128	134	205	71
South	280	51	139	145	219	75

Source: table elaborated by the author.

From each one of these observations were removed values according to the $ZTBin(6, p)$ distribution considering the three values of p previously defined ($p = 0.02, 0.03$ or 0.04) and so the missing dataset was prepared to evaluate the different imputation methods. As the results of all methods discussed in this paper were very similar by different values of p , we decided to illustrate it with the scenarios where $p = 0.03$.

Starting with the results of single imputation, in Figure 3.3 one can observe the boxplot of both metrics calculated RMSE and Willmott index on the results of SAA and MI imputation methods in the 1000 simulations.

Figure 3.3: Boxplots of the results of RMSE and Willmott index on the imputed values with the SAA and MI method.

Source: figure elaborated by the author.

It can be seen that SAA presented more precise results than MI in both metrics evaluated, with lower RMSE (between 1.1 and 5.4) and higher Willmott index that was very close to 1 indicating a great match to the original values ($d > 0.99$) (Figure 3.3).

In Figure 3.4 below we have the results of the MAA method for the three different number of imputed values evaluated ($m = 3, 5$ and 10), where we can note that we almost can see no difference between these scenarios. This outcome shows that the way how Rapid AA was used was very effective in the selection of the archetypes and that updating its value has made a minimum or no significant improvement, so the estimate of the imputed value was very similar.

Figure 3.4: Boxplots of the results of RMSE and Willmott index on the imputed values with the MAA method for the different number of imputed values $m = 3, 5$ and 10.

Source: figure elaborated by the author.

Again we can see an outstanding Willmott index of $d > 0.99$ and the RMSE varied between 1 and 5.4, which was very close to the result obtained with SAA (Figures 3.3 and 3.4). Next figure shows the results of the PMM method also for the three different number of imputed values considered (Figure 3.5).

Figure 3.5: Boxplots of the results of RMSE and Willmott index on the imputed values with the PMM method for the different number of imputed values $m = 3, 5$ and 10.

Source: figure elaborated by the author.

PMM method has performed slightly better with greater number of imputed values as we can observe in Figure 3.5. The Willmott index was excellent regardless the number of imputed values ($d > 0.99$), but the RMSE reduced from the interval 1.4 to 8.2 where $m = 3$ to 1 to 5.8 where $m = 10$. Thus, can be interesting to use this method with more imputed values in order to achieve more accurate results. Next, in Figure 3.6 we selected the most interesting results of each method and put it together to facilitate the comparison, they are: SAA, MAA ($m = 3$), PMM ($m = 10$) and MI.

Figure 3.6: Boxplots of the results of RMSE and Willmott index on the imputed values with the SAA, MAA ($m = 3$), PMM ($m = 10$) and MI.

Source: figure elaborated by the author.

First, we highlight that SAA, MAA and PMM have presented an exceptional Willmott index indicating a great match between the imputed and the original values and were higher than MI. Then, analyzing the RMSE one can see a contrast from MI to the other three methods, which had very similar results again. The RMSE median of the results of SAA, MAA, PMM and MI were respectively 3.2, 3.2, 3.7 and 6.6 (Figure 3.6).

It was already expected that MI were the method with the worst metrics as it doesn't take into account the relation between the variables like the other methods do, but even so it performed well because has considered the differences between the islands, which was not interesting to do in the other methods because of the small number of observations in the North Island.

The AA based imputation methods have presented excellent results in all scenarios evaluated corroborating to Epifanio et al. (2020) that have also used archetypes to impute missing values. These methods have the advantage to take into account the covariance structure between the variables not only the values observed in each variable. So the archetypes acts like a measure of location and dispersion at the same time which leads to the success of its use in the imputation approach.

Comparing the both methods proposed among themselves shows us that performing Rapid AA just one time (SAA) or m times updating the matrix $\mathbf{Z}$ (MAA) can lead to approximately the same result (Figure 3.6). It means that performing the Rapid AA four times and selecting the best set of archetypes (the one with the lowest RSS) it's enough to represent the dataset and impute the missing values accurately.

The imputation methods via AA proved to be competitive to PMM as we can see in Figure 3.6 and they also have the advantage to present the same or better results with a lower number of imputations considering MAA or just one (single imputation) in the case of SAA. Therefore, among the studied methods we recommend SAA which is the simplest one. According to this recommendation it was chosen a scenario where the missing values were imputed by SAA and in Table 3.2 it can be seen the nine observations selected from the original dataset with the imputed values highlighted in bold type.

Table 3.2: Observations randomly selected from the original dataset with imputed values (bold) by SAA.

Island	CBL	RWM	RL	ZW	ML	RWB
North	299.6	61.0	146.0	153.0	246.0	85.0
North	284.0	57.0	143.0	148.0	232.0	83.0
South	285.0	46.0	139.0	141.0	223.0	75.0
South	285.0	52.0	138.0	147.0	228.6	83.0
South	277.8	52.0	131.0	143.0	220.0	83.0
South	284.0	49.0	134.0	136.0	222.0	73.9
South	287.0	46.0	141.1	139.0	230.0	76.0
South	264.0	47.0	128.0	133.6	205.0	71.0
South	280.0	49.9	139.0	145.0	219.0	75.0

Source: table elaborated by the author.

The results are noteworthy as the imputed data were very close to the original data. Next, in Figure 3.7, one can see the biplot with the results of PCA for the filled dataset, where we can see the complete data (red and blue dots), their archetypes (purple dots) and the imputed values (red and blue triangles).

Figure 3.7: Biplot of the first two principal components PC1 and PC2 resulting of the PCA on the filled dataset. Red dots represent the North Island dolphins of the complete data, blue dots represent the South Island dolphins of the complete data, red triangles represent the North Island dolphins of the imputed values, blue triangles represent the South Island dolphins of the imputed values, and purple dots represent the archetypes of the complete data.

Source: figure elaborated by the author.

The first two principal components explained together 93.64% of the total variance, almost the same percentage of the PCA on the original dataset (93.6%). One more time, it can be noted that the imputed data were similar to the original ones and corroborated with the interpretation of the dataset (Figure 3.7).

As in Figure 3.2, in Figure 3.7 we can observe the North Island dolphins (red dots and triangles) in the first and fourth quadrants of the graph and the South Island dolphins (blue dots and triangles) in

the second and third quadrants. Now one can also see the three archetypes: archetype 2 had the lowest values of all measurements, representing the smallest dolphins, archetype 1 had the higher values of all variables, representing the largest dolphins, and archetype 3 represents the medium-size dolphins. The combination of these three archetypes allowed to impute the missing values with exactitude.

3.3 Conclusions

The AA based imputation methods showed promising results in a MCAR situation, proving to be competitive to other methods like PMM and the body measurement of dolphins were successfully imputed via archetypes.

In future work we aim to address the imputation method via AA on other types of missing data, like MAR and MNAR.

References

- Baker, A. N., Smith, A. N., and Pichler, F. B. (2002). Geographical variation in hector’s dolphin: Recognition of new subspecies of *cephalorhynchus hectori*. *Journal of the Royal Society of New Zealand*, 32(4):713–727.
- Bauchhage, C., Kersting, K., Hoppe, F., and Thureau, C. (2015). Archetypal analysis as an autoencoder. In *Workshop New Challenges in Neural Computation*, page 8. Citeseer.
- Clarkson, K. L. (2010). Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):1–30.
- Cutler, A. and Breiman, L. (1994). Archetypal Analysis. *Technometrics*, 36(4):338–347.
- Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.
- Epifanio, I., Ibáñez, M. V., and Simó, A. (2020). Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles. *The American Statistician*, 74(2):169–183.
- Eugster, M. J. A. and Leisch, F. (2009). From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software*, 30(8):1–23.
- Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110.
- Morup, M. and Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63.
- Noghrehchi, F., Stoklosa, J., and Penev, S. (2020). Multiple imputation and functional methods in the presence of measurement error and missingness in explanatory variables. *Computational Statistics*, pages 1–27.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician*, 58(4):298–302.
- Santos, I. K. S. d. and Conde, W. L. (2020). Predictive mean matching como método de imputação alternativo ao hot deck no vigitel. *Cadernos de Saúde Pública*, 36:e00167219.

- Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman & Hall/CRC press, 2 edition.
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, pages 1–67.
- Willmott, C. J. (1981). On the validation of models. *Physical geography*, 2(2):184–194.

4 FINAL CONSIDERATIONS

The use of the Archetypal Analysis for sample augmentation and as an imputation method both in multivariate scenarios were proposed and showed how versatile and efficient archetypes can be. Multivariate sample augmentation via archetypes is efficient regardless the covariance structure among the variables which reinforces the advantage of this method has no strong assumption about the data. And the AA based imputation methods performs great in a Missing Completing At Random situation as it can achieve a great agreement between the original and imputed values, successfully filling up the missing data.

The results were noteworthy and indicated that the archetypes can properly represent the dataset and so express it as a new data or filling up possible missing values. As missing or incomplete data are common in several areas of knowledge, this study is relevant because it brings two new tools to deal with this type of situation and so it can be widely applied in different contexts. Future works should evaluate such tools on different datasets, varying the number of variables and its probability distributions or the types of missing data.