UNIVERSIDADE FEDERAL DE ALFENAS
UNIFAL-MG

PÓRTYA PISCITELLI CAVALCANTI

PROPOSTA DE ALGORITMOS PARA AUMENTO DE DADOS VIA ARQUÉTIPOS

ALFENAS - MG
2016

PÓRTYA PISCITELLI CAVALCANTI

PROPOSTA DE ALGORITMOS PARA AUMENTO DE DADOS VIA ARQUÉTIPOS

Dissertação apresentada à Universidade Federal de Alfenas, como parte dos requisitos para obtenção do título de Mestre em Estatística Aplicada e Biometria.
Área de concentração: Estatística Aplicada e Biometria.
Linha de pesquisa: Modelagem Estatística e Estatística Computacional.

Orientador: Dr. Eric Batista Ferreira

ALFENAS - MG
2016

Dados Internacionais de Catalogação-na-Publicação (CIP)
Biblioteca Central da Universidade Federal de Alfenas

Cavalcanti, Pórtya Piscitelli.
Proposta de algoritmos para aumento de dados via arquétipos / Pórtya Piscitelli Cavalcanti. -- 2016.
56 f. : il.

Orientador: Eric Batista Ferreira.
Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Alfenas, 2016.
Bibliografia.

1. Análise multivariada. 2. Estatística matemática. 3. Monte Carlo, Método de. 4. Ausência de dados (Estatística). I. Ferreira, Eric Batista. II. Título.

CDD 519.535

MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Alfenas / UNIFAL-MG
Programa de Pós-graduação em Estatística Aplicada e Biometria
Rua Gabriel Monteiro da Silva, 700. Alfenas - MG CEP 37130-000
Fone: (35) 3299-1392 (Secretaria) (35) 3299-1121(Coordenação)
https://www.unifal-mg.edu.br/ppgeab/

1914 100 anos 2014 Universidade Federal de Alfenas

PÓRTYA PISCITELLI CAVALCANTI

"PROPOSTA DE ALGORITMOS PARA AUMENTO DE DADOS VIA ARQUÉTIPOS"

A Banca Examinadora, abaixo assinada, aprova a Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Linha de Pesquisa: Modelagem Estatística e Estatística Computacional.

Aprovado em: 11 de julho de 2016.

Prof. Dr. Eric Batista Ferreira
Instituição: UNIFAL-MG
Assinatura: _____

Prof. Dr. Denismar Alves Nogueira
Instituição: UNIFAL-MG
Assinatura: _____

Profa. Dra. Roberta Bessa Veloso Silva
Instituição: UNFENAS
Assinatura: Roberta Bessa V. Silva

A todos os amigos especiais do PPGEAB.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me guiar em mais essa etapa da minha vida.

Ao meu orientador e amigo Prof. Eric, por me fazer enxergar a Estatística com outros olhos, que agora brilham por ela, e me inspirar a seguir essa carreira tão bela que é ser professor.

A todos professores que me guiaram e me ajudaram a traçar esse caminho com segurança, especialmente aos Profs. Luiz Beijo, Denismar, Roberta e aos demais membros da banca por todo suporte na etapa final.

Aos meus pais Ângela e Fernando, pelo apoio imensurável, por sempre acreditarem em meu potencial e por me mostrarem o verdadeiro significado da vida.

Aos meus irmãos Vytória, Crystal e Arthur pela atenção e carinho em todas as ocasiões, vocês são minhas preciosidades.

Aos meus avós e familiares, que mesmo distantes fisicamente, se fazem presentes em cada pensamento.

Ao meu namorado Antonio por estar ao meu lado literalmente em todos os momentos, fazendo com que cada dificuldade se tornasse apenas mais um degrau para subirmos juntos.

Aos meus amigos pelos momentos de descontração, que tornaram essa fase mais leve. Em especial à Isabela, Larissa, Renata, Gustavo, Cássia e Helen, vocês são pessoas especiais e diferenciadas. Vou levar sempre comigo!

Enfim, a todos que fizeram e fazem parte da minha vida, muito obrigada!

RESUMO

Arquétipos, na estatística, são os elementos extremos mais representativos de uma amostra ou população, a partir dos quais todos os outros podem ser reescritos. A Análise de Arquétipos (AA) é uma técnica multivariada que visa reduzir a dimensionalidade dos dados, por meio de combinações convexas dos próprios dados, proporcionando encontrar e selecionar seus arquétipos. Existem aplicações da AA em diversas áreas do conhecimento, contudo ainda não foi explorado o seu potencial no aumento de dados amostrais. Quando um conjunto de dados é caracterizado como incompleto ou não possui o tamanho necessário para cometer o erro desejado no procedimento de inferência estatística, surge a ideia, ou necessidade, de aumentar essa amostra. Para esse fim, a técnica de aumento de dados consiste em introduzir dados não observados ou variáveis latentes por meio de métodos iterativos ou algoritmos de amostragem. Sendo assim, como os arquétipos permitem reescrever os elementos amostrais com um erro mínimo, gerando elementos não observados, esses poderiam ser utilizados para o aumento de dados. Então, o objetivo deste trabalho foi propor e avaliar a eficiência do aumento de dados por meio dos arquétipos. Foram programados três algoritmos para aumento de dados amostrais via arquétipos (Algoritmos 1, 2 e 3 - A1, A2 e A3, respectivamente), e foram realizados dois estudos de simulação para avaliar e comparar cada algoritmo quanto à sua eficiência; sendo testada a distribuição da variável aleatória e as estimativas de seus parâmetros, e também para verificar se esse aumento pode ser executado sucessivas vezes. Além disso, foi feita a aplicação dos algoritmos em um conjunto de dados reais sobre análise sensorial. Os três algoritmos apresentaram resultados semelhantes, destacando-se o A3, por ter apresentado um desempenho apropriado em todos os cenários. Esse algoritmo permitiu aumentar 10% do tamanho da amostra inicial, sem alterar a distribuição de probabilidade, bem como as estimativas de seus parâmetros. O estudo sobre aumentos sucessivos de dados também indicou o A3 como o mais eficiente, que foi capaz de aumentar a amostra em 110% de seu tamanho inicial, através de 11 aumentos sucessivos de 10% cada. O estudo com dados reais permitiu aumentar o tamanho da amostra e proporcionar maior precisão na inferência praticada. Portanto, parece seguro realizar o aumento de dados via arquétipos sugerindo-se o algoritmo 3.

Palavras-chave: Análise de Arquétipos. Dados aumentados. Estatística multivariada.

ABSTRACT

In statistics, archetypes are the most representative extreme observations of a sample or population, from which all others can be written. The Archetypal Analysis (AA) is a multivariate technique that aims to reduce the dimensionality of data through convex combinations of data itself, providing to find and select their archetypes. There are applications of AA in several areas of knowledge, but its potential in sample data augmentation still has not been exploited. When our data set is characterized as missing data or does not have the size needed to make the desired error in statistical inference procedure, there is the idea or need to increase this sample. For this purpose, data augmentation technique consists to introduce non observed data or latent variables by iterative methods or sampling algorithms. Thus, as archetypes allow rewriting the sample elements with a minimum error, generating elements not observed, these could be used to augment data. So the aim of this work was to propose and evaluate the efficiency of data augmentation through archetypes. Three algorithms were programmed to augment sample data using the archetypes (Algorithms 1, 2 and 3 - A1, A2 and A3, respectively), and two simulation studies were conducted to assess and compare the algorithms about the efficacy; testing the random variable distribution, and the estimatives of its parameters, and also to check whether this augment can be run successive times. In addition, was made an application of the algorithms into a real sensory analysis data. All algorithms showed similar results, highlighting the A3, that present an appropriate performance in all scenarios. This algorithm allowed to augment 10% of the initial sample size, without changing the probability distribution, as well as estimatives of its parameters. The study about successive augments also indicated A3 as the most efficient, that was able to augment the sample up to 110% of their initial size by 11 successive augments of 10%. The study with real data allowed to augment the sample size and improve the precision in practiced inference. So it seems safe to perform data augmentation by archetypes suggesting the algorithm 3.

Keywords: Archetypal Analysis. Augmented data. Multivariate statistics.

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Um arquétipo pode ser definido como um padrão, modelo, protótipo, entre outros. O conceito de arquétipos é utilizado em diferentes áreas, como literatura, filosofia, psicologia e inclusive na estatística. Nesta última, os arquétipos seriam os elementos extremos mais representativos de uma amostra ou população, a partir dos quais todos os outros podem ser reescritos.

A Análise de Arquétipos (AA) é uma técnica multivariada proposta por Cutler e Breiman (1994) com o objetivo de reduzir a dimensão dos dados através de seus arquétipos. Os arquétipos são obtidos por combinações lineares dos dados e são selecionados minimizando-se a soma de quadrados dos erros cometidos na reconstrução de cada observação dos dados originais como combinação dos arquétipos.

As aplicações da AA vêm sendo estudadas em vários campos de atuação como na astrofísica, sensometria, *marketing*, biologia, esportes, aprendizado de máquinas, entre outras.

Entretanto, apesar das diversas aplicações encontradas para os arquétipos, existe um grande potencial, ainda não explorado, para seu uso: o aumento de dados amostrais.

Quando um conjunto de dados apresenta dificuldades durante o procedimento de inferência estatística por estar incompleto ou por possuir menos observações do que o desejado, o pesquisador pode ter interesse em completar essa amostra. Com essa finalidade, a técnica de aumento de dados se baseia na introdução de dados não observados ou variáveis latentes por meio de métodos iterativos ou algoritmos de amostragem.

Essa técnica foi introduzida com a proposta de um algoritmo determinístico, utilizado para calcular estimativas de máxima verossimilhança de dados incompletos. No entanto, o termo aumento de dados teve sua origem em um contexto bayesiano, com a implementação de um algoritmo estocástico. Desta forma, existem aplicações do aumento de dados tanto na inferência frequentista, quanto na bayesiana, e com diversos propósitos.

Tendo em vista o que foi supracitado, considerando uma amostra representativa da população e a capacidade dos arquétipos reescreverem os elementos amostrais com um erro mínimo, esses poderiam ser utilizados para o aumento de dados, ao gerarem elementos não observados na amostra original.

Dessa afirmação, surge a hipótese de que é possível obter elementos não observados em uma amostra da população por meio dos seus arquétipos e, portanto, esse trabalho tem por objetivo propor e avaliar a eficiência do aumento de dados via arquétipos. Mais especificamente,

tem-se os seguintes objetivos:

- Propor algoritmos que realizem o aumento de um conjunto de dados por meio de seus arquétipos.
- Analisar a performance dos algoritmos em dois estudos computacionais, comparando o desempenho destes com um controle positivo e uma testemunha.
- Comparar os resultados de aumentos sucessivos realizados utilizando apenas os arquétipos da amostra inicial com os de aumentos sucessivos recalculando os arquétipos nas amostras aumentadas.
- Eleger o algoritmo mais eficaz e recomendar o maior aumento de dados, conforme o tamanho da amostra e o número de variáveis, visto que foram considerados como vindos do mesmo mecanismo gerador de dados.
- Verificar o desempenho do algoritmo eleito em dados reais sobre análise sensorial.

2 REVISÃO DE LITERATURA

Nesta seção foi detalhada a teoria da Análise de Arquétipos, explicado o seu modelo matemático e apresentadas as suas aplicações. Em seguida, foi explanado sobre o aumento de dados e seus propósitos.

2.1 ANÁLISE DE ARQUÉTIPOS

A estatística multivariada tem como propósito a análise, descrição e inferência de diversas variáveis medidas simultaneamente em experimentos, através da estrutura de correlação entre essas variáveis, a fim de proporcionar uma análise mais informativa e com resultados otimizados (FERREIRA, 2011).

De acordo com Cutler e Breiman (1994), a Análise de Arquétipos é uma técnica multivariada que tem por objetivo a redução da dimensão dos dados através de combinações convexas dos arquétipos - sendo esta combinação convexa uma particularidade da combinação linear, em que os coeficientes são valores não negativos e o somatório destes totaliza uma unidade. Como consequência, a AA facilita a interpretação dos resultados (MARTINS JÚNIOR et al., 2015b).

A seleção dos arquétipos ocorre conforme a minimização da soma de quadrados de resíduos (SQR), ou seja, na reconstrução de cada observação dos dados originais como combinação convexa dos arquétipos, serão escolhidos aqueles que proporcionarem menor SQR (CUTLER; BREIMAN, 1994).

Quando um arquétipo selecionado é uma própria observação da amostra, tem-se um arquétipo puro. Em situações cuja interpretação de uma mistura de observações não tem sentido prático, a Análise de Arquetipóides permite selecionar apenas arquétipos puros, neste caso chamados arquetipóides (VINUÉ; EPIFANIO; ALEMANY, 2014).

Ao selecionar um número de arquétipos maior que um, estes se encontram na fronteira do fecho convexo dos dados (CUTLER; BREIMAN, 1994). Eddy (1977) define o fecho convexo de um conjunto de pontos como o menor polígono convexo que contenha todos os pontos em seu interior. Desta forma, os pontos reconstruídos nunca estarão fora do fecho convexo do conjunto original dos dados (STONE; OLSON, 1999).

É importante ressaltar que quanto maior o número de arquétipos selecionados menor

será a SQR, contudo a redução da dimensão dos dados também será menor, ficando a critério do pesquisador a escolha do número ideal de arquétipos (K) para cada situação particular, contanto que $1 \leq K \leq N$, sendo N o número de elementos na fronteira do fecho convexo (MARTINS JÚNIOR et al., 2015b). Para auxiliar essa decisão sobre o número de arquétipos, é recomendada a construção de um gráfico *scree plot* (CUTLER; BREIMAN, 1994), onde pode-se observar a quantidade da variação explicada de acordo com o número de arquétipos.

2.1.1 Modelo Matemático

Segundo Bauckhage e Thurau (2009), considerando uma matriz ${}_{n}\boldsymbol{X}_{\boldsymbol{p}}$ representando um conjunto de dados multivariados com n observações e p variáveis, tal que $\boldsymbol{x}_i \in \mathbb{R}^p$ e $i = 1,\ldots,n$. A AA visa encontrar um conjunto de arquétipos $\boldsymbol{z}_k \in {}_{\boldsymbol{K}}\boldsymbol{Z}_{\boldsymbol{p}}$, sendo $k = 1,\ldots,K$ e $K < n$, que sejam combinações lineares dos dados

$$\boldsymbol{z}_k = \sum_{i=1}^{n} \boldsymbol{x}_i \beta_{ik} \tag{2.1}$$

em que $\beta_{ik} \geq 0$ e $\sum_{i=1}^{n} \beta_{ik} = 1$, caracterizando a combinação convexa (EUGSTER; LEISCH, 2009). Desta forma, para um determinado conjunto de arquétipos, a AA minimiza

$$||\boldsymbol{x}_i - \sum_{k=1}^{K} \alpha_{ik} \boldsymbol{z}_k||^2 \tag{2.2}$$

que também ocorre por combinação convexa, ou seja $\alpha_{ik} \geq 0$ e $\sum_{k=1}^{K} \alpha_{ik} = 1$. Sendo assim, com o intuito de selecionar um conjunto de arquétipos que melhor reescreva os dados originais, deve-se minimizar a soma de quadrados dos resíduos (SQR) (CUTLER; BREIMAN, 1994)

$$SQR = \sum_{i=1}^{n} ||\boldsymbol{x}_i - \sum_{k=1}^{K} \alpha_{ik} \boldsymbol{z}_k||^2 = \sum_{i=1}^{n} ||\boldsymbol{x}_i - \sum_{k=1}^{K} \alpha_{ik} \sum_{l=1}^{n} \beta_{lk} \boldsymbol{x}_l||^2. \tag{2.3}$$

Conforme Bauckhage e Thurau (2009), a Equação (2.3) também pode ser escrita matricialmente. De modo que o conjunto de dados $\boldsymbol{x}_i \in \mathbb{R}^p$ esteja representado pela matriz ${}_{n}\boldsymbol{X}_{\boldsymbol{p}}$ e os arquétipos $\boldsymbol{z}_k \in \mathbb{R}^p$ pela matriz ${}_{\boldsymbol{K}}\boldsymbol{Z}_{\boldsymbol{p}}$,

$$SQR = ||\boldsymbol{X} - \boldsymbol{AZ}||^2 = ||\boldsymbol{X} - \boldsymbol{ABX}||^2, \tag{2.4}$$

em que $\boldsymbol{A} \in \mathbb{R}^{n \times K}$ e $\boldsymbol{B} \in \mathbb{R}^{K \times n}$ são as matrizes que contêm os coeficientes α_{ik} e β_{ik}, respectivamente.

A Equação 2.3 pode ser solucionada pelo algoritmo descrito em Martins Júnior et al. (2015b), que utiliza procedimentos iterativos alternados. Os princípios básicos desse algoritmo se baseiam em dois fundamentos: encontrar os melhores coeficientes α_{ik} para determinado conjunto de arquétipos $\boldsymbol{z}_k$ e encontrar os melhores arquétipos $\boldsymbol{z}_k$ para um dado conjunto de coeficientes α_{ik} (EUGSTER; LEISCH, 2009). Em cada etapa, é necessário resolver diversos problemas de quadrados mínimos convexos fazendo com que a SQR seja reduzida sucessivamente (CUTLER; BREIMAN, 1994).

Segundo Bauckhage e Thurau (2009), a AA fornece resultados de fácil compreensão, principalmente quando comparada com com outros métodos de redução de dimensionalidade ou técnicas de agrupamento. Outra vantagem desta análise é permitir uma simples classificação ou agrupamento dos dados, ao considerar os coeficientes α_{ik} de cada observação $\boldsymbol{x}_i$ como probabilidade condicional de um arquétipo $\boldsymbol{z}_k$, ou seja, a $p(\boldsymbol{x}_i|\boldsymbol{z}_k)$ indica qual a classe (arquétipo $\boldsymbol{z}_k$) mais provável de representar o ponto.

2.1.2 Aplicações

Nos últimos anos, vários estudos vem sendo realizados sobre a aplicação da AA em diferentes áreas do conhecimento. Dentre essas áreas, pode-se citar astrofísica, aprendizado de máquinas, economia, *marketing*, reconhecimento de padrões, e análises esportivas (MARTINS JÚNIOR et al., 2015b). A seguir serão apresentados alguns dos estudos mais relevantes.

Em 1996, foi feita uma comparação entre as principais características, vantagens e desvantagens da AA e da ACP (STONE; CUTLER, 1996). Em 1999, foi proposto um método híbrido de Arquétipos e Componentes Principais para análise de sistemas dinâmicos por Stone e Olson (1999). Os autores testaram dados de espaços de dimensões acima de 500, utilizando primeiramente a ACP para reduzir a dimensão e, em seguida, a AA. O método foi nomeado Arquétipos Móveis e através dele foi possível diferenciar situações de movimentação de situações de inércia.

Já em 2003, com o intuito de melhorar a interpretação dos resultados, foi publicado um estudo que combinou a AA e a ACP por Chan, Mitchell e Cram (2003). Estes autores demons-

traram que a AA é um método eficaz na classificação de espetros das galáxias e mostraram sua robustez na presença de observações extremas (*outliers*).

Na área de *marketing*, mais especificamente na segmentação de mercado, a AA foi utilizada no lugar Análise de *Clusters* (Agrupamentos) devido à vantagem de proporcionar informações sobre consumidores extremos e consumidores médios isolados, não apenas consumidores médios - como ocorre na outra técnica, e também por oferecer uma nova perspectiva de segmentação e heterogeneidade de consumidores (LI et al., 2003; D'ESPOSITO; PALUMBO; RAGOZINI, 2006; RIEDESEL, 2014). Riedesel (2014) ainda ressalta o grande benefício de se conhecer os arquétipos puros, que possibilitam encontrar o consumidor alvo ao invés de combinações dos consumidores.

Em sensometria, D'Esposito, Palumbo e Ragozini (2011) estenderam a aplicação da AA para dados intervalares. Na análise sensorial de queijos, os autores relataram que a técnica permitiu realçar os atributos sensoriais dos produtos avaliados. Essa mesma técnica também obteve bons resultados em análises exploratórias; em um experimento sobre morcegos, foi possível identificar as espécies que representavam as demais (D'ESPOSITO; PALUMBO; RAGOZINI, 2012).

Em estudos sobre atletas, Eugster (2011) propõe uma nova perspectiva. Segundo o autor, a AA pode ser utilizada na identificação dos atletas mais habilidosos, o que foi exemplificado com jogadores de basquete e futebol. Ainda na área esportiva, um estudo sobre a movimentação de jogadores de futebol verificou que a AA permite analisar o desempenho de um time, após uma partida, o que contribui para a tomada de decisões dos técnicos (MARTINS JÚNIOR et al., 2015a).

Já no ambiente científico, a AA foi aplicada em uma base de dados com quase trinta mil economistas a fim de descobrir os arquétipos de cientistas da área de economia, verificando a quantidade de artigos publicados, suas citações e *downloads* (SEILER; WOHLRABE, 2013).

Com o propósito de averiguar o interesse de consumidores de jogos eletrônicos, Sifa, Bauckhage e Drachen (2014) apresentaram um modelo em função do tipo de jogo e do tempo jogado utilizando a AA.

Além do explanado, a AA ainda possui aplicações em dados de expressão gênica, dados antropométricos, na área de matemática, inteligência artificial, *Data Mining* e Aprendizado de Máquinas (COSTANTINI et al., 2012; MORUP; HANSEN, 2012; EPIFANIO; VINUÉ; ALEMANY, 2013; SIFA; BAUCKHAGE, 2013; THOGERSEN et al., 2013); e os arquétipos

também foram utilizados como pontos de referência (*benchmark*) (PORZIO; RAGOZINI; VISTOCCO, 2008).

Uma nova abordagem que pode ser incluída nesse leque de possíveis aplicações dos arquétipos, é a sua utilização no aumento de dados amostrais, devido à sua capacidade de reescrever os elementos amostrais com um erro mínimo, gerando elementos não observados na amostra original.

2.2 AUMENTO DE DADOS

Frequentemente são encontrados conjuntos de dados caracterizados como *missing data* (dados faltantes, ausentes ou incompletos), ou com menos dados do que o desejado (ou necessário). Nesse contexto, surge a ideia de completar a amostra e, portanto, a técnica de aumento de dados consiste em aumentar de modo estatisticamente adequado um conjunto de dados observados, de modo a torná-lo mais propício para analisar (TANNER; WONG, 1987). Sendo assim, o aumento de dados pode ser empregado em diversas situações de pesquisa cotidianas, exemplificadas a seguir.

Quando são realizadas pesquisas que utilizam questionários, pode acontecer de informantes se recusarem ou esquecerem de responder determinada pergunta, ou de pesquisadores não salvarem corretamente os dados coletados e perderem arquivos (PIGOTT, 2001). Também pode haver perda de parcelas experimentais em experimentos cujas unidades são seres vivos, que podem vir a óbito, gerando dados desbalanceados que podem ocasionar dificuldades na análise estatística.

Além disso, em alguns casos deseja-se utilizar o menor número possível de parcelas: como em pesquisas com animais, por questões éticas (SCHNAIDER; SOUZA, 2003); como quando se tem amostras destrutivas, como é o caso do palito de fósforo; ou amostras caras e frágeis, como antígenos utilizados em vacinas. Ainda na área da saúde, as pesquisas que dependem de voluntários - pessoas que possuam determinada doença grave ou mulheres gestantes, por exemplo, podem não apresentar um tamanho amostral suficiente (VIEIRA; HOSSNE, 2015).

Outras ocasiões em que também pode ser interessante aumentar uma amostra incluem o problema de Behens-Fisher (BEHRENS, 1929; FISHER, 1939), ou seja, quando se tem heterogeneidade das matrizes de covariâncias em distribuições normais multivariadas, e quando se

almeja aumentar o poder do teste F da análise de variância e de testes de comparação múltipla de médias.

Em todos esses cenários, com o aumento de dados, pode-se melhorar a inferência estatística, aumentando a precisão da estimação intervalar, reduzindo o erro padrão, aumentando o poder do teste, etc.

O aumento de dados refere-se a métodos para construção de otimizações iterativas ou algoritmos de amostragem, pela introdução de dados não observados ou variáveis latentes (VAN DYK; MENG, 2001).

Essa técnica foi inicialmente introduzida por Dempster, Laird e Rubin (1977), que propuseram um algoritmo determinístico utilizado para calcular estimativas de máxima verossimilhança de dados incompletos. Esse algoritmo foi denominado EM (*Expectation - Maximization*), pois consiste em dois passos:

1. Calcular o valor esperado do logaritmo natural da verossimilhança (log-verossimilhança) dos dados completos (passo E).

2. Maximizar a log-verossimilhança para obter o valor do parâmetro atualizado (passo M) (TANNER; WONG, 1987).

O algoritmo EM é caracterizado por apresentar uma infinidade de aplicações, dentre elas, Dempster, Laird e Rubin (1977) exemplificam a utilização desse algoritmo em métodos para a manipulação de dados faltantes, procedimentos adequados para dados arbitrariamente censurados e truncados, métodos de estimação para misturas finitas de famílias paramétricas, análise fatorial, técnicas de estimação robustas com base em mínimos quadrados iterativos reponderados, componentes de variância e estimação de hiperparâmetros de distribuições *a priori* na inferência bayesiana. Além disso, os autores ressaltam que existe uma gama de aplicações potenciais do algoritmo EM e deixam claro que mesmo em casos que parecem não ser um problema de dados incompletos, pode haver um lucro em utilizar essa ferramenta para facilitar a estimativa da máxima verossimilhança (TANNER; WONG, 1987).

Embora essa abordagem tenha sido inicialmente difundida com um algoritmo determinístico, é importante salientar que o termo aumento de dados (*data augmentation*) originou-se com Tanner e Wong (1987) num contexto bayesiano, com a implementação de um algoritmo estocástico que visa o cálculo da distribuição *a posteriori* dos parâmetros de interesse. Este algoritmo, nomeado DA - *Data Augmentation*, se baseia em aumentar um conjunto de dados

observados com a introdução de dados latentes. Assumindo uma amostra final, composta por dados observados e latentes, então a distribuição *a posteriori* pode ser analisada diretamente, ou seja, a distribuição *a posteriori* dos dados aumentados pode ser calculada (VAN DYK; MENG, 2001).

Então, a teoria do aumento de dados é fundamentada por um algoritmo determinístico (EM) e um algoritmo estocástico (DA), o que possibilita o emprego dessa metodologia em diferentes contextos. Embora a técnica seja comumente difundida na inferência bayesiana, e com a finalidade de inteirar dados incompletos, pode ser utilizada na inferência frequentista e aplicada a conjuntos de dados balanceados, em que o único objetivo é o aumento do tamanho amostral e consequente aumento de precisão na inferência praticada.

3 METODOLOGIA

Esta seção é composta pela descrição das propostas metodológicas de três algoritmos para o aumento de dados amostrais utilizando os arquétipos, e pela descrição de como os algoritmos foram avaliados e comparados quanto à sua eficiência através de dois estudos de simulação: o primeiro sobre o aumento de dados em amostras com diferentes tamanhos e diferentes números de variáveis, e o segundo sobre o aumento sucessivo de dados em amostras com um tamanho e número de variáveis específico. Além disso, foi feita a aplicação dos algoritmos em um conjunto de dados reais sobre análise sensorial de bebida láctea achocolatada.

3.1 ALGORITMOS PARA AUMENTO DE DADOS VIA ARQUÉTIPOS

Partindo de uma amostra aleatória p-variada $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ independente e identicamente distribuída, foram adicionadas variáveis latentes $\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_a$ calculadas pelos algoritmos propostos, em que a é o número de dados aumentados.

Selecionados os arquétipos $\boldsymbol{z}_k$ e seus coeficientes α_{ik} descritos anteriormente (Equação 2.2), os algoritmos foram programados para sortear coeficientes referentes a cada arquétipo, denominados α_{*k} em que $* \in \{1, 2, \ldots, n\}$. Em seguida, os coeficientes sorteados são multiplicados pelos respectivos arquétipos, resultando nos dados não observados, logo

$$\boldsymbol{w}_j = \sum_{k=1}^{K} \alpha_{*k} \times \boldsymbol{z}_k. \tag{3.1}$$

Os algoritmos 1 e 2 realizam o sorteio dos coeficientes considerando a exigência da combinação convexa e, dessa forma, possuem diferentes restrições para garantir que a soma dos coeficientes seja uma unidade. Já o algoritmo 3, não cumpre esse requisito, sendo uma alternativa mais simples computacionalmente. A seguir, tem-se o detalhamento de cada passo dos algoritmos.

3.1.1 Algoritmo 1

O primeiro algoritmo consiste dos seguintes passos, ilustrados pela Figura 1:

(i) Sortear um dos arquétipos $\boldsymbol{z}_k$, em que $k = 1, \ldots, K$;

(ii) Sortear um coeficiente α_{ik} referente ao primeiro arquétipo sorteado, sendo o coeficiente sorteado denominado α_{*k};

(iii) Calcular a diferença $d = 1 - \sum_{k=1}^{K} \alpha_{*k}$;

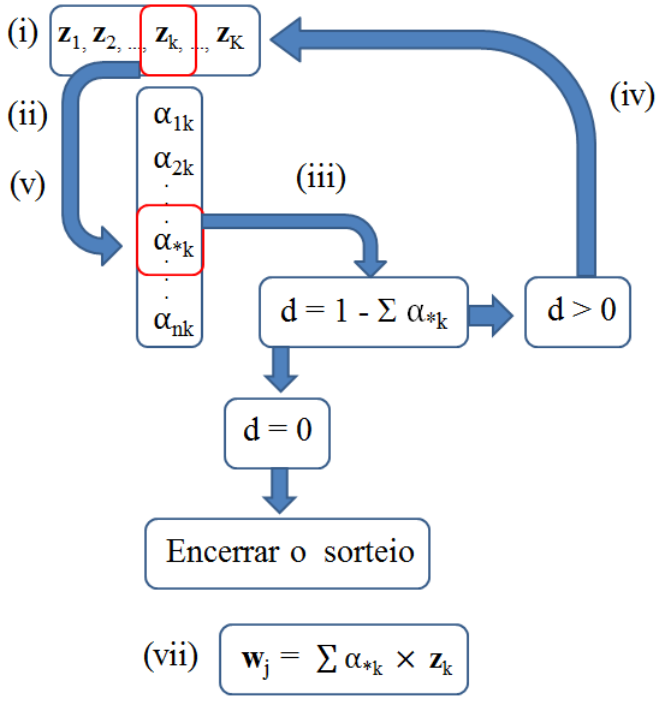
(iv) Se $d = 0$ e, consequentemente, $\sum_{k=1}^{K} \alpha_{*k} = 1$, encerrar o sorteio e zerar os coeficientes seguintes; se $d > 0$, sortear outro arquétipo dentre os restantes;

(v) Sortear um coeficiente referente ao arquétipo sorteado, desde que $\alpha_{*k} \leq d$;

(vi) Repetir os itens (iii) a (v) até chegar no último arquétipo (sendo o último coeficiente encontrado por diferença).

(vii) Multiplicar os coeficientes sorteados pelos respectivos arquétipos.

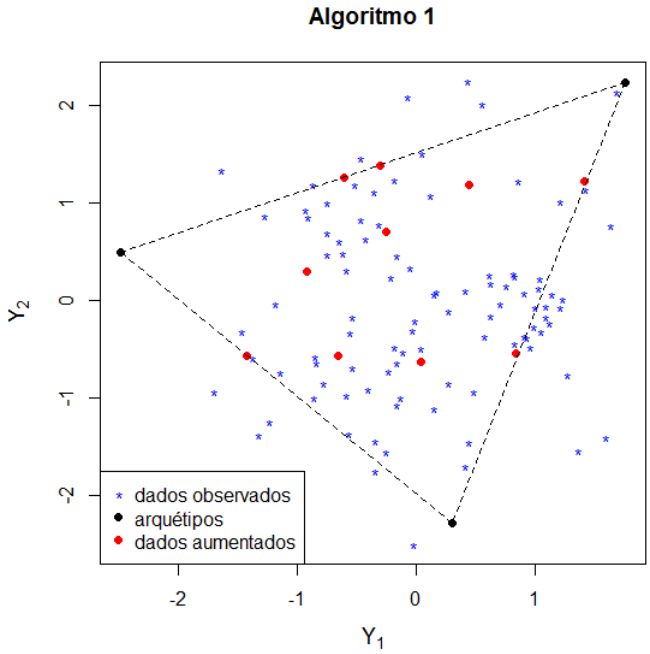
Figura 1 – Ilustração dos passos do Algoritmo 1.



Fonte: Da autora.

O resultado da utilização do Algoritmo 1 está exemplificado na Figura 2 por um conjunto de dados bivariados de tamanho inicial $n = 100$, em que foram selecionados três arquétipos, o que faz com que o fecho convexo seja um triângulo.

Figura 2 – Ilustração de dados aumentados pelo Algoritmo 1.



Fonte: Da autora.

Sendo assim, a partir dos arquétipos foram aumentados dez dados dentro do fecho convexo, respeitando a combinação convexa.

3.1.2 Algoritmo 2

O segundo algoritmo se baseia nos seguintes passos, ilustrados pela Figura 3:

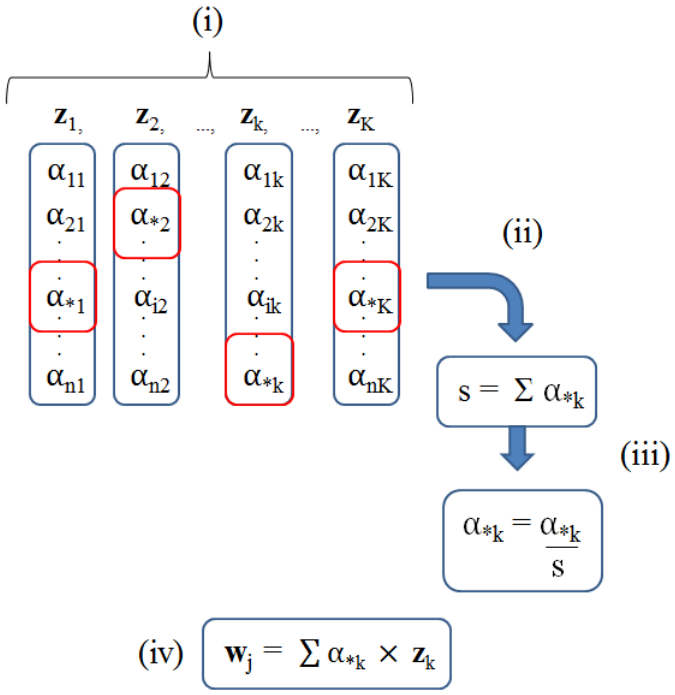
(i) Sortear um coeficiente referente a cada arquétipo;

(ii) Somar os coeficientes $s = \sum_{k=1}^{K} \alpha_{*k}$;

(iii) Fazer a proporção, ou seja, dividir cada coeficiente pela soma anterior s.

(iv) Multiplicar os coeficientes sorteados pelos respectivos arquétipos.

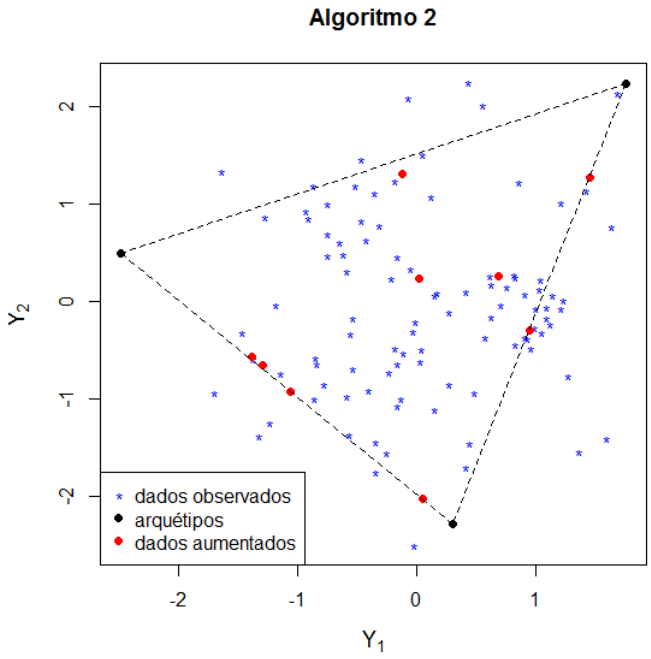
Figura 3 – Ilustração dos passos do Algoritmo 2.



Fonte: Da autora.

O resultado da utilização do Algoritmo 2 está exemplificado na Figura 4 também por um conjunto de dados bivariados de tamanho inicial $n = 100$ e com a seleção de três arquétipos.

Figura 4 – Ilustração de dados aumentados pelo Algoritmo 2.



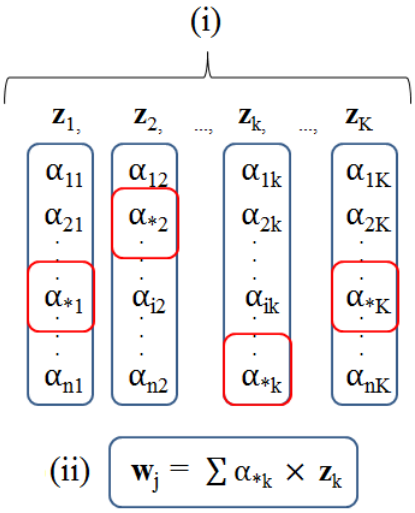
Fonte: Da autora.

Neste caso, os dez dados aumentados também se encontram dentro do fecho convexo devido à combinação convexa.

3.1.3 Algoritmo 3

O terceiro algoritmo, se resume a sortear um coeficiente referente a cada arquétipo e multiplicar os coeficientes sorteados pelos respectivos arquétipos, isto é, apenas os passos (i) e (iv) do Algoritmo 2, como pode-se observar na Figura 5.

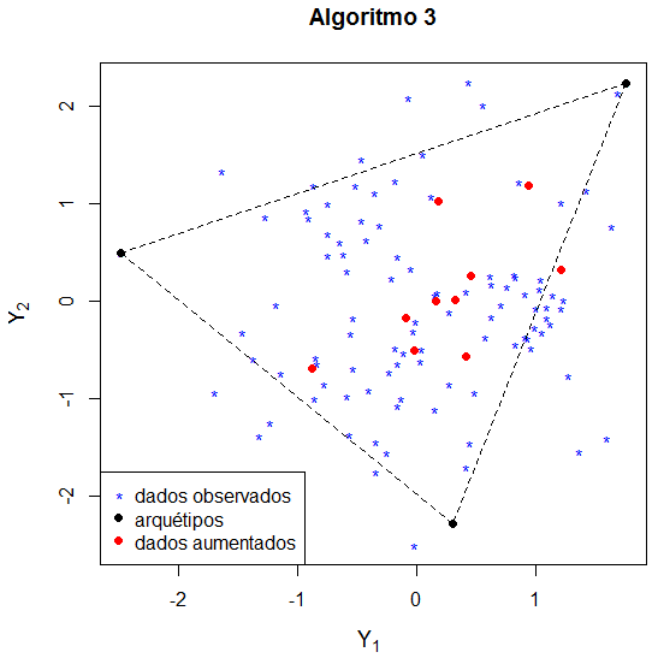
Figura 5 – Ilustração dos passos do Algoritmo 3.



Fonte: Da autora.

Na Figura 6 encontra-se o resultado da utilização do Algoritmo 3, novamente exemplificando-se por um conjunto de dados bivariados de tamanho inicial $n = 100$ e com a seleção de três arquétipos.

Figura 6 – Ilustração de dados aumentados pelo Algoritmo 3.



Fonte: Da autora.

Esse algoritmo não possui a restrição da combinação convexa, logo permite que sejam aumentados dados fora do fecho convexo como pode-se observar (FIGURA 6).

3.2 AVALIAÇÃO COMPUTACIONAL DOS ALGORITMOS

Com o intuito de comparar e avaliar a eficiência de cada algoritmo no aumento dos dados de uma amostra, foi realizado o seguinte estudo de simulação: a partir de um conjunto de dados com distribuição de probabilidade e parâmetros conhecidos, foram encontrados e selecionados seus arquétipos para, posteriormente, obter elementos não observados com os três algoritmos propostos.

Além da comparação dos algoritmos entre si, foi realizada ainda uma comparação destes com um controle positivo e uma testemunha. Esses métodos foram usados como referências e consistem em sortear elementos amostrais da seguinte forma:

Controle positivo: sorteia elementos conhecendo a distribuição de probabilidade da variável aleatória e o verdadeiro valor dos seus parâmetros;

Testemunha: inspirada no método de reamostragem *Bootstrap* não paramétrico, sorteia apenas elementos presentes na amostra.

Sendo assim, o controle positivo é o padrão ouro do estudo de simulação e a testemunha é um método simples utilizado como baliza dos resultados.

Ao final do estudo, foram obtidos conjuntos de dados aumentados por cada algoritmo e pelos métodos de referência, compostos pela amostra inicial mais um determinado número de dados aumentados, proporcional ao tamanho da amostra.

A distribuição de probabilidade adotada foi a normal multivariada com vetor de médias nulo e estrutura de covariância dada pela identidade. Fixada essa distribuição, foram avaliados: o tamanho da amostra (n = 5, 10, 30, 50 e 100 observações); o número de variáveis utilizadas (p = 2, 5, 10 e 20); e a proporção do aumento realizado (aumento = 10%, 30%, 50%, 80% e 100% de n) arredondada, quando necessário. É importante ressaltar que só foram realizados os cenários que possuem tamanho da amostra maior que o número de variáveis ($n > p$), o que proporciona um total de 75 cenários. Cada cenário foi simulado com 1000 repetições de Monte Carlo.

Para avaliar a eficiência dos algoritmos, do controle positivo e da testemunha, em cada cenário, foi testado se, após o aumento, a variável aleatória p-variada segue uma distribuição normal com vetor de médias nulo e com matriz de covariâncias identidade. Sendo assim, foram utilizados os testes multivariados: normalidade de Shapiro-Wilk generalizado por Royston

(1983), T^2 de Hotelling (HOTELLING, 1951) e de igualdade de matriz de covariâncias de Ledoit e Wolf (2002), considerando um nível de 5% de significância, com os respectivos pares de hipóteses:

$$\begin{cases} H_0 : \text{a variável aleatória } p\text{-variada segue distribuição de probabilidade normal.} \\ H_1 : \text{a variável aleatória } p\text{-variada não segue distribuição de probabilidade normal.} \end{cases}$$

$$\begin{cases} H_0 : \boldsymbol{\mu} = \emptyset \\ H_1 : \boldsymbol{\mu} \neq \emptyset \end{cases}$$

$$\begin{cases} H_0 : \boldsymbol{\Sigma} = \boldsymbol{I} \\ H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{I} \end{cases}$$

A escolha do teste de normalidade multivariada foi feita com base no trabalho de Korkmaz, Goksuluk e Zararsiz (2014), que apontam o teste de Royston como um dos mais amplamente utilizados e caracterizado por ser rigoroso. O teste sobre vetor de médias adotado (T^2 de Hotelling) é uma extensão do teste t univariado, que também já está bem consolidado na literatura (HAIR JR. et al., 2005; FERREIRA, 2011). E o teste de matriz de covariâncias de Ledoit e Wolf (2002) mostrou-se robusto quando testada a esfericidade, ou seja, se $\boldsymbol{\Sigma}$ possui covariâncias nulas e variâncias iguais, que é o caso da matriz identidade (FERREIRA, 2011).

A programação dos algoritmos e todas as análises foram executadas no software R (R CORE TEAM, 2016). Foram simuladas amostras normais com o auxílio do pacote *mvtnorm* (GENZ; BRETZ, 2009) e só foram utilizadas nas simulações as amostras consideradas normais, com vetor de médias nulo e matriz de covariâncias identidade, conformes os três testes utilizados, ao nível de 5% de significância.

Para a Análise de Arquétipos, utilizou-se o pacote *archetypes* (EUGSTER; LEISCH, 2009) e, portanto, foi executada através de funções prontas que não foram modificadas. A definição do número de arquétipos utilizados foi feita previamente com base no gráfico *scree plot* construído para cada amostra inicial. Sendo assim, foi fixado um número de arquétipos (na) para cada cenário, conforme o número de variáveis: na = 3, 6, 11 e 17 arquétipos, para p = 2, 5, 10 e 20 variáveis, respectivamente.

Finalmente, para a realização do teste T^2 de Hotelling foi utilizado o pacote *ICSNP* (NORDHAUSEN et al., 2015).

3.3 ESTUDO COMPUTACIONAL DE AUMENTOS SUCESSIVOS

A fim de estudar a realização de aumentos sucessivos com cada um dos três algoritmos, foi realizado um segundo estudo de simulação, seguindo os mesmos critérios do primeiro. Sendo assim, partindo de uma amostra inicial com distribuição de probabilidade e parâmetros conhecidos, foram realizados aumentos consecutivos e avaliadas: a distribuição de probabilidade da variável aleatória e as estimativas dos seus parâmetros na amostra aumentada.

Assim como no estudo anterior, foram realizados aumentos sucessivos também no controle positivo e na testemunha explicados previamente.

Novamente foi adotada a distribuição de probabilidade normal multivariada com vetor de médias nulo e matriz de covariâncias identidade, fixando-se agora um tamanho de amostra $n = 30$, $p = 2$ variáveis e um aumento final de $40 \times n$ (1200 dados aumentados) por aumentos sucessivos de 10%, 33%, 67% e 100% de n, ou seja, aumentos de 3, 10, 20 e 30 dados. Desta forma, com cada algoritmo, com o controle positivo e com a testemunha, foram simulados quatro cenários:

Cenário 1: aumentos sucessivos de 10% de n;

Cenário 2: aumentos sucessivos de 33% de n;

Cenário 3: aumentos sucessivos de 67% de n;

Cenário 4: aumentos sucessivos de 100% de n.

Para avaliar a eficiência dos algoritmos, do controle positivo e da testemunha, em cada cenário, foi testado se, após os aumentos consecutivos, a variável aleatória bivariada seguia distribuição normal com vetor de médias nulo e com matriz de covariâncias identidade. E para isso, foram utilizados os mesmos testes multivariados: normalidade de Shapiro-Wilk generalizado por Royston (1983), T^2 de Hotelling (HOTELLING, 1951) e de igualdade de matriz de covariâncias de Ledoit e Wolf (2002), considerando 5% de significância.

Os aumentos foram realizados manuseando-se cada um dos três algoritmos de duas formas:

1. utilizando os arquétipos da amostra inicial em todos os aumentos realizados sucessivamente;

2. utilizando os arquétipos da amostra inicial no primeiro aumento e recalculando os arquétipos a cada nova amostra aumentada.

Sendo assim, os algoritmos manuseados conforme o item 1 foram denominados A1_i, A2_i e A3_i, referentes aos algoritmos 1, 2 e 3, respectivamente. E os algoritmos manuseados de acordo com o item 2, denominados A1_a, A2_a e A3_a, respectivamente.

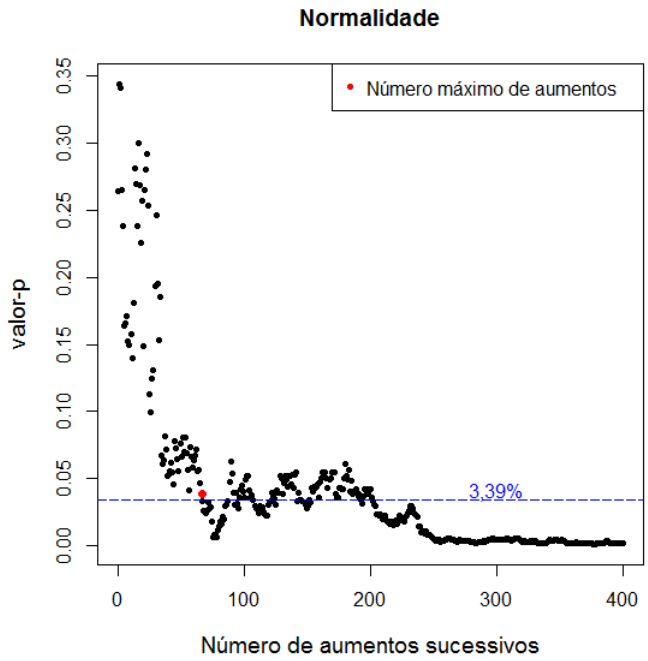
É importante explicar também que os aumentos realizados com a testemunha foram análogos ao item 2, isto é, no primeiro aumento foram sorteados elementos da amostra inicial e, nos aumentos seguintes, foram sorteados elementos pertencentes às amostras aumentadas.

Neste estudo, cada cenário foi simulado com 100 repetições de Monte Carlo. Foi necessário utilizar menos repetições do que o desejado, porque recalcular os arquétipos a cada aumento sucessivo demandou muito tempo, fato que não ocorreu ao realizar todos os aumentos consecutivos com os arquétipos da amostra inicial. Desta forma, em cada simulação, eram verificados os valores-p dos testes realizados a cada aumento sucessivo, a fim de avaliar a partir de qual aumento a distribuição da variável aleatória deixava de ser a normal multivariada e a partir de qual momento as estimativas dos parâmetros eram diferentes do verdadeiro valor dos parâmetros.

Como o nível de significância adotado foi 5%, foi definido como o número máximo de aumentos sucessivos o número anterior ao primeiro que proporcionou valor-p dos testes realizados abaixo do limite inferior do intervalo de confiança binomial exato $IC_{99\%}(p)$ = [3,39%; 7,05%], ou seja o número antecessor ao que obteve valor-p abaixo de 3,39%.

Na Figura 7 tem-se um exemplo da definição do número máximos de aumentos consecutivos considerando o teste de normalidade de Royston.

Figura 7 – Gráfico de dispersão que relaciona os valores-p do teste de Royston (Normalidade) conforme o número de aumentos sucessivos realizados.



Fonte: Da autora.

Pode-se observar, neste exemplo, que o número máximo de aumentos sucessivos estabelecido para garantir distribuição normal encontra-se em torno de 70, mesmo que com aproximadamente 100 aumentos a distribuição de probabilidade tenha voltado a ser a normal multivariada (FIGURA 7). Esse padrão foi mantido em todos os cenários com todos os algoritmos e métodos de referências e para todos os testes realizados.

Ao final, foram obtidas sequências de 100 números máximos de aumentos, referentes às 100 repetições de Monte Carlo. Como a variável resposta número máximo de aumentos sucessivos é quantitativa discreta e as sequências apresentaram valores discrepantes e não foram unimodais, foi escolhida a medida de posição mediana para apresentação dos resultados, bem como seu intervalo de confiança.

Assim como no primeiro estudo computacional, a programação dos algoritmos e todas as análises também foram executadas no software R (R CORE TEAM, 2016), utilizando-se dos mesmos pacotes: *mvtnorm* (GENZ; BRETZ, 2009), *archetypes* (EUGSTER; LEISCH, 2009) e *ICSNP* (NORDHAUSEN et al., 2015).

3.4 ESTUDO COM DADOS REAIS

Em um experimento sobre análise sensorial de bebida láctea achocolatada, a marca comercial Toddy foi submetida à Análise Descritiva Quantitativa por 17 provadores recrutados entre os alunos que cursavam a disciplina "Análise Sensorial de Alimentos e Bebidas", na Universidade Federal de Alfenas, em Maio de 2015, os quais receberam um breve treinamento durante a disciplina.

Foram analisados os termos descritivos: aroma de chocolate e sabor de chocolate, ou seja, o aroma e o sabor associados ao chocolate em pó solúvel. A ficha de avaliação escolhida foi a de escala não-estruturada (variando de a 0 a 10 cm), permitindo quantificar, de forma contínua, a intensidade desses atributos sensoriais.

Com os dados coletados, primeiramente foi verificado se a variável aleatória em estudo seguia distribuição normal multivariada através do teste de Royston (1983) e, na sequência, foram estimados o seu vetor de médias e sua matriz de covariâncias amostral.

Então, foi realizado o aumento de dados com o algoritmo 3, proposto nesse trabalho, de duas maneiras, seguindo as recomendações dos dois estudos computacionais: realizando um aumento de dois dados (aproximadamente 10% de n) com A3; e realizando 11 aumentos sucessivos de 2 dados (aproximadamente 130% de n ao final) com A3$_i$.

Por fim, foram estimados pontualmente o vetor de médias e a matriz de covariâncias das amostras aumentadas e foram realizados os testes multivariados de Royston, T^2 de Hotelling e de Box (1949), a fim de testar a distribuição de probabilidade e comparar o vetor de médias e a matriz de covariâncias da amostra inicial com os das amostras aumentadas.

4 RESULTADOS E DISCUSSÃO

Os resultados dos 75 cenários do estudo de simulação foram organizados em cinco tabelas. Na primeira tabela encontram-se os resultados do controle positivo em todos os cenários (TABELA 1), já nas tabelas seguintes, encontram-se os resultados obtidos com os algoritmos 1, 2 e 3 (A1, A2 e A3, respectivamente) e com a testemunha (T), sendo uma tabela para cada número de variáveis (p = 2, 5, 10 e 20 variáveis). Os resultados apresentados são referentes à proporção de rejeição (%) da hipótese nula de cada teste realizado nas amostras aumentadas. Como foi adotado o nível de 5% de significância, consideram-se ideais os resultados dentro do intervalo de confiança binomial exato $IC_{99\%}(p)$: [3,39%; 7,05%].

Tabela 1 – Proporção (%) de rejeição da hipótese de que a variável aleatória segue distribuição normal (N), com vetor de médias nulo (M) e matriz de covariâncias identidade (C), após o aumento com o controle positivo, nos cenários com p = 2, 5, 10 e 20 variáveis.

		$p = 2$			$p = 5$			$p = 10$			$p = 20$		
n	aumento	N	M	C	N	M	C	N	M	C	N	M	C
	10%	1,9	2,6	1,3	..	..	..	..	..	..	..	..	..
	30%	2,8	3,1	2,2	..	..	..	..	..	..	..	..	..
5	50%	3,3	5,1	1,8	..	..	..	..	..	..	..	..	..
	80%	3,8	3,6	2,4	..	..	..	..	..	..	..	..	..
	100%	2,9	4,9	4,2	..	..	..	..	..	..	..	..	..
	10%	1,3	2,1	1,0	1,8	1,3	2,3	..	..	..	..	..	..
	30%	2,0	3,6	1,6	2,1	2,8	3,5	..	..	..	..	..	..
10	50%	3,1	3,0	2,9	2,9	3,8	3,3	..	..	..	..	..	..
	80%	2,1	3,4	2,3	3,4	3,9	3,7	..	..	..	..	..	..
	100%	4,3	5,4	2,7	4,3	4,0	3,5	..	..	..	..	..	..
	10%	1,3	2,4	1,4	2,1	1,6	1,3	2,0	2,1	1,8	2,0	3,6	3,0
	30%	2,9	1,6	2,9	1,9	2,6	3,1	3,4	1,7	2,8	3,7	4,4	3,8
30	50%	3,1	3,0	3,8	3,5	3,7	3,9	3,3	4,1	3,5	2,8	3,8	4,5
	80%	3,2	4,3	3,8	3,6	3,1	3,1	2,8	5,0	4,5	3,6	4,0	5,2
	100%	3,3	3,8	3,9	3,1	3,6	4,5	3,8	3,7	3,8	3,7	4,3	4,9
	10%	1,2	1,4	0,9	1,1	1,8	1,5	1,3	2,2	1,4	1,7	2,8	2,4
	30%	2,9	2,7	3,0	2,3	2,9	3,0	3,8	2,3	2,2	2,1	3,6	3,8
50	50%	3,6	3,2	4,0	2,3	1,7	3,2	3,0	3,3	3,5	3,2	4,4	4,3
	80%	2,7	2,0	3,4	3,4	3,9	2,5	3,4	3,5	3,6	3,9	3,7	6,1
	100%	2,8	4,0	3,7	4,4	3,5	3,7	3,6	3,1	4,3	3,5	3,0	4,6
	10%	1,3	1,8	2,3	2,1	2,0	1,6	1,5	1,8	1,6	1,7	2,3	2,7
	30%	2,7	1,9	2,3	2,1	2,1	2,9	2,1	2,7	3,2	2,8	2,3	3,7
100	50%	2,6	3,2	2,3	2,6	3,0	2,5	2,8	2,6	2,3	2,7	4,1	3,7
	80%	3,0	3,6	2,8	3,5	2,6	4,3	3,4	4,1	4,3	3,7	3,5	3,4
	100%	3,8	3,3	3,9	3,8	3,7	3,4	3,5	4,2	3,2	4,3	5,4	4,4

Fonte: Da autora.

Nota: Sinal convencional utilizado:
.. Não se aplica dado numérico.

Na Tabela 1 verifica-se que o controle positivo permite aumentar a amostra inicial sem alterar a distribuição normal de probabilidade e nem as estimativas de seus parâmetros, pois apresentou baixíssimas taxas de rejeição de H_0 em todos os cenários, entre 0,9% e 6,1%. Pode-se notar também que os três testes foram conservadores, pois foram observadas proporções de rejeição abaixo de 3,39% muitas vezes.

Esses resultados confirmam a propensão do controle positivo a padrão ouro desse estudo e, portanto, optou-se por subtrair a proporção de rejeição deste controle em todos os resultados de A1, A2, A3 e T, como forma de corrigir um erro aleatório cometido por todos os algoritmos e pela testemunha e também verificar o quanto eles rejeitam H_0 a mais que o controle positivo. Dessa forma, em cada cenário e para cada teste, foi feita a seguinte subtração

$$\hat{p}^*_m = \hat{p}_m - \hat{p}_{cp}, \forall\ m \in \{A1; A2; A3; T\}$$

em que $\hat{p}_m$ é a proporção de rejeição de H_0 obtida pelo método m, $\hat{p}_{cp}$ é a proporção de rejeição de H_0 obtida pelo controle positivo e $\hat{p}^*_m$ a proporção de rejeição de H_0 do método m corrigida. Essa correção também foi realizada no intervalo de confiança para o nível de significância estabelecido (5%), também corrigido pela proporção de rejeição obtida pelo controle positivo em cada cenário, isto é,

$$IC_{99\%}(p)^* = IC_{99\%}(p) - \hat{p}_{cp},$$

em que $IC_{99\%}(p)^*$ é o intervalo de confiança corrigido. Desta forma, foram destacadas em todas as tabelas seguintes (em cinza) as proporções de rejeição de H_0 corrigidas ($\hat{p}^*_m$) inferiores ao limite superior do $IC_{99\%}(p)^*$ para cada cenário.

Na Tabela 2, tem-se a proporção corrigida de rejeição das hipóteses nulas nos cenários bivariados. Nesta, verifica-se que os três algoritmos propostos apresentaram melhor desempenho que a testemunha, quando testada a distribuição de probabilidade da variável aleatória; destacando-se o A3, que chegou a apresentar proporção de rejeição inferior à do controle positivo em alguns cenários.

Tabela 2 – Proporção corrigida (%) de rejeição das hipóteses nulas de que a variável aleatória segue distribuição normal (Normalidade), com vetor de médias nulo (Médias) e matriz de covariâncias identidade (Covariâncias), após o aumento com os algoritmos 1, 2 e 3 (A1, A2 e A3) e com a testemunha (T), nos cenários de $p = 2$ variáveis.

		Normalidade				Médias				Covariâncias			
n	aumento	A1	A2	A3	T	A1	A2	A3	T	A1	A2	A3	T
	10%	5,7	4,7	1,5	8,9	5,2	5,2	2,9	5,3	0,9	1,0	1,5	2,1
	30%	13,2	8,5	4,8	25,2	13,0	13,0	10,3	13,1	2,2	1,6	0,8	3,0
5	50%	25,5	16,9	8,6	46,3	20,3	18,8	15,5	17,9	3,8	3,9	2,9	5,7
	80%	37,4	22,5	10,9	62,1	30,0	29,8	24,0	25,3	6,1	6,0	4,4	7,9
	100%	49,7	28,7	13,5	77,4	35,7	36,3	29,0	31,1	10,7	9,3	6,7	11,0
	10%	2,0	1,1	1,2	4,0	0,7	1,1	0,4	0,5	0,7	1,1	0,2	1,5
	30%	5,8	5,0	1,9	19,6	7,1	6,6	6,8	7,1	2,6	2,7	2,5	3,5
10	50%	10,7	7,5	1,5	39,8	15,9	16,5	14,5	14,1	7,7	5,8	5,1	8,0
	80%	22,4	17,4	6,1	65,7	30,6	25,9	24,7	21,3	14,8	14,8	14,1	13,5
	100%	29,5	17,2	3,7	75,5	34,0	30,8	33,1	23,3	20,1	21,5	19,1	20,6
	10%	1,0	1,2	1,1	5,2	2,4	1,9	1,8	1,7	1,4	1,0	0,7	1,8
	30%	3,4	1,2	-0,0	20,9	11,2	9,6	9,1	7,1	5,1	4,9	4,4	6,4
30	50%	6,9	3,6	-0,0	41,8	22,7	19,9	20,4	13,8	14,8	14,5	13,9	13,5
	80%	18,0	10,5	0,8	68,3	37,1	32,1	31,1	19,7	28,4	27,5	32,1	20,5
	100%	25,8	16,7	1,1	78,2	48,4	38,4	40,0	24,5	36,4	37,0	43,7	26,9
	10%	0,6	0,8	-0,0	4,8	2,6	2,8	1,7	1,8	2,9	1,7	2,6	1,9
	30%	3,7	1,3	-0,0	23,1	13,3	9,2	12,2	9,1	9,0	7,7	7,6	7,7
50	50%	7,6	5,2	-0,0	43,9	26,8	19,7	21,2	14,7	19,9	17,5	22,0	13,4
	80%	19,6	11,5	0,1	71,9	44,9	33,0	37,4	17,7	34,8	34,1	44,2	23,1
	100%	28,6	16,0	2,2	79,0	54,9	42,8	46,5	23,6	46,4	43,4	54,8	28,6
	10%	0,6	0,7	-0,0	5,3	3,2	3,4	3,4	2,0	2,6	1,1	1,6	2,8
	30%	5,1	2,3	-0,0	24,8	17,1	11,8	13,5	8,3	11,6	10,4	13,6	10,3
100	50%	11,2	6,1	-0,0	47,6	33,4	24,2	25,2	12,6	28,1	24,8	37,4	18,8
	80%	25,9	15,8	0,4	70,8	54,7	41,7	42,4	22,5	47,3	44,0	65,8	24,7
	100%	37,1	22,6	0,7	79,9	63,5	47,0	50,9	23,2	58,2	54,8	77,4	29,4

Fonte: Da autora.

Notas: Destacadas em cinza as proporções abaixo do limite superior do $IC_{99\%}(p)^*$.
Sinal convencional utilizado:
-0,0 Dado numérico igual a zero resultante de arredondamento de um dado numérico originalmente negativo.

Avaliando a proporção do aumento realizado, tem-se que o A3 permitiu aumentar até 100% do tamanho da amostra inicial quando $n \geq 30$, mantendo a distribuição de probabilidade normal, isto é, com proporções de rejeição de H_0 inferiores ao limite superior do $IC_{99\%}(p)^*$, e rejeitando no máximo 2,2% a mais que o controle positivo. Quando $n = 10$, permitiu aumentar até 50% e, quando $n = 5$, 10%. Com A2, foi possível aumentar somente 10% quando $n = 5$, 30% quando $n = 10$, 50 e 100, e até 50% quando $n = 30$. Já com A1, foi possível aumentar em 30% da amostra inicial quando $n = 30$ e 50, apenas 10% quando $n = 10$ e 100, e não foi possível realizar nenhum aumento quando $n = 5$ (TABELA 2).

Os resultados dos testes sobre o vetor de médias em geral foram semelhantes entre os três algoritmos e a testemunha. Considerando o $IC_{99\%}(p)^*$, foi possível aumentar a amostra inicial em 10% em quase todos os cenários, excetuando-se os cenários com $n = 5$, no qual foi possível realizar esse aumento de 10% apenas com A3, sem alterar a estimativa do vetor de médias (TABELA 2).

Nos testes sobre a matriz de covariâncias também foram observados resultados próximos entre A1, A2, A3 e T, permitindo aumentos entre 10 e 80%, dependendo do tamanho amostral (TABELA 2). Pode-se notar que, conforme n aumenta, a proporção do aumento deve ser menor para que não se altere a estimativa da matriz de covariâncias. Assim, quando $n = 5$, foi possível aumentar 50% de n com A1 e A2, e 80% de n com A3; quando $n = 10$, foi possível aumentar 30% com todos os métodos comparados, e quando $n \geq 30$, apenas 10% de aumento. Essas porcentagens levam a aumentos com A3, respectivamente, de 4, 3, 3, 5 e 10 dados na amostra inicial, ou seja, pôde-se inserir um número parecido de elementos independente do tamanho da amostra inicial.

Sendo assim, em conjuntos de dados com 2 variáveis, pode-se realizar aumento de 10% do tamanho da amostra inicial com os algoritmos sem alterar, simultaneamente, a distribuição de probabilidade normal da variável aleatória, nem as estimativas de $\boldsymbol{\mu}$ e de $\boldsymbol{\Sigma}$. Ressaltando-se que quando $n = 5$ é necessário utilizar A3.

A seguir, podem ser observados os resultados obtidos com os cenários que utilizaram 5 variáveis. Mais uma vez nota-se que os três algoritmos propostos apresentaram melhor desempenho que a testemunha quando avaliada a distribuição de probabilidade. Enquanto T não permitiu aumentos em nenhum cenário, sem alterar a distribuição de probabilidade, os três algoritmos garantiram 10% em todos eles. Destaca-se o A3, que permitiu aumentar 30% do tamanho inicial quando $n = 30$ (TABELA 3).

Tabela 3 – Proporção corrigida (%) de rejeição das hipóteses nulas de que a variável aleatória segue distribuição normal (Normalidade), com vetor de médias nulo (Médias) e matriz de covariâncias identidade (Covariâncias), após o aumento com os algoritmos 1, 2 e 3 (A1, A2 e A3) e com a testemunha (T), nos cenários de $p = 5$ variáveis.

		Normalidade				Médias				Covariâncias			
n	aumento	A1	A2	A3	T	A1	A2	A3	T	A1	A2	A3	T
	10%	4,1	3,4	1,7	7,1	3,5	3,7	1,4	3,4	3,8	2,5	4,5	4,5
	30%	18,3	14,0	5,2	36,8	17,6	16,6	9,6	15,9	17,0	12,9	17,0	20,7
10	50%	39,4	21,6	11,7	69,9	34,5	31,3	22,0	30,5	34,3	27,8	30,5	35,9
	80%	63,7	40,3	22,1	91,1	59,7	52,7	41,9	45,6	58,8	53,7	57,0	55,1
	100%	73,9	49,5	29,0	94,1	70,6	63,4	50,6	55,6	74,1	69,8	69,3	67,4
	10%	2,2	1,2	0,5	7,8	3,4	3,2	2,5	2,8	4,2	4,2	2,1	7,8
	30%	10,3	7,1	4,8	42,3	22,0	18,0	16,8	11,8	29,2	22,7	24,1	24,2
30	50%	25,4	14,0	10,9	73,2	43,2	32,3	31,5	23,9	54,8	50,6	50,9	42,9
	80%	51,0	24,9	23,9	93,4	69,5	57,8	56,4	38,9	88,5	83,0	85,5	64,9
	100%	60,3	34,1	31,2	95,8	81,7	70,0	68,1	49,2	92,6	89,9	89,4	74,4
	10%	2,0	1,0	0,2	9,4	4,5	3,4	2,7	2,0	4,8	4,5	4,6	6,4
	30%	12,4	6,9	5,6	45,5	28,5	22,1	20,4	13,4	32,7	29,7	30,3	26,9
50	50%	30,6	16,4	15,0	80,4	55,4	40,1	35,8	24,8	70,5	64,5	67,1	47,6
	80%	58,6	32,6	30,1	93,4	78,5	60,4	60,8	39,4	95,8	93,2	93,1	67,6
	100%	69,5	40,1	42,1	94,9	86,8	72,5	70,9	45,3	95,8	95,1	95,6	73,9
	10%	1,8	0,7	0,8	10,4	5,4	4,1	4,0	2,9	6,1	5,6	3,9	7,4
	30%	18,3	10,3	9,1	51,4	40,3	23,3	24,2	14,2	55,8	46,7	49,3	28,0
100	50%	41,1	21,1	22,0	77,2	68,0	47,1	45,8	21,8	89,3	83,8	86,1	47,7
	80%	74,3	47,4	51,2	93,7	89,7	72,2	71,3	37,4	95,7	95,2	95,6	67,0
	100%	84,6	64,6	68,1	95,8	92,1	80,1	79,0	45,1	96,6	96,6	96,6	76,8

Fonte: Da autora.

Nota: Destacadas em cinza as proporções abaixo do limite superior do $IC_{99\%}(p)^*$.

Ao avaliar os resultados dos testes sobre o vetor de médias e sobre a matriz de covariâncias, verifica-se novamente uma semelhança entre a proporção de rejeição dos três algoritmos e da testemunha. De acordo com ambos os testes, pôde-se aumentar em 10% do tamanho amostral sem prejuízos nas estimativas dos parâmetros com todos os métodos comparados, quando $n = 10$ e 50; no outros cenários, exclui-se a testemunha, pois apresentou proporção de rejeição da hipótese de esfericidade acima do limite superior do $IC_{99\%}(p)^*$. Além disso, ressalta-se que, quando $n = 100$, somente o A3 permitiu aumentar 10% de n sem alterar as estimativas de $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ simultaneamente (TABELA 3).

Portanto, quando um estudo possui 5 variáveis e $5 \leq n \leq 50$, pode-se aumentar n em 10% com os três algoritmos mantendo, ao mesmo tempo, a distribuição de probabilidade normal da variável aleatória e as estimativas de seus parâmetros. Já quando $n = 100$, deve-se utilizar A3 para obter tal resultado.

É importante salientar que com o aumento no número de variáveis, vai se tornando cada vez mais difícil aumentar a amostra inicial sem interferir tanto na distribuição de probabilidade,

quanto nas estimativas de seus parâmetros, assim, quando $p = 5$, aumentos a partir de 50% de n, geraram taxas de rejeição das hipóteses nulas cada vez maiores (TABELA 3).

Os resultados dos cenários que utilizaram amostras com 10 variáveis encontram-se na Tabela 4. A hipótese de que a variável aleatória segue distribuição normal foi rejeitada em proporção equivalente nos três algoritmos, sendo esta proporção inferior à da testemunha. A1, A2 e A3 garantiram aumentar 10% de n sem interferir na distribuição de probabilidade, ao passo que, com esse aumento, T já apresentou entre 16,6 e 19% de rejeição de H_0 a mais que o controle positivo.

Tabela 4 – Proporção corrigida (%) de rejeição das hipóteses nulas de que a variável aleatória segue distribuição normal (Normalidade), com vetor de médias nulo (Médias) e matriz de covariâncias identidade (Covariâncias), após o aumento com os algoritmos 1, 2 e 3 (A1, A2 e A3) e com a testemunha (T), nos cenários de $p = 10$ variáveis.

		Normalidade				Médias				Covariâncias			
n	aumento	A1	A2	A3	T	A1	A2	A3	T	A1	A2	A3	T
30	10%	3,7	2,5	1,8	16,6	5,5	5,0	4,0	4,3	14,7	12,9	17,8	23,3
	30%	29,5	10,9	14,4	70,2	36,7	31,4	22,5	23,5	64,9	55,6	63,3	65,3
	50%	64,2	25,4	36,2	92,9	64,7	52,6	44,5	41,8	92,5	88,4	89,5	86,4
	80%	91,2	51,0	67,9	97,2	86,9	79,0	69,3	62,4	95,4	95,5	95,4	94,6
	100%	94,4	59,9	82,2	96,2	93,8	84,9	79,7	71,0	96,2	96,2	96,2	95,6
50	10%	2,4	1,2	1,6	16,6	6,6	6,3	5,0	5,1	13,5	11,9	13,4	19,3
	30%	30,8	12,6	19,8	71,0	40,4	31,0	27,2	24,1	74,2	62,8	69,1	65,2
	50%	68,8	28,8	47,4	94,2	72,8	54,2	49,0	38,7	95,4	93,7	94,8	87,0
	80%	93,5	58,8	82,0	96,5	92,4	81,1	75,4	58,7	96,4	96,4	96,4	95,7
	100%	95,5	74,0	92,6	96,4	96,3	89,3	86,5	69,8	95,7	95,7	95,7	95,3
100	10%	4,1	1,8	3,4	19,0	7,9	6,5	5,7	4,3	18,4	14,1	15,7	20,9
	30%	48,5	21,8	35,2	73,4	53,3	34,5	35,6	22,3	91,3	84,2	88,2	65,8
	50%	90,0	57,6	77,6	94,8	85,3	65,2	63,8	38,9	97,7	97,7	97,7	89,5
	80%	96,4	88,3	96,0	96,6	95,2	87,5	87,1	56,8	95,7	95,7	95,7	94,7
	100%	96,5	94,1	96,3	96,5	95,6	92,1	92,0	65,9	96,8	96,8	96,8	96,6

Fonte: Da autora.

Nota: Destacadas em cinza as proporções abaixo do limite superior do $IC_{99\%}(p)^*$.

A proporção de rejeição da hipótese de que a amostra aumentada possui vetor de médias nulo foi parecida nos três algoritmos e na testemunha, contudo, dentre os algoritmos, somente A3 permitiu aumentar a amostra inicial em 10% e apenas quando $n = 10$, mantendo as estimativas do vetor de médias (TABELA 4).

Os testes sobre a matriz de covariâncias indicaram resultados semelhantes em todos os métodos de aumento de dados, com taxas de rejeição de H_0 acima do limite superior do $IC_{99\%}(p)^*$ já nos aumentos de 10% de n, chegando às porcentagens máximas de rejeição ao aumentar 50% de n ou mais. Quando $n = 100$, por exemplo, as taxas foram de 97,7%, 95,7% e 96,8%, respectivamente, ou seja, 100% corrigida pela proporção de rejeição obtida pelo controle

positivo: 2,3%, 4,3% e 3,2%, respectivamente (TABELAS 4 e 1). Logo, quando $p = 10$, não foi possível aumentar a amostra inicial sem interferir na estimativa da matriz de covariâncias.

Desta forma, quando trata-se de dados com 10 variáveis e $n = 10$, pode-se realizar aumento de 10% do tamanho da amostral apenas com A3 sem alterar, simultaneamente, a distribuição de probabilidade normal da variável aleatória e a estimativa de $\boldsymbol{\mu}$. Contudo, a estimativa de $\boldsymbol{\Sigma}$ pode não ser a mesma.

Por fim, na Tabela 5 estão apresentados os resultados obtidos nos cenários com p = 20 variáveis. Com estes resultados, fica consolidada a superioridade dos algoritmos frente à testemunha, quando testada a hipótese nula de que a variável aleatória segue distribuição normal. Contudo, somente A2 e A3 apresentaram proporção de rejeição em concordância com o nível de significância adotado ao aumentar 10% do tamanho amostral. Quando $n = 50$, ambos algoritmos permitiram realizar esse aumento, já com $n = 100$, foi possível apenas com A2.

Tabela 5 – Proporção corrigida (%) de rejeição das hipóteses nulas de que a variável aleatória segue distribuição normal (Normalidade), com vetor de médias nulo (Médias) e matriz de covariâncias identidade (Covariâncias), após o aumento com os algoritmos 1, 2 e 3 (A1, A2 e A3) e com a testemunha (T), nos cenários de $p = 20$ variáveis.

		Normalidade				Médias				Covariâncias			
n	aumento	A1	A2	A3	T	A1	A2	A3	T	A1	A2	A3	T
	10%	10,6	6,1	5,9	30,0	10,4	10,0	3,3	8,9	41,4	34,3	43,6	50,9
	30%	67,1	26,0	32,2	91,0	55,2	50,7	29,8	44,8	95,8	93,6	94,7	95,4
30	50%	95,2	50,8	72,0	97,0	86,1	81,2	60,6	73,3	95,5	95,5	95,5	95,5
	80%	96,4	80,5	93,2	96,4	95,5	94,4	85,0	89,4	94,8	94,8	94,8	94,8
	100%	96,3	85,9	95,5	96,3	95,7	94,9	92,6	93,5	95,1	95,1	95,1	95,1
	10%	8,0	2,4	4,4	29,3	10,7	9,0	6,1	7,8	40,6	34,1	43,3	53,3
	30%	70,3	26,9	47,2	93,4	59,1	48,9	37,2	39,3	95,7	95,2	95,7	95,4
50	50%	95,6	54,9	86,6	96,8	87,1	77,8	67,4	61,9	95,7	95,7	95,7	95,7
	80%	96,1	83,5	96,0	96,1	96,0	94,0	91,2	85,4	93,9	93,9	93,9	93,9
	100%	96,5	93,1	96,5	96,5	96,9	96,2	94,7	90,5	95,4	95,4	95,4	95,4
	10%	8,9	3,5	7,8	32,8	12,0	9,7	8,5	7,9	42,3	34,5	42,0	52,9
	30%	87,1	51,2	76,2	93,0	70,4	54,8	48,4	37,5	96,3	96,1	96,1	95,1
100	50%	97,3	92,9	97,0	97,3	92,3	82,5	80,1	58,6	96,3	96,3	96,3	96,3
	80%	96,3	96,3	96,3	96,3	96,5	95,4	95,3	79,6	96,6	96,6	96,6	96,6
	100%	95,7	95,7	95,7	95,7	94,6	94,4	94,2	85,9	95,6	95,6	95,6	95,6

Fonte: Da autora.

Nota: Destacadas em cinza as proporções abaixo do limite superior do $IC_{99\%}(p)^*$.

Avaliando os resultados sobre o vetor de médias da amostra aumentada, tem-se resultados parecidos entre todos os métodos de aumento de dados, porém somente A3 permitiu aumentar a amostra inicial em 10% quando $n = 30$, pois foi o único que apresentou proporção de rejeição dentro do $IC_{99\%}(p)^*$ estabelecido (TABELA 5).

Os resultados dos testes sobre a matriz de covariâncias novamente foram parecidos entre

os algoritmos e a testemunha e com taxas de rejeição máximas a partir de aumentos de 30% da amostra inicial, quando $n \geq 50$. Confirmando que, com $p \geq 10$, os métodos avaliados nesse estudo não permitiram realizar aumento do tamanho amostral sem alterar a estimativa da matriz de covariâncias (TABELA 5).

Logo, em conjuntos de dados com 20 variáveis não é possível fazer nenhum recomendação simultânea sobre a distribuição de probabilidade e as estimativas de seus parâmetros. Pode-se aumentar a amostra inicial em 10% com A2 quando $n = 50$ e 100, mantendo a distribuição de probabilidade e realizar o mesmo aumento com A3, quando $n = 30$, mantendo a estimativa do vetor de médias. Novamente, não se pode garantir que a estimativa de $\mathbf{\Sigma}$ seja a mesma.

Face ao exposto, o estudo de simulação sugere que os três algoritmos propostos apresentaram resultados semelhantes entre si, sendo observado um melhor desempenho dos Algoritmos 2 e 3 nos testes multivariados realizados. Dentre eles, destaca-se o Algoritmo 3 que mostrou-se competente em todos os cenários, principalmente nos bivariados e na conservação da distribuição de probabilidade normal, em que permite aumentos de até 100% de n.

Com o estudo de simulação verificou-se que o aumento de dados via arquétipos é eficaz, permitindo aumentar 10% de n, sem alterar a distribuição de probabilidade normal, bem como as estimativas de seus parâmetros. Entretanto, conforme o número de variáveis aumenta ($p \geq 10$), vão surgindo limitações a esse método e deve-se ter cautela ao utilizá-lo.

Comparando o aumento de dados via arquétipos com métodos de reamostragem existentes, como Monte Carlo, *jackknife* e *bootstrap*, por exemplo, em ambos os casos, a principal vantagem é que só existem ganhos ao utilizá-los, pois independente das pressuposições de um procedimento de estimação ou decisão serem atendidas, seus resultados serão os mais confiáveis (BASTOS, 2013).

Vale notar que os arquétipos conseguem representar bem a amostra de determinada população, de forma análoga a estatísticas suficientes, e sem nenhum conhecimento prévio sobre a distribuição da variável aleatória e nem sobre o verdadeiro valor do parâmetro. Levando em conta que a testemunha avaliada nesse estudo consistiu de um método inspirado no *boostrap* não paramétrico, verifica-se que o aumento de dados via arquétipos pode ser mais vantajoso que esse técnica de reamostragem, o que pode ser testado de maneira mais ampla em trabalhos futuros.

4.1 AUMENTOS SUCESSIVOS

Os resultados das simulações dos quatro cenários estão apresentados, sendo uma tabela para cada algoritmo e três gráficos por cenário, referentes aos testes sobre a distribuição normal multivariada, sobre o vetor de médias e sobre a matriz de covariâncias realizados. Em cada tabela encontram-se a mediana dos aumentos máximos obtidos nas 100 simulações, com cada método de aumento, e seu intervalo de confiança $IC_{95\%}$(Md), agrupando-se os métodos da seguinte forma: testemunha (T) e controle positivo (CP), $A1_i$ e $A1_a$, $A2_i$ e $A2_a$, e $A3_i$ e $A3_a$.

Na Tabela 6 encontram-se os resultados dos aumentos máximos possíveis com a testemunha e com o controle positivo, sendo que para esse último foi apresentada somente sua mediana, visto que o limite inferior e o superior do intervalo de confiança foram iguais ao valor mediano em todos os casos.

Tabela 6 – Mediana dos aumentos máximos (Md) com a testemunha (T) e com o controle positivo (CP) sem alterar a distribuição de probabilidade da variável aleatória (Normalidade) e as estimativas do vetor de médias (Médias) e da matriz de covariâncias (Covariâncias) da amostra inicial para cada cenário.

Teste	Cenário 1			Cenário 2		
	T		CP	T		CP
	Md	$IC_{95\%}$(Md)	Md	Md	$IC_{95\%}$(Md)	Md
Normalidade	5	[4; 5]	400	1	[1; 2)	120
Médias	25	[15; 31]	400	4	[3; 5]	120
Covariâncias	16	(13; 21)	400	5	[4; 7]	120
Teste	Cenário 3			Cenário 4		
	T		CP	T		CP
	Md	$IC_{95\%}$(Md)	Md	Md	$IC_{95\%}$(Md)	Md
Normalidade	0	[0; 1]	60	0	[0; 0]	40
Médias	3	[2; 4]	60	2	[1; 3]	40
Covariâncias	2	[1; 2]	60	2	[1; 2]	40

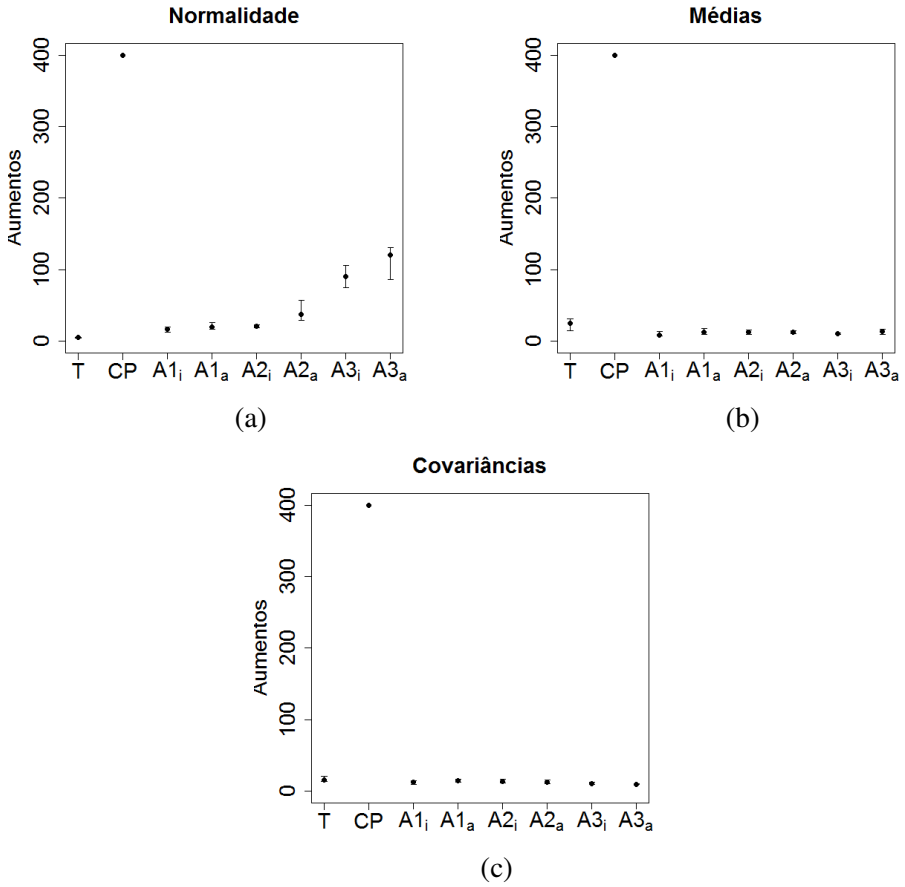
Fonte: Da autora.

Pode-se notar que o controle positivo apresentou mediana dos aumentos máximos igual ao número total de aumentos consecutivos realizados em todos os cenários, confirmando o seu papel como padrão ouro do estudo de simulação. Já a testemunha, obteve valores medianos inferiores, principalmente quando testada a distribuição de probabilidade através da amostra aumentada, não permitindo nenhum aumento nos cenários 3 e 4 (TABELA 6). Esse fato ocorre devido à testemunha permitir sortear apenas elementos presentes na amostra, o que ocasiona picos de frequência de determinadas observações do espaço amostral.

Na Figura 8 encontram-se os gráficos do cenário 1 (aumentos de 10% de n) contendo os

valores medianos dos aumentos máximos com cada algoritmo comparado no estudo de simulação.

Figura 8 – Mediana dos aumentos máximos no cenário 1 com a testemunha (T), controle positivo (CP) e os algoritmos 1, 2 e 3, sendo $A1_i$, $A2_i$ e $A3_a$ utilizando arquétipos da amostra inicial e $A1_a$, $A2_a$ e $A3_a$ das amostras aumentadas, sem modificar: a distribuição de probabilidade (a) e as estimativas do vetor de médias (b) e da matriz de covariâncias (c) da amostra inicial.

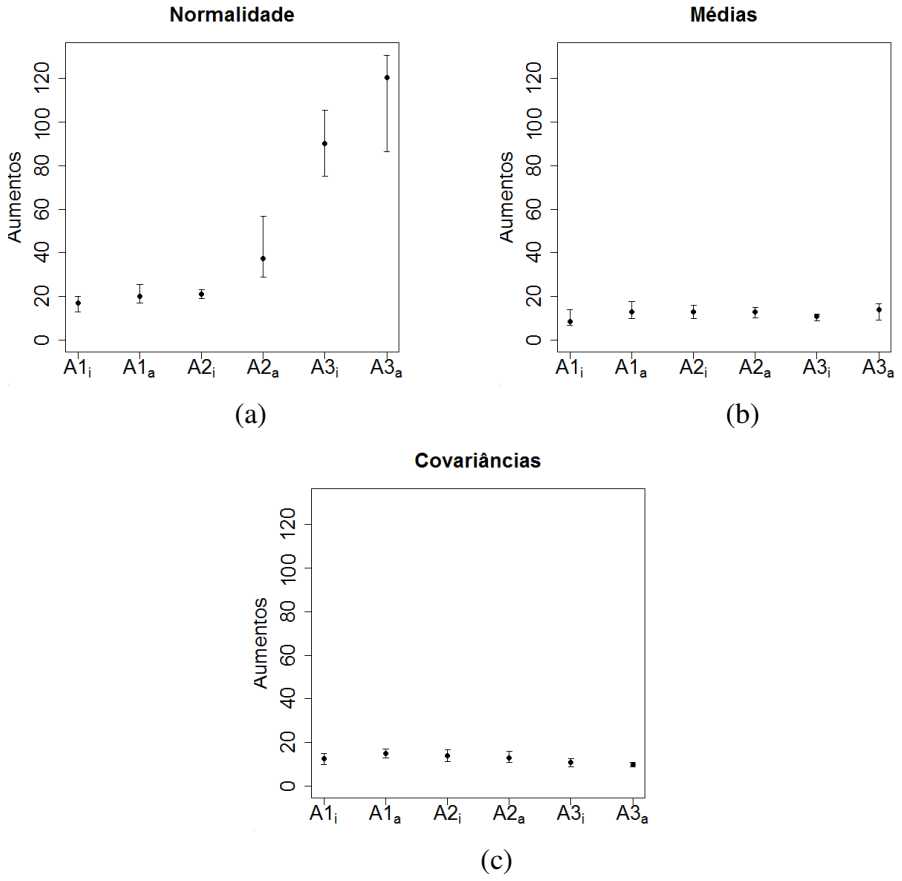


Fonte: Da autora.

Primeiramente, verifica-se nestes gráficos que o controle positivo destoa dos demais, inclusive dificultando a comparação dos outros algoritmos; em seguida, comparando os algoritmos 1, 2 e 3 com a testemunha, observam-se resultados bem próximos, muitas vezes iguais, quanto aos aumentos mantendo as estimativas dos parâmetros, e um número inferior de aumentos com a testemunha quando avaliada a distribuição normal de probabilidade, o que ocorreu em todos os cenários (FIGURA 8). Sendo assim, a testemunha foi considerada inferior aos demais, pois deturpa a distribuição de probabilidade da amostra, o que costuma ser mais grave do que alterar a estimativa de seus parâmetros.

Tendo em vista essas considerações, os gráficos do cenário 1 foram refeitos removendo a testemunha e o controle positivo, padrão que foi mantido nos outros 3 cenários. Portanto, na Figura 9 é possível visualizar melhor a comparação dos algoritmos 1, 2 e 3.

Figura 9 – Mediana dos aumentos máximos no cenário 1 com os algoritmos 1, 2 e 3, sendo $A1_i$, $A2_i$ e $A3_a$ utilizando arquétipos da amostra inicial e $A1_a$, $A2_a$ e $A3_a$ das amostras aumentadas, sem modificar: a distribuição de probabilidade (a) e as estimativas do vetor de médias (b) e da matriz de covariâncias (c) da amostra inicial.



Fonte: Da autora.

Comparando cada um dos três algoritmos utilizando os arquétipos da amostra inicial e utilizando os arquétipos das amostras aumentadas, na maioria dos vezes, pode-se equiparar os resultados. Comparando agora os três algoritmos entre si, tem-se aumentos medianos semelhantes quando avaliadas as estimativas dos parâmetros, sendo possível realizar aproximadamente 13 aumentos garantindo a estimativa do vetor de médias e 12 aumentos garantindo a estimativa da matriz de covariâncias. Como são aumentos de três dados cada, é possível aumentar 130% (39 dados) e 120% (36 dados) do tamanho da amostra inicial, respectivamente. Já na garantia da distribuição de probabilidade, o Algoritmo 3 se destaca dos demais, permitindo 120 aumentos com $A3_a$, ou seja, um total de 1200% de n (FIGURA 9).

Complementando a informação dos gráficos acima, a Tabela 7 apresenta os resultados dos aumentos máximos com o Algoritmo 1 nos quatro cenários.

Tabela 7 – Mediana dos aumentos máximos (Md) com o Algoritmo 1 sem alterar a distribuição de probabilidade da variável aleatória (Normalidade) e as estimativas do vetor de médias (Médias) e da matriz de covariâncias (Covariâncias) da amostra inicial para cada cenário.

Teste	Cenário 1				Cenário 2			
	$A1_i$		$A1_a$		$A1_i$		$A1_a$	
	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)
Normalidade	17	[13; 20]	20	[17; 26)	4	[3; 5]	6	(4; 7]
Médias	8	[7; 14]	13	[10; 18)	2	[2; 3]	3	[2; 4]
Covariâncias	12	[10; 15]	15	[13; 17]	4	[3; 4]	4	[3; 6]
Teste	Cenário 3				Cenário 4			
	$A1_i$		$A1_a$		$A1_i$		$A1_a$	
	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)
Normalidade	2	[2; 2]	3	[2; 4]	1	[1; 1]	1	[1; 2]
Médias	1	[1; 2]	2	[1; 2]	1	[0; 1]	0	[0; 1)
Covariâncias	2	[1; 2]	1	[1; 2]	1	[1; 1]	1	[1; 1]

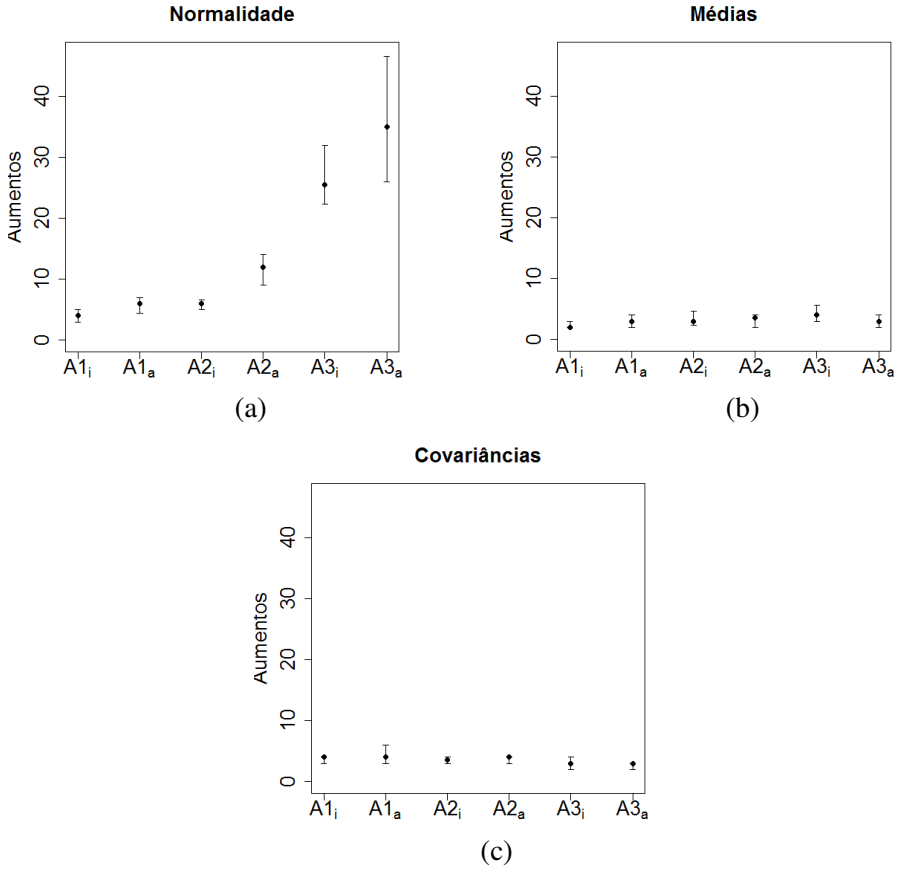
Fonte: Da autora.

Nota: $A1_i$: arquétipos da amostra inicial; $A1_a$: arquétipos das amostras aumentadas.

Na Tabela 7 fica mais nítida a equivalência entre $A1_i$ e $A1_a$, pois existe interseção na grande maioria dos intervalo de de confiança. Além disso, é interessante notar que o número de dados aumentados em cada cenário é bem próximo, contudo é mais vantajoso fazer um número maior de aumentos de poucos dados. Tomando $A1_i$, por exemplo, é possível realizar 17 aumentos de três dados, quatro aumentos de 10 dados, dois aumentos de 20 dados e um aumento de 30 dados, mantendo a distribuição de probabilidade normal. Desta forma, é possível aumentar 51, 40, 40 e 30 dados, ou 170%, 133%, 133% e 100% de n respectivamente (TABELA 7).

Em seguida, tem-se os gráficos do cenário 2 (aumentos de 33% de n) que se encontram na Figura 10.

Figura 10 – Mediana dos aumentos máximos no cenário 2 com os algoritmos 1, 2 e 3, sendo $A1_i$, $A2_i$ e $A3_a$ utilizando arquétipos da amostra inicial e $A1_a$, $A2_a$ e $A3_a$ das amostras aumentadas, sem modificar: a distribuição de probabilidade (a) e as estimativas do vetor de médias (b) e da matriz de covariâncias (c) da amostra inicial.



Fonte: Da autora.

Os resultados obtidos no segundo cenário são semelhantes aos do primeiro, ou seja, os três algoritmos são equiparáveis em relação ao número de aumentos que garantem as estimativas dos parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ e o Algoritmo 3 apresentou-se superior quanto ao número de aumentos garantindo normalidade. Neste cenário, os algoritmos 1, 2 e 3 obtiveram valor mediano dos aumentos máximos em torno de 3, quando avaliados o vetor de médias e a matriz de covariâncias, o que permite aumentar 100% do tamanho da amostra inicial - lembrando que cada aumento possui 10 dados. Para garantir a distribuição de probabilidade, com $A3_a$ é possível realizar 35 aumentos, ou seja, aproximadamente 1167% de n (FIGURA 10).

Analisando em mais detalhes o Algoritmo 2, na Tabela 8 tem-se os resultados dos aumentos máximos com este algoritmo.

Tabela 8 – Mediana dos aumentos máximos (Md) com o Algoritmo 2 sem alterar a distribuição de probabilidade da variável aleatória (Normalidade) e as estimativas do vetor de médias (Médias) e da matriz de covariâncias (Covariâncias) da amostra inicial para cada cenário.

Teste	Cenário 1				Cenário 2			
	$A2_i$		$A2_a$		$A2_i$		$A2_a$	
	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)
Normalidade	21	[19; 23]	38	[29; 57)	6	[5; 7)	12	[9; 14]
Médias	13	[10; 16]	13	(10; 15]	3	(2; 5)	4	[2; 4]
Covariâncias	14	(11; 17)	13	[11; 16]	3	[3; 4]	4	[3; 4]
Teste	Cenário 3				Cenário 4			
	$A2_i$		$A2_a$		$A2_i$		$A2_a$	
	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)
Normalidade	2	[2; 3]	4	[3; 6]	2	[1; 2]	2	[2; 3)
Médias	2	[1; 2]	2	[1; 3)	1	[0; 1]	1	[0; 1]
Covariâncias	1	[1; 2]	2	[1; 2]	1	[1; 1]	1	[1; 1]

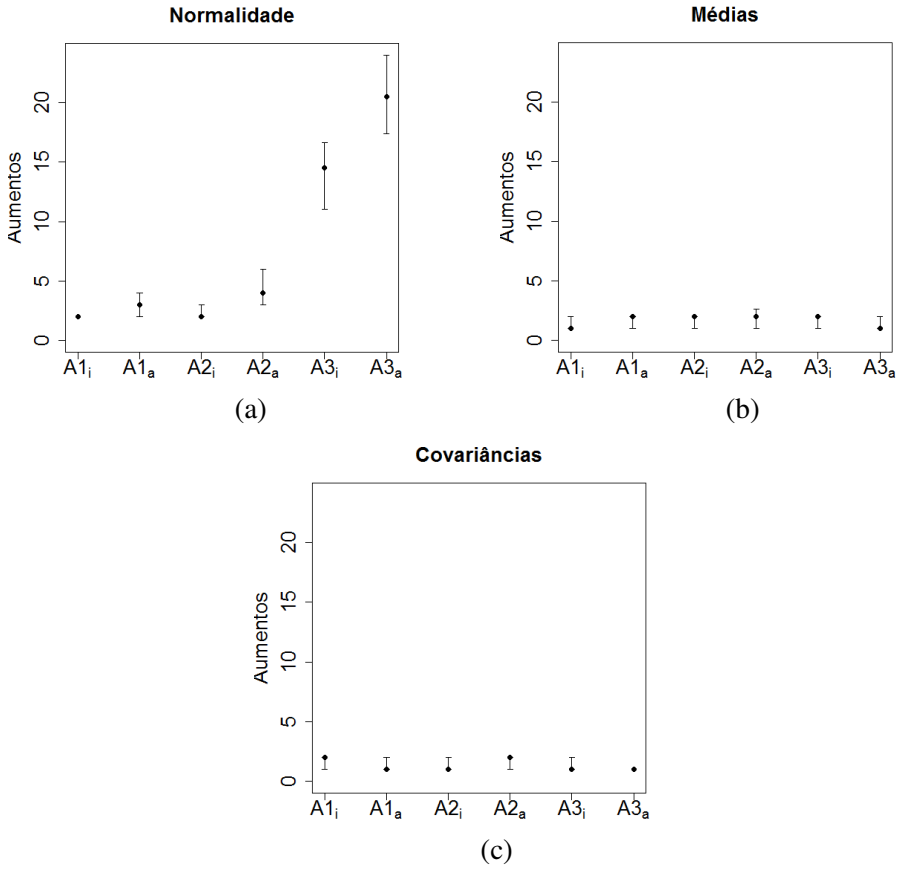
Fonte: Da autora.

Nota: $A2_i$: arquétipos da amostra inicial; $A2_a$: arquétipos das amostras aumentadas.

Comparando $A2_i$ e $A2_a$, verifica-se que existe interseção entre os intervalos de confiança referentes aos aumentos possíveis sem deturpar as estimativas do vetor de médias e da matriz de covariâncias, entretanto, não existe interseção entre os intervalos de confiança referentes aos aumentos possíveis sem deturpar a distribuição normal. Nesta última situação, $A2_a$ se destaca permitindo 12 aumentos de 10 dados (cenário 2), isto é, 400% de n. Nesse algoritmo, o número de dados aumentados em cada cenário também foi bem parecido, mas ainda existe vantagem em se fazer aumentos com menos dados. Exemplificando com os aumentos possíveis garantindo normalidade em $A2_a$, tem-se aumentos de 380% (114 dados) e 400% (120 dados) da amostra inicial com os cenários 1 e 2, respectivamente, contra 267% (80 dados) e 200% (60 dados) nos cenários 3 e 4 (TABELA 8).

Na sequência, encontram-se os gráficos do terceiro cenário (aumentos de 67% de n) na Figura 11.

Figura 11 – Mediana dos aumentos máximos no cenário 3 com os algoritmos 1, 2 e 3, sendo $A1_i$, $A2_i$ e $A3_a$ utilizando arquétipos da amostra inicial e $A1_a$, $A2_a$ e $A3_a$ das amostras aumentadas, sem modificar: a distribuição de probabilidade (a) e as estimativas do vetor de médias (b) e da matriz de covariâncias (c) da amostra inicial.



Fonte: Da autora.

Mais uma vez, observa-se o mesmo padrão no terceiro cenário simulado, sendo possível realizar entre 1 e 2 aumentos, mantendo as estimativas dos parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ com os três algoritmos. Em outras palavras, como cada aumento possui 20 dados, é possível aumentar entre 67% e 133% do tamanho da amostra inicial sem deturpar essas estimativas. Como o Algoritmo 3 destaca-se novamente, com $A3_a$ é possível fazer 20 aumentos, mantendo a distribuição de probabilidade da amostra inicial, o que significa aumentar n em 1333% (FIGURA 11).

Na Tabela 9, a seguir, pode-se explorar os resultados obtidos com o Algoritmo 3.

Tabela 9 – Mediana dos aumentos máximos (Md) com o Algoritmo 3 sem alterar a distribuição de probabilidade da variável aleatória (Normalidade) e as estimativas do vetor de médias (Médias) e da matriz de covariâncias (Covariâncias) da amostra inicial para cada cenário.

Teste	Cenário 1				Cenário 2			
	$A3_i$		$A3_a$		$A3_i$		$A3_a$	
	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)
Normalidade	90	[75; 106)	120	(86; 131)	26	(22; 32)	35	[26; 47)
Médias	11	[9; 12]	14	(9; 17)	4	[3; 6)	3	[2; 4]
Covariâncias	11	[9; 13)	10	[9; 11]	3	[2; 4]	3	[2; 3]
Teste	Cenário 3				Cenário 4			
	$A3_i$		$A3_a$		$A3_i$		$A3_a$	
	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)	Md	$IC_{95\%}$(Md)
Normalidade	14	[11; 17)	20	(17; 24]	7	[6; 8]	14	(11; 15]
Médias	2	[1; 2]	1	[1; 2]	1	[0; 1]	1	[0; 1]
Covariâncias	1	[1; 2]	1	[1; 1]	1	[0; 1]	1	[0; 1]

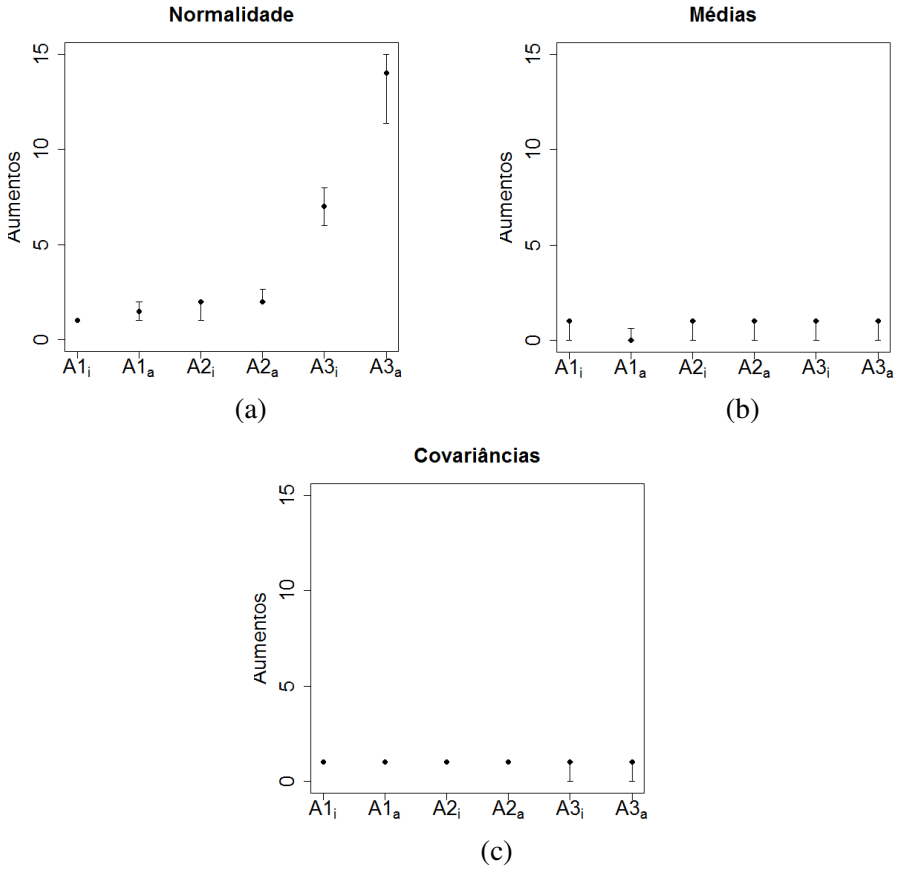
Fonte: Da autora.

Nota: $A3_i$: arquétipos da amostra inicial; $A3_a$: arquétipos das amostras aumentadas.

O Algoritmo 3 merece destaque, pois seus resultados indicam que este proporciona um maior aumento de dados sem deturpar a distribuição de probabilidade da amostra inicial. Dentre $A3_i$ e $A3_a$, nota-se que, nos dois primeiros cenários, ambos apresentaram resultados equiparáveis, devido à interseção existente entre os intervalos de confiança, enquanto nos outros dois houve diferença entre os resultados, principalmente com relação à distribuição de probabilidade. Novamente, o número de dados aumentados em cada cenário foi semelhante, com exceção dos resultados sobre a distribuição de probabilidade, o que pode ser exemplificado por $A3_i$: no cenário 1 foram possíveis 90 aumentos de três dados, no cenário 2, 26 aumentos de 10 dados, no cenário 3, 14 aumentos de 20 dados e, no cenário 4, 7 aumentos de 30 dados. Sendo assim, é possível aumentar o tamanho da amostra inicial em 900%, 867%, 933% e 700%, respectivamente.

Por fim, tem-se os gráficos do quarto e último cenário (aumentos de 100% de n) na Figura 12.

Figura 12 – Mediana dos aumentos máximos no cenário 4 com os algoritmos 1, 2 e 3, sendo $A1_i$, $A2_i$ e $A3_a$ utilizando arquétipos da amostra inicial e $A1_a$, $A2_a$ e $A3_a$ das amostras aumentadas, sem modificar: a distribuição de probabilidade (a) e as estimativas do vetor de médias (b) e da matriz de covariâncias (c) da amostra inicial.



Fonte: Da autora.

Nos gráficos da Figura 12 também foi verificado que os três algoritmos apresentaram desempenho semelhante quanto à garantia das estimativas do vetor de médias e da matriz de covariâncias, permitindo realizar 1 aumento de 30 dados, ou seja, 100% de n. Como o Algoritmo 3 destaca-se novamente, com $A3_a$ foi possível fazer 14 aumentos mantendo a distribuição de probabilidade da amostra inicial, o que significa aumentar n em 1400%.

Portanto, os algoritmos 1, 2 e 3 apresentaram resultados semelhantes à testemunha, quanto aos aumentos possíveis sem modificar as estimativas dos parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, e se mostraram superiores por manter a distribuição de probabilidade normal após um número maior de aumentos consecutivos. Nesse aspecto, não há dúvidas quanto à eficiência do Algoritmo 3.

Comparando os resultados de aumentos realizados utilizando os arquétipos da amostra inicial com os de aumentos realizados com os arquétipos das amostras aumentadas, verifica-se que ambos foram semelhantes e, muitas vezes equivalentes. Sendo assim, pode-se reduzir o custo computacional utilizando os arquétipos da amostra inicial em todos os aumentos.

Com essas considerações, e levando em conta o princípio da parcimônia, para a reco-

mendação do maior número de aumentos sucessivos sem deturpar simultaneamente a distribuição normal de probabilidade e as estimativas do vetor de médias e da matriz de covariâncias, foi utilizado o Algoritmo 3, apenas com os arquétipos da amostra inicial: A3$_i$. Com esse algoritmo, foi possível realizar 11 aumentos de três dados (110% de n), três aumentos de 10 dados (100% de n), um aumento de 20 dados (67% de n) e um aumento de 30 dados (100% de n). Sendo assim, recomenda-se realizar 11 aumentos de 10% de n.

4.2 ESTUDO COM DADOS REAIS

Antes da aplicação do aumento de dados, verificou-se que a variável aleatória referente ao conjunto de dados reais sobre bebida láctea achocolatada ($n = 17$ provadores) apresentou distribuição de probabilidade normal, de acordo com o teste de Royston a 5% de significância (valor-$p = 0{,}4436$).

Após a realização do aumento de dados de duas maneiras, conforme as recomendações do trabalho, ou seja, um aumento de 10% de n com o A3 ($a = 2$), totalizando 19 provadores, e 11 aumentos sucessivos de 10% de n com A3$_i$ ($a = 22$), totalizando 39 provadores, pôde-se observar as estimativas pontuais do vetor de médias dos dados completos e dos dados aumentados que se encontram na Tabela 10.

Tabela 10 – Estimativa pontual da média ($\bar{x}$) e erro padrão da média (EP$_{\bar{x}}$) das notas de cada atributo sensorial da amostra inicial (AI) e das amostras aumentadas por A3 e A3$_i$.

	Aroma Chocolate		Sabor Chocolate	
Amostra	$\bar{x}$	EP$_{\bar{x}}$	$\bar{x}$	EP$_{\bar{x}}$
AI	4,18	1,01	4,94	1,20
A3	4,05	0,86	4,60	0,98
A3$_i$	4,41	0,71	6,24	1,00

Fonte: Da autora.

Pode-se notar que as estimativas das médias foram semelhantes em todos os casos e que os erros padrão diminuíram (TABELA 10). A seguir, tem-se as matrizes de covariância da amostra inicial e das amostras aumentadas na Figura 13 abaixo.

Figura 13 – Matrizes de covariâncias da amostra inicial e das amostras aumentadas por A3 e A3$_i$.

$$\begin{pmatrix} 3{,}24 & 0{,}22 \\ 0{,}22 & 3{,}64 \end{pmatrix} \qquad \begin{pmatrix} 3{,}09 & 0{,}60 \\ 0{,}60 & 4{,}26 \end{pmatrix} \qquad \begin{pmatrix} 6{,}27 & 3{,}23 \\ 3{,}23 & 7{,}90 \end{pmatrix}$$

(a) AI. (b) A3. (c) A3$_i$.

Fonte: Da autora.

Na Figura 13 tem-se que a matriz de covariâncias dos dados aumentados por A3 foi muito parecida com a dos dados iniciais e que a matriz de covariâncias dos dados aumentados por A3$_i$ apresentou valores superiores. Logo, foi necessária a realização do teste de Box para correta interpretação desses resultados.

O valor-p dos testes multivariados realizados se encontram na Tabela 11 e indicam que a distribuição de probabilidade normal e as estimativas de seus parâmetros foram mantidas (valor-$p \geq 0{,}05$).

Tabela 11 – Valor-p dos testes de Royston, T^2 de Hotelling e de Box após o aumento realizado por A3 e A3$_i$.

Teste	A3	A3$_i$
Royston	0,4338	0,3533
T^2 de Hotelling	0,8716	0,2227
Box	0,9808	0,1716

Fonte: Da autora.

O resultado do teste de Royston, permitiu concluir que a distribuição de probabilidade foi mantida após o aumento de dados realizado com ambos A3 e A3$_i$. O teste T^2 de Hotelling indicou que os vetores de médias das amostras aumentadas foram iguais ao vetor de médias da amostra inicial. E o teste de Box concluiu que as matrizes de covariâncias das amostras aumentadas foram iguais à da amostra inicial, pois, em todos os casos, o valor-p foi maior que o nível de significância adotado (5%), aceitando as hipóteses nulas (TABELA 11).

Portanto, existem indícios de que há um ganho na precisão da inferência praticada ao aumentar o tamanho da amostra através dos arquétipos, pois a distribuição de probabilidade e as estimativas de seus parâmetros foram mantidas e, além disso, houve uma redução no erro padrão da média (TABELA 10). Esse fato mostra que é possível realizar o aumento de dados conforme as recomendações dos estudos computacionais.

No caso particular da análise sensorial da bebida achocolatada, foi possível aumentar o número de provadores, mostrando ser muito útil quando não se possui um número suficiente de provadores devidamente treinados.

5 CONCLUSÕES

Foram propostos três algoritmos para aumento de dados via arquétipos, que apresentaram melhor desempenho que a testemunha, quando avaliado se a variável aleatória p-variada apresentava distribuição de probabilidade normal após o aumento realizado, inclusive com resultados competitivos aos do controle positivo nos cenários bivariados.

Dentre os algoritmos propostos, o algoritmo 3 foi eleito o mais eficaz nos dois estudos computacionais, por ter apresentado melhor desempenho que os demais em todos os cenários, permitindo aumentar o tamanho da amostra inicial, quando $p \leq 5$, sem alterar a distribuição de probabilidade original, bem como as estimativas de seus parâmetros.

Pelo princípio da parcimônia, concluiu-se que os aumentos sucessivos de dados podem ser realizados utilizando apenas os arquétipos da amostra inicial em cada aumento consecutivo.

Em uma situação real foi possível aumentar o tamanho da amostra e proporcionar maior precisão na inferência sobre o vetor de médias e a matriz de covariâncias.

Portanto, parece ser seguro aumentar dados por meio de seus arquétipos sugerindo-se o Algoritmo 3 e dentro das seguintes recomendações:

- Em cenários bivariados com $n \leq 100$, um aumento de até 10%;
- Em cenários bivariados com $n = 30$, 11 aumentos sucessivos de 10% de n;
- Em cenários com cinco variáveis e $n \leq 100$, um aumento de até 10%.
- Não é recomendado o aumento de dados via arquétipos para conjuntos de dados com $p \geq 10$.

REFERÊNCIAS

BASTOS, R. L. **Proposição de Testes Bootstrap para o Índice de Qualidade Sensorial**. 2013. 125 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras - UFLA, Lavras, MG.

BAUCKHAGE, C.; THURAU, C. Making archetypal analysis practical. In: **Pattern Recognition**. Springer, 2009. p. 272–281. Disponível em: <http://dx.doi.org/10.1007/978-3-642-03798-6_28>. Acesso em: 06 mar. 2016.

BEHRENS, W. V. Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. **Landwirtschaftliche Jahrbücher**, Berlin, v. 68, p. 807–837, 1929.

BOX, G. E. P. A general distribution theory for a class of likelihood criteria. **Biometrika**, Londres, v. 36, n. 3/4, p. 317–346, 1949.

CHAN, B. H. P.; MITCHELL, D. A.; CRAM, L. E. Archetypal analysis of galaxy spectra. **Monthly Notices of the Royal Astronomical Society**, Oxford, v. 338, n. 3, p. 790–795, 2003.

COSTANTINI, P. et al. Archetypal functions. In: **Analysis and Modeling of Complex Data in Behavioural and Social Sciences**. Anacapri, Italy: Springer, 2012. p. 4.

CUTLER, A.; BREIMAN, L. Archetypal analysis. **Technometrics**, Washington, v. 36, n. 4, p. 338–347, 1994.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm (with discussion). **Journal of the Royal Statistical Society**, Londres, v. 39, n. 1, p. 1–38, 1977.

D'ESPOSITO, M. R.; PALUMBO, F.; RAGOZINI, G. Archetypal analysis for interval data in marketing research. **Statistica Applicata**, v. 18, n. 2, p. 343–358, 2006.

____. On the use of archetypes and interval coding in sensory analysis. In: FICHET, B. et al. (Ed.). **Classification and Multivariate Analysis for Complex Data Structures**. Italy: Springer Berlin Heidelberg, 2011, (Studies in Classification, Data Analysis, and Knowledg Organization). p. 353–361.

____. Interval archetypes: a new tool for interval data analysis. **Statistical Analysis and Data Mining**, Wiley Online Library, v. 5, n. 4, p. 322–335, 2012.

EDDY, W. F. A new convex hull algorithm for planar sets. **ACM Transactions on Mathematical Software**, Nova York, v. 3, n. 4, p. 398–403, 1977.

EPIFANIO, I.; VINUÉ, G.; ALEMANY, S. Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. **Computers & Industrial**

Engineering, Elsevier, Nova York, v. 64, n. 3, p. 757–765, 2013.

EUGSTER, M. J. A. **Archetypal athletes**. 2011. Disponível em: <http://arxiv.org/pdf/1110.1972.pdf>. Acesso em: 09 set. 2016.

EUGSTER, M. J. A.; LEISCH, F. From Spider-Man to Hero - Archetypal Analysis in R. **Journal of Statistical Software**, v. 30, n. 8, p. 1–23, 2009.

FERREIRA, D. F. **Estatística Multivariada**. 2. ed. Lavras: Ed. UFLA, 2011. 676 p.

FISHER, R. A. The comparison of samples with possibly unequal variances. **Annals of Eugenics**, Londres, v. 9, p. 174–180, 1939.

GENZ, A.; BRETZ, F. Computation of multivariate normal and t probabilities. **Lecture Notes in Statistics**, Berlin, v. 195, 2009.

HAIR JR., J. F. et al. **Análise Multivariada de Dados**. 5. ed. Porto Alegre: Bookman, 2005. 593 p.

HOTELLING, H. A generalized t-test and measure of multivariate dispersion. In: **Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.: s.n.], 1951. p. 23–41.

KORKMAZ, S.; GOKSULUK, D.; ZARARSIZ, G. MVN: An R package for assessing multivariate normality. **The R Journal**, v. 6, n. 2, p. 151–162, 2014. Disponível em: <http://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf>.

LEDOIT, O.; WOLF, M. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. **The Annals of Statistics**, v. 30, n. 4, p. 1081–1102, 2002.

LI, S. et al. Archetypal analysis: a new way to segment markets based on extreme individuals. In: **A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution. Proceedings of the ANZMAC 2003 Conference**. [S.l.: s.n.], 2003. p. 1674–1679.

MARTINS JÚNIOR, J. M. et al. Análise de arquétipos na avaliação da movimentação de jogadores de futebol. **Revista Brasileira de Biometria**, São Paulo, v. 33, n. 1, p. 30–41, 2015a.

____. A análise de arquétipos: uma revisão bibliográfica. **Revista Brasileira de Biometria**, São Paulo, v. 33, n. 2, p. 156–169, 2015b.

MORUP, M.; HANSEN, L. K. Archetypal analysis for machine learning and data mining. **Neurocomputing**, v. 80, p. 54–63, 2012. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231211006060>. Acesso em: 06 mar. 2016.

NORDHAUSEN, K. et al. **ICSNP: Tools for Multivariate Nonparametrics**. [S.l.], 2015. R package version 1.1-0. Disponível em: <http://CRAN.R-project.org/package=ICSNP>.

PIGOTT, T. D. A review of methods for missing data. **Educational Research and Evaluation**, v. 7, n. 4, p. 353–383, 2001.

PORZIO, G. C.; RAGOZINI, G.; VISTOCCO, D. On the use of archetypes as benchmarks. Wiley Online Library, v. 24, n. 5, p. 419–437, 2008.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2016. Disponível em: <http://www.R-project.org/>. Acesso em: 06 mar. 2016.

RIEDESEL, P. Archetypal analysis in marketing research: a new way of understanding consumer heterogeneity. **Action Marketing Research**, v. 24, 2014.

ROYSTON, J. P. Some techniques for assesing multivariate normality based on the shapiro-wilk w. **Journal of the Royal Statistical Society**, Londres, v. 32, n. 2, p. 121–133, 1983.

SCHNAIDER, T. B.; SOUZA, C. de. Aspectos Éticos da experimentação animal. **Revista Brasileira de Anestesiologia**, Rio de Janeiro, v. 53, n. 2, p. 278 – 285, 2003.

SEILER, C.; WOHLRABE, K. Archetypal scientists. **Journal of Informetrics**, Elsevier, v. 7, n. 2, p. 345–356, 2013.

SIFA, R.; BAUCKHAGE, C. Archetypical motion: Supervised game behavior learning with archetypal analysis. In: **IEEE Conference on Computational Intelligence in Games (CIG)**. Dortmund: IEEE, 2013. p. 1–8.

SIFA, R.; BAUCKHAGE, C.; DRACHEN, A. The playtime principle: Large-scale cross-games interest modeling. In: **IEEE Conference on Computational Intelligence and Games (CIG)**. Dortmund: IEEE, 2014. p. 1–8.

STONE, E.; CUTLER, A. Archetypal analysis of spatio-temporal dynamics. **Physica D**, Amsterdã, v. 90, p. 209–224, 1996.

STONE, E.; OLSON, B. **Archetypal Analysis of Cellular Flame Data**. Utah State University, 1999. Technical report.

TANNER, M. A.; WONG, W. H. The calculation of posterior distributions by data augmentation (with discussion). **Journal of the American Statistical Association**, Nova York, v. 82, n. 398, p. 528–550, 1987.

THOGERSEN, J. C. et al. Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. **BMC Bioinformatics**, v. 279, n. 14, p. 1–15, 2013.

VAN DYK, D. A.; MENG, X.-L. The art of data augmentation. **Journal of Computational and Graphical Statistics**, Alexandria, v. 10, n. 1, p. 1–50, 2001.

VIEIRA, S.; HOSSNE, W. S. **Metodologia Científica para a área da Saúde**. 2. ed. Rio de Janeiro: Elsevier, 2015.

VINUÉ, G.; EPIFANIO, I.; ALEMANY, S. Archetypoids: a new approach to define representative archetypal data. **Computational Statistics & Data Analysis**, Amsterdã, v. 87, p. 102–115, 2014.