UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

ADENICE VASCONCELOS FERREIRA

Distâncias Estocásticas para Agrupamentos de Formas em 2D e Dados Direcionais

Recife

2024

ADENICE VASCONCELOS FERREIRA

Distâncias Estocásticas para Agrupamentos de Formas em 2D e Dados Direcionais

Tese apresentada, como requisito parcial para obtenção do título de Doutora em Estatística, ao Programa de Pós-Graduação em Estatística, da Universidade Federal de Pernambuco.

Área de Concentração: Estatística Aplicada

Orientador: Prof. Dr. Getúlio José Amorim do Amaral

Coorientador: Prof. Dr. Abraão David Costa do Nascimento

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Ferreira, Adenice Vasconcelos.
Distâncias Estocásticas para Agrupamentos de Formas em 2D e Dados Direcionais / Adenice Vasconcelos Ferreira. - Recife, 2024.
87p.: il.

Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-graduação em Estatística, Recife, 2024.
Orientação: Getúlio José Amorim do Amaral.
Coorientação: Abraão David Costa do Nascimento.

1. Agrupamento; 2. Análise de formas; 3. Distribuição Bingham; 4. Distâncias; 5. Dados direcionais; 6. Espaço multidimensional. I. Amaral, Getúlio José Amorim do - Orientador. II. Abraão David Costa do Nascimento - Coorientador. III. Título.

UFPE-Biblioteca Central CDD 310

ADENICE GOMES DE OLIVEIRA FERREIRA

Distâncias Estocásticas para Agrupamento de Formas em 2D e Dados Direcionais

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Estatística.

Aprovada em 20 de junho de 2024.

BANCA EXAMINADORA

Prof. Dr. GETULIO JOSE AMORIM DO AMARAL, UFPE
Presidente

Profª Drª LUCIA PEREIRA BARROSO, IME/USP
Examinadora Externa à Instituição

Prof. Dr. ROBERTO FERREIRA MANGHI, UFPE
Examinador Interno

Prof. Dr. ALEX DIAS RAMOS, UFPE
Examinador Interno

Profª Drª RENATA MARIA CARDOSO RODRIGUES DE SOUZA, UFPE
Examinadora Externa

AGRADECIMENTOS

Primeiramente, agradeço a Deus, por me guiar em cada passo, me dar força e condições de cumprir mais esta missão que me foi concebida. Agradeço pelo consolo em momentos difíceis, pelas palavras de incentivo e pela fé.

Agradeço ao meu esposo por acreditar em mim e dar suporte durante minha caminhada acadêmica. Nossa caminhada, na verdade! Começamos juntos na graduação e permanecemos dando suporte um ao outro até aqui. Partilhar a vida com você, Diogo, tornam os desafios mais leves e as vitórias mais significativas.

Sou grata aos meus pais (Gilvandro e Adelma) e ao meu irmão (João Marcos) por todo apoio e cumplicidade. Enfrentamos muitas lutas juntos, inclusive os desafios advindos juntamente com a pandemia. Temos vencido até aqui, pela graça e misericórdia do Senhor.

Agradeço aos amigos, familiares e irmãos em Cristo, que de forma direta ou indireta contribuíram com meu aprendizado nesta etapa da vida. Principalmente, aos meus colegas de sala no doutorado - sou feliz em termos construído uma amizade sólida.

Agradeço aos professores Getúlio e Abraão pela confiança, incentivo e orientação nesta etapa de conclusão do doutorado. Findo esta etapa com o privilégio de poder partilhar da construção deste trabalho com pessoas completamente humanizadas. Estimo bençãos sobre suas vidas!

Gratidão aos professores que fizeram parte da minha banca examinadora. As contribuições sugeridas foram fundamentais para o aprimoramento desta obra.

Por fim, meus agradecimentos à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

"The true foundation of theology is to ascertain the character of God. It is by the aid of Statistics that law in the social sphere can be ascertained and codified, and certain aspects of the character of God thereby revealed. The study of Statistics is thus a religious service." (DAVID, 1998).

RESUMO

Métodos não supervisionados são utilizados na Estatística para distribuir dados em grupos distintos entre si, e com elementos com maior similaridade dentro dos grupos. Novas técnicas têm sido propostas a fim de melhorar a eficácia do agrupamento em diversos cenários. Neste trabalho, propusemos três distâncias para dados direcionais baseadas na distribuição Von Mises-Fisher matricial e dois algoritmos: um para cenários de baixa concentração de dados de pré-formas, considerando vetores complexos, e outro para dados do tipo direcional, dentro do espaço multidimensional de Stiefel múltiplo. O algoritmo proposto para agrupar pré-formas possui nove possíveis variações, por conter duas possibilidades de critério de parada e três possibilidades de distâncias baseadas na ditribuição Bingham (Battacharyya, Rényi e Hellinger), além da distância de procrustes completo que também foi utilizada em algumas variações do algoritmo para calcular a estatística de convergência.As nove adaptações propostas apresentaram desempenho semelhante entre si. Quando comparados com um algoritmo pré-existente na literatura, as variações do algoritmos mostraram desempenho melhor nos cenários em que haviam baixa concentração dos dados. Foram feitas três aplicações em base de dados reais, nas quais se destacaram os algoritmos com distância de Battacharyya, Rényi e o algoritmo já existente na literatura. No contexto de dados direcionais, a exemplo de latitude e longitude, propusemos três novas distâncias (Battacharyya, Rényi e Hellinger) baseadas na distribuição Von Mises-Fisher matricial. Além disso, propusemos um algoritmo, contendo as novas distâncias, com seis possíveis variações. Nas análises, iniciamos com a avaliação gráfica do desempenho das distâncias propostas, visando identificar mudanças em uma matriz previamente fixada. Em adição, verificamos o impacto do número de direções nos desempenhos dos algoritmos, bem como suas performances em três cenários com características distintas. Foram realizadas aplicações em base de dados reais com informações tridimensionais de remanências magnéticas de rochas. As seis adaptações obtiveram desempenhos semelhantes, com índices de Rand e coeficientes Kappa variando de tal modo a alcançar valores máximos, nos cenários com dois grupos.

Palavras-chaves: Agrupamento. Análise de formas. Distribuição Bingham. Distâncias. Dados direcionais. Espaço multidimensional. Von Mises-Fisher matricial.

ABSTRACT

Unsupervised methods are used in Statistics to distribute data into distinct groups, where elements within each group exhibit greater similarity. New techniques have been proposed to enhance the effectiveness of clustering in various scenarios. In this work, we propose three distance metrics for directional data based on the matrix von Mises-Fisher distribution, and two algorithms: one for scenarios with low concentration of preform data, considering complex vectors, and another for directional data within the multidimensional Stiefel manifold. The proposed algorithm for clustering preforms offers nine possible variations, due to two stopping criteria and three distance metrics based on the Bingham distribution (Bhattacharyya, Rényi, and Hellinger), along with the full Procrustes distance, which was also used in some algorithm variations to calculate convergence statistics. The nine proposed adaptations demonstrated similar performance. When compared to a pre-existing algorithm in the literature, the variations of the proposed algorithm performed better in scenarios with low data concentration. Three applications on real-world datasets were conducted, in which the algorithms using Bhattacharyya and Rényi distances, and the algorithm already present in the literature, stood out. In the context of directional data, such as latitude and longitude, we proposed three new distance metrics (Bhattacharyya, Rényi, and Hellinger) based on the matrix von Mises-Fisher distribution. Additionally, we proposed an algorithm, incorporating the new distance metrics, with six possible variations. In the analyses, we began with a graphical evaluation of the proposed distances' performance, aiming to identify changes in a previously fixed matrix. Moreover, we examined the impact of the number of directions on the algorithms' performance, as well as their effectiveness across three scenarios with distinct characteristics. Applications on real-world datasets involving three-dimensional magnetic remanence data from rocks were conducted. The six adaptations showed similar performance, with Rand index and Kappa coefficient values varying in such a way as to reach maximum values in scenarios with two groups.

Keywords: Clustering. Shape analysis. Bingham distribution. Distances. Directional data. Multidimensional space. Matrix Von Mises-Fisher.

LISTA DE FIGURAS

Figura 1 – Ilustração de configurações originais, helmertizadas e pré-forma de um crânio de chimpanzé macho. 19

Figura 2 – Ilustração das pré-formas de um grupo de 29 gorilas do sexo masculino (à esquerda) e sua respectiva forma média de procrustes completo (à direita). Dados disponibilizados por Dryden e Mardia (2016). 20

Figura 3 – Baricentros das pré-formas. 33

Figura 4 – Gráficos dos índices de Rand ajustados *versus* números de K vizinhos. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899.900)$, vetores de rotação $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2. 35

Figura 5 – Representação gráfica dos marcos anatômicos de duas crianças da cidade de Ann Arbor, ambas com 8 anos de idade. 39

Figura 6 – Exemplo gráfico dos *landmarks* dos rostos de motoristas captados com a posição da cabeça para frente, direita ou esquerda. Base de dados *DrivFace*, 2016. 40

Figura 7 – Exemplo gráfico dos *landmarks* no cenário 1: motoristas olhando para direita (grupo 1) e motoristas olhando para frente (grupo 2). Base de dados *DrivFace*, 2016. 42

Figura 8 – Exemplo gráfico dos *landmarks* no cenário 2: motoristas olhando para esquerda (grupo 1) e motoristas olhando para frente (grupo 2). Base de dados *DrivFace*, 2016. 42

Figura 9 – Exemplo gráfico dos *landmarks* no cenário 3: motoristas olhando para direita (grupo 1) e motoristas olhando para esquerda (grupo 2). Base de dados *DrivFace*, 2016. 42

Figura 10 – Representação gráfica dos cenários 1 (a), 2 (b) e 3 (c). 53

Figura 11 – Distâncias mudando x_{11}. 56

Figura 12 – Comparação entre as médias dos índices de Rand ajustados para os valores de $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ em vMFR1 e vMFR2, por n números de direções. 57

Figura 13 – Comparação das médias de índices de Rand ajustados por tamanhos amostrais $n_1 = n_2 \in \{5, 10, 15, 20, 30, 50, 100, 150\}$. 58

Figura 14 – Representação gráfica bidimensional e tridimensional das bases de dados reais B2, B6, B7 e B9 (FISHER; LEWIS; EMBLETON, 1993). 69

LISTA DE TABELAS

Tabela 1 – Definições e nomenclatura dos algoritmos propostos. 31

Tabela 2 – Índice de Rand e coeficiente Kappa das variações do algoritmo com base na distância de Rényi. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2. . . . 36

Tabela 3 – Média, mediana e Desvio Padrão (DP) das acurácias, índice de Rand e coeficiente Kappa obtidos nas $MC = 100$ simulações de Monte Carlo, sob as seguintes condições: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2. 37

Tabela 4 – Índice de Rand ajustado - Estatísticas e valores p dos testes de normalidade Shapiro-Wilk - H_0: admite distribuição normal dos dados. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2. 37

Tabela 5 – Índices de Rand ajustados - Estatísticas e valores p de testes Wilcoxon-Man-Whitney - H_1 two-sided. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2. . . . 38

Tabela 6 – Base de dados das crianças com 8 e 14 anos de idade. Índice de Rand ajustado e coeficiente Kappa dos algoritmos de agrupamento. 40

Tabela 7 – Coeficiente Kappa e índice de Rand ajustado dos algoritmos aplicados na base *DrivFace* - Cenários 1, 2, 3 e 4. 41

Tabela 8 – Definições e nomenclaturas das variações do algoritmo proposto. 52

Tabela 9 – Matriz de Confusão - Exemplo para $k = 3$. 54

Tabela 10 – Medidas de desempenho nos cenários 1, 2 e 3, para $n = 1$. 59

Tabela 11 – Medidas de desempenho nos cenários 1, 2 e 3, para $n = 2$. 60

Tabela 12 – Medidas de desempenho nos cenários 1, 2 e 3, para $n = 3$. 60

Tabela 13 – Base de dados B2. 62

Tabela 14 – Base de dados B7. 63

Tabela 15 – Base de dados B6. 64

Tabela 16 – Base de dados B9. 65

Tabela 17 – Base de dados B2 transformados para $\mathbb{R}^3$. 66

Tabela 18 – Base de dados B7 transformados para $\mathbb{R}^3$. 66

Tabela 19 – Base de dados B6 transformados para $\mathbb{R}^3$. 67
Tabela 20 – Base de dados B9 transformados para $\mathbb{R}^3$. 68
Tabela 21 – Estatísticas descritivas das bases B2, B6, B7 e B9 em $\mathbb{R}^3$. 69
Tabela 22 – Desempenho do agrupamento com k=2. 71
Tabela 23 – Tempos de processamento, Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B2, B6 e B7. 72
Tabela 24 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B2, B7 e B9. 73
Tabela 25 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B2, B6 e B9. 74
Tabela 26 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B6, B7 e B9. 75
Tabela 27 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com os quatro grupos: B2, B6, B7 e B9. 76

SUMÁRIO

1 AGRUPAMENTO DE FORMAS PLANAS UTILIZANDO DISTÂNCIAS BASEADAS NA DISTRIBUIÇÃO BINGHAM COMPLEXA . 14

1.1 INTRODUÇÃO 14

1.2 ANÁLISE ESTATÍSTICA DE FORMAS 15

1.2.1 Distância de procrustes para formas planas 19

1.3 DISTRIBUIÇÃO BINGHAM COMPLEXA 21

1.3.1 Simulação de dados com distribuição Bingham complexa 22

1.3.2 Distâncias baseadas na distribuição Bingham 22

1.4 K VIZINHOS MAIS PRÓXIMOS - KNN 24

1.5 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO 25

1.5.1 Análise de agrupamentos 26

1.6 KNN k-MÉDIAS BASEADO EM DISTÂNCIAS PARA DISTRIBUIÇÃO BINGHAM 28

1.7 METODOLOGIA 31

1.8 RESULTADOS NUMÉRICOS 34

1.9 APLICAÇÕES EM BASES DE DADOS REAIS 38

1.9.1 Marcos craniofaciais de crianças 38

1.9.2 Base de dados *DrivFace* 40

1.10 CONSIDERAÇÕES FINAIS 42

2 DISTÂNCIAS PARA DADOS DIRECIONAIS EM ESPAÇO MULTIDIMENSIONAL BASEADAS NA DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL 44

2.1 INTRODUÇÃO 44

2.2 DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL 45

2.2.1 Função Hipergeometrica e Função Geradora de Momentos 47

2.3 DISTÂNCIAS BASEADAS NA DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL 47

2.4 ALGORITMO DE AGRUPAMENTO: NOSSA PROPOSTA 49

2.5 METODOLOGIA 52

2.5.1 Simulação de dados Von Mises-Fisher matriciais 54

2.6 RESULTADOS NUMÉRICOS 56
2.7 BASES DE DADOS REAIS 61
2.8 CONSIDERAÇÕES FINAIS 76
REFERÊNCIAS 79

1 AGRUPAMENTO DE FORMAS PLANAS UTILIZANDO DISTÂNCIAS BASEADAS NA DISTRIBUIÇÃO BINGHAM COMPLEXA

1.1 INTRODUÇÃO

A Estatística possui ferramentas poderosas que podem ser utilizadas em diversas áreas de pesquisa e desenvolvimento de novas tecnologias (SOBHANI et al., 2021; RODRÍGUEZ et al., 2020). Análise de agrupamento, redução de dimensionalidade de dados, métodos de classificação de objetos, regressão e análises descritivas básicas são exemplos de recursos estatísticos amplamente utilizados para construir conhecimento científico (IZBICKI; SANTOS, 2020).

Assim como os algoritmos de classificação, as técnicas em análise de agrupamentos são ferramentas importantes no campo da mineração de dados e reconhecimento de padrões (KAUFMAN; ROUSSEEUW, 2009). Os métodos de análise de agrupamento dividem a amostra em grupos, alocando no mesmo grupo objetos que possuem características semelhantes, e em grupos diferentes, os objetos que possuem características divergentes entre si (IZBICKI; SANTOS, 2020).

Alguns pesquisadores têm utilizado técnicas de agrupamento na área de análises de forma, com a finalidade de categorizar as formas de maneira não supervisionada. A exemplo de Vinué, Simó e Alemany (2016), que utilizaram o algoritmo k-médias para forma em terceira dimensão (denotado por 3D) em uma base de dados com informação de landmarks do corpo de 6013 mulheres europeias, a fim de categorizá-las pelo tamanho e forma corporal. Por sua vez, Huang, Styner e Zhu (2015) implementaram agrupamentos de dados de formas bidimensionais com base em pontos de referência de alta dimensão para agrupar imagem de corpos calosos de crianças com e sem Transtorno de Déficit de Atenção e Hiperatividade (TDAH). Shen et al. (2013) apresentam técnicas para descobrir uma estrutura comum que capte a informação estrutural intrínseca de formas pertencentes ao mesmo grupo, e utiliza bases de dados com formas diversas, dentre elas, formas de objetos inanimados, animais e seres humanos.

De acordo com Srivastava et al. (2005) existem diversas maneiras de caracterizar um objeto em uma imagem, seja por localização, textura, cor, forma ou movimento. Uma abordagem muito comum no tratamento de imagens é o uso de pontos de referência na forma, conhecidos como *landmarks* (DRYDEN; MARDIA, 2016).

Em seu trabalho, Assis, Souza e Amaral (2021) obtiveram bons resultados em agrupamentos, utilizando *landmarks*, ao implementar quatro diferentes critérios de convergência em um

algoritmo k-médias, adaptado para análise de formas planas. Entretanto, no cenário em que há baixa concentração dos dados, tais algoritmos não atuaram bem. Isso é de certa forma esperado, pois nessa condição os grupos encontram-se dispersos ao ponto de misturarem-se parcialmente entre si, dificultando o agrupamento. Considerando esse cenário, achamos coerente fazer o uso de divergências com abordagens diferentes na adaptação do k-médias, na tentativa de melhorar o agrupamento. A abordagem considerada foi a distância entre amostras, tomando como base os K-vizinhos mais próximos de cada observação e centroides. Para tanto, considerou-se o estudo de Lima, Nascimento e Amaral (2021), Félix (2019), que apresenta em sua tese a divergência de Rényi, Bhattacharyya e Hellinger derivadas para a distribuição Bingham, aplicadas ao contexto de *landmarks*. E, como se tratam de divergências utilizadas para comparação entre amostras de distribuições, dentro do algoritmo adaptado k-médias, fez-se uso de um algoritmo de captação dos vizinhos mais próximos (KNN).

Tomando um cenário de baixa concentração, o principal objetivo desta pesquisa foi avaliar o desempenho dos algoritmos de agrupamento k-médias adaptados, utilizando as divergências propostas por Félix (2019), aliadas ao KNN, considerando critérios como o índice de Rand ajustado e o coeficiente Kappa. Além disso, também é apresentada uma análise sobre as médias de tais índices, e aplicação em base de dados reais.

O Capítulo 1 desta tese está organizado da seguinte forma: A Seção 1.2 apresenta uma breve explicação sobre os principais conceitos na análise Estatística de formas; a Seção 1.3 traz uma breve revisão sobre a distribuição de Bingham complexa, bem como as divergências baseadas na mesma; o método dos K vizinhos mais próximos (KNN) é explanado na Seção 1.4; na Seção 1.5 é apresentada uma breve revisão sobre inteligência artificial e aprendizado de máquina não supervisionado; a Seção 1.6 descreve o algoritmo de agrupamento proposto e suas variações; a Seção 1.7 delineia a metodologia aplicada; os resultados numéricos obtidos são apresentados e discutidos na Seção 1.8, seguidos pelas aplicações em bases de dados reais, na Seção 1.9; as considerações finais do Capítulo 1 são apresentadas na Seção 1.10.

1.2 ANÁLISE ESTATÍSTICA DE FORMAS

A área de análise estatística de formas (*Statistical Shape analysis - SSA*) tem como objetivo principal tratar de formas com o fator da aleatoriedade presente. Para tanto, se faz imprescindível o uso de métodos inferenciais, a fim de obter informações relevantes sobre uma população, por meio de uma amostra, parte dessa mesma população. Assim, técnicas as

quais minimizem o erro de estimação dos parâmetros de interesse na análise de amostras são ferramentas essenciais (DRYDEN; MARDIA, 2016).

O conceito de forma é uma das definições importantes dentro da análise estatística de formas: forma são todas as informações geométricas que permanecem quando os efeitos de localização, escala e rotação são removidos de um objeto (DRYDEN; MARDIA, 2016). Além disso, entre as formas oriundas de mesma população, há pontos finitos que se correspondem, sendo estes chamados pontos de referência, marcos anatômicos ou *landmarks*.

Em seu livro, Bookstein (1997) relata três tipos comuns de landmarks utilizados na morfometria multivariada[1], são eles:

- Justaposições de tecidos - envolve pontos no espaço de estudo onde estruturas se encontram (Exemplo: as suturas ósseas sob a ponte do nariz em humanos);
- Máximos de curvatura - estruturas anatômicas que incluem pontas de extrusões e vales de invaginações;
- Pontos extremos - são aquelas cujas definições se referem a informações em locais diversos e finitamente separados. Esta categoria envolve pontos finais de diâmetros, centroides, interseções de segmentos, pontos mais distantes. É a mais comum em morfometria multivariada.

Outra forma de categorizar os pontos de referência é dada por Dryden e Mardia (1998):

- Marcos anatômicos ou científicos - São os pontos atribuídos ao objeto por um especialista, tais quais se correspondem entre os organismos homólogos de alguma maneira biologicamente significativa;
- Marcos matemáticos - São pontos identificados nos objetos de acordo com alguma propriedade matemática ou geométrica;
- Pseudo-marcos - São pontos localizados entre os marcos anatômicos, seguindo o contorno e a curvatura do objeto.

Tomando a classificação de *landmarks* definida por Dryden e Mardia (1998), nesta tese foram utilizados marcos anatômicos e matemáticos. As bases de dados utilizadas possuem pontos de referências previamente definidos pelos profissionais da área que os coletaram.

[1] Dryden e Mardia (2016) definem morfometria como o estudo de objetos ou sujeitos pertencentes à mesma população, aplicando técnicas estatísticas multivariadas, utilizando distâncias entre pontos de referências ou ângulos.

Os *landmarks* são comumente representados por configuração. Uma configuração é a representação matemática de pontos de referência em um objeto, e, uma matriz de configuração $\mathbf{X}$ corresponde a uma matriz que contém as coordenadas cartesianas dos marcos, e possui dimensão $(k+1) \times m$, na qual $(k+1)$ é o numero de *landmarks* e m, o numero dimensões. Um espaço de configuração é o espaço de todas as possíveis coordenadas dos marcos. No capítulo 1 desta tese consideramos o espaço de formas planas, portanto, foram utilizadas matrizes de configurações com m igual a 2.

$$\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2]^\top = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{(k+1),1} & x_{(k+1),2} \end{pmatrix} \tag{1.1}$$

Kent (1994) relata que distribuições simétricas no conjunto dos números complexos se tornam uma ferramenta natural no tratamento e análises de configurações bidimensionais. Considerando tal fato, comumente se realiza uma transformação nos dados reais, trazendo-os para a esfera complexa. Assim, no contexto das formas planas, é possível representar a matriz de configuração por meio de um vetor complexo $\mathbf{z}^0$ (configuração complexa), no qual $\mathbf{x}_1$ corresponde aos valores reais e $\mathbf{x}_2$ aos complexos:

$$\mathbf{z}^0 = [\mathbf{z}_1^0, \mathbf{z}_2^0, ..., \mathbf{z}_{(k+1)}^0]^\top = [x_{1,1} + ix_{1,2}, ..., x_{(k+1),1} + ix_{(k+1),2}]^\top, \tag{1.2}$$

para $i = \sqrt{-1}$.

As coordenadas complexas precisam receber um tratamento antes da análise estatística, e para isso existem algumas técnicas conhecidas, como o cálculo das coordenadas de Bookstein, coordenadas de Kent ou coordenadas de Kendall (DRYDEN; MARDIA, 2016). O objetivo principal é retirar os efeitos de rotação, escala e locação, para que reste apenas as formas propriamente ditas. Neste trabalho foram utilizadas as coordenadas de Kendall pois são as que mais se encontram em aplicações na literatura. Para realizar o cálculo é preciso utilizar a submatriz de Helmert, a qual pode ser vista na Definição 1.2.1.

Definição 1.2.1. *A **matriz completa de Helmert** $\mathbf{H}^F$ é uma matriz $(k+1) \times (k+1)$ ortogonal em que a primeira linha possui todos os elementos iguais a $1/\sqrt{k+1}$ e a j-ésima linha, para $j = 2, 3, ..., (k+1)$, é dada por $(h_j, ..., h_j, -(j-1)h_j, 0, ..., 0)$, na qual $h_j =$*

$-[j(j-1)]^{-1/2}$. *Assim, a* j*-ésima linha de* $\mathbf{H}^F$ *é composta por* h_j *repetido* $(j-1)$ *vezes, seguido por um* $-(j-1)h_j$ *e* $(k-j)$ *zeros.*

A ***submatriz de Helmert*** $\mathbf{H}$ *corresponde à matriz completa de Helmert* $\mathbf{H}^F$ *sem a primeira linha.*

Exemplo 1.2.1. *Para* $(k+1)=4$*, a matriz completa de Helmert é dada por:*

$$\mathbf{H}^F = \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{bmatrix},$$

e a submatriz de Helmert resultante é

$$\mathbf{H} = \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{bmatrix}.$$

Com a utilização da submatriz de Helmert, é possível retirar o efeito de locação do vetor complexo $\mathbf{z}^0$, resultando no que se conhece como configuração Helmertizada:

$$\mathbf{w} = \mathbf{H}\mathbf{z}^0. \tag{1.3}$$

Posteriormente, calcula-se o vetor de pré-formas $\mathbf{z}$.

Definição 1.2.2. *A* ***pré-forma*** *de uma configuração complexa* $\mathbf{z}^0$ *é dada por*

$$\mathbf{z} = \frac{\mathbf{H}\mathbf{z}^0}{\sqrt{(\mathbf{H}\mathbf{z}^0)^* \mathbf{H}\mathbf{z}^0}}, \tag{1.4}$$

em que $(\cdot)^*$ *representa a operação de transposto conjugado. Assim,* $(\mathbf{H}\mathbf{z}^0)^*$ *é o transposto conjugado de* $\mathbf{H}\mathbf{z}^0$*.*

Um vetor complexo de pré-formas é invariante sob locação e escala da configuração original e possui dimensão $k \times 1$*.*

O espaço de pré-formas denotado por $\mathbf{S}_m^k$ *(*k *landmarks e* m *dimensões) é o espaço de todas as pré-formas possíveis, ou seja, de todos os possíveis vetores de dimensão* k *que não possuem informação de locação nem de escala. No caso das formas planas, esse espaço corresponde a uma hiperesfera complexa de dimensão* k*:*

$$\mathbb{C}\mathbf{S}^{k-1} = \{\mathbf{z} \in \mathbb{C}^k; \mathbf{z}^*\mathbf{z} = 1\}, \tag{1.5}$$

em que $\mathbb{C}^k$ é o espaço complexo de dimensão k.

A Figura 1 ilustra uma configuração individual com $k+1$ igual a 8 marcos anatômicos do crânio de um chimpanzé (DRYDEN; MARDIA, 2016). Note que a configuração Helmertizada possui $k = 7$ pontos de referência e está centralizada (sem efeito de locação), e, que a pré-forma não contém efeito de locação nem de escala.

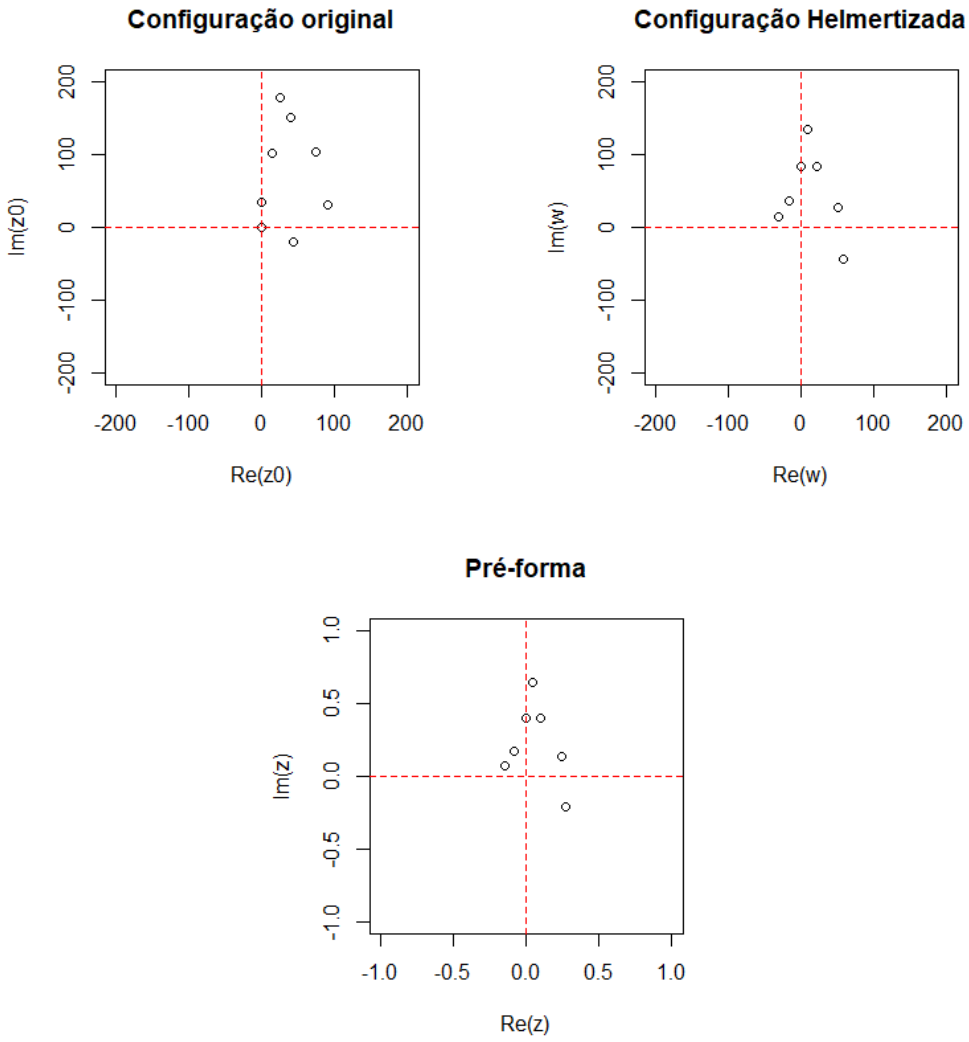


Figura 1 – Ilustração de configurações originais, helmertizadas e pré-forma de um crânio de chimpanzé macho.

1.2.1 Distância de procrustes para formas planas

Tratando-se de amostras aleatórias, existem alguns conceitos e medidas relevantes no contexto de formas planas. Dentre elas, encontram-se a forma média de procrustes e algumas distâncias associadas (DRYDEN; MARDIA, 2016).

Definição 1.2.3. *Seja* $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ *uma amostra aleatória de matrizes de configurações, com tamanho* n*, e,* $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n$ *suas respectivas pré-formas. A* ***forma média de procrustes completo*** $\hat{\boldsymbol{\mu}}$ *corresponde ao autovetor associado ao maior autovalor da matriz produto*

$$\mathbf{S} = \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^*, \tag{1.6}$$

em que n *é o tamanho amostral, e* $\mathbf{z}_i$*, para* $i = 1, 2, ..., n$*, os vetores complexos de pré-formas (KENT, 1994).*

A Figura 2 ilustra uma representação gráfica das pré-formas para uma amostra de 29 gorilas do sexo masculino e exemplifica como a forma média de procrustes completo (gráfico com pontos em azul) representa o conjunto de dados da qual originou-se.

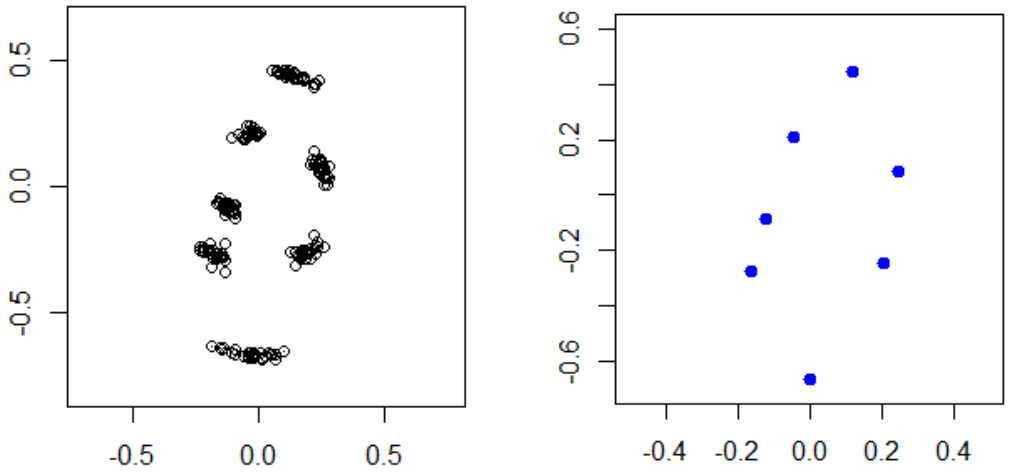


Figura 2 – Ilustração das pré-formas de um grupo de 29 gorilas do sexo masculino (à esquerda) e sua respectiva forma média de procrustes completo (à direita). Dados disponibilizados por Dryden e Mardia (2016).

A forma média de procrustes completo é uma medida interessante para descrever um padrão entre os landmarks num dado grupo de formas. Entretanto, para comparar formas entre si, é necessário o uso de outras medidas, mais especificamente, as distâncias.

Em seu livro sobre aplicações de análises estatísticas de formas em ambiente de programação R, Dryden e Mardia (2016) trazem algumas distâncias bastante utilizadas nesta área: distância de Procrustes completo, distância de Procrustes parcial, distância de forma Riemanniana, distância de tamanho e forma Riemanniana/Procrustes. Neste trabalho, a distância de procrustes completo, descrita a seguir, foi utilizada no espaço das pré-formas pela sua simplicidade em cômputo.

Definição 1.2.4. *Sejam* $\mathbf{Y}_1$ *e* $\mathbf{Y}_2$ *duas configurações com* $\mathbf{z}_1$ *e* $\mathbf{z}_2$ *como suas respectivas pré-formas. No contexto das formas planas, a* ***distância de procrustes completo*** *entre* $\mathbf{z}_1$ *e*

$\mathbf{z}_2$ *é dada por:*

$$d_{FP}(\mathbf{z}_1, \mathbf{z}_2) = \left[1 - \left(\frac{\mathbf{z}_1^* \mathbf{z}_2 \mathbf{z}_2^* \mathbf{z}_1}{\mathbf{z}_2^* \mathbf{z}_2 \mathbf{z}_1^* \mathbf{z}_1}\right)^2\right]^{\frac{1}{2}}. \tag{1.7}$$

1.3 DISTRIBUIÇÃO BINGHAM COMPLEXA

Para lidar com dados de natureza complexa e aplicar as técnicas estatísticas inferenciais é necessário considerar uma distribuição de probabilidade que consiga descrever o comportamento de tais dados. Dryden e Mardia (1998) apresentam algumas distribuições no contexto formas planas, dentre elas, a uniforme complexa, Bingham complexa, Watson complexa e a gaussiana angular central complexa. Neste trabalho tomamos a Bingham complexa como base no âmbito das pré-formas, por ser uma das principais utilizadas no contexto de formas planas e já possuir propriedades muito bem conhecidas na literatura.

Considerando a esfera unitária complexa $\mathbb{C}S^{k-1} = \{\mathbf{z} = (z_1, ..., z_k)^\top : \sum |z_j|^2 = 1\} \subset \mathbb{C}^k$, para $k \geq 2$, a distribuição complexa de Bingham tem função densidade de probabilidade (f.d.p.)

$$f(\mathbf{z}) = c(\mathbf{A})^{-1} \exp\{\mathbf{z}^\star \mathbf{A} \mathbf{z}\}, \tag{1.8}$$

na qual $\mathbf{A}$ é uma matriz de parâmetros, hermitiana, ou seja $\mathbf{A} = \mathbf{A}^*$, de dimensão $k \times k$ e $c(\mathbf{A})$ é uma constante de normalização (DRYDEN; MARDIA, 2016; KENT, 1994). Assim, dizemos que $\mathbf{z} \sim \mathbb{C}_{k-1}B(\mathbf{A})$, ou seja, o vetor $\mathbf{z}$ segue distribuição Bingham complexa com matriz de parâmetros $\mathbf{A}$.

Sejam $\lambda_1 < \lambda_2 < ... < \lambda_k = 0$ os autovalores de $\mathbf{A}$ e $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, ..., \boldsymbol{\gamma}_k$ seus autovetores associados, então $c(\mathbf{A})$ é dada por

$$c(\mathbf{A}) = 2\pi^k \sum_{i=1}^{k} a_i \exp \lambda_i, \tag{1.9}$$

em que $a_i = \left[\prod_{j \neq i}(\lambda_i - \lambda_j)\right]^{-1}$. Compreende-se que $c(\mathbf{A}) = c(\mathbf{\Lambda})$ dependem apenas dos autovalores de $\mathbf{A}$, sendo $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_k)$. Quando partimos para o contexto inferencial, os autovalores de $\mathbf{A}$ podem ser estimados por meio da matriz produto $\mathbf{S}$, em que $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n$, $n \geq k$, sejam amostras aleatórias que seguem distribuição complexa de Bingham. As provas de tais resultados podem ser vistas no livro escrito por Dryden e Mardia (2016).

Sejam $0 < l_1 < l_2 < ... < l_k$ os autovalores da matriz produto $\mathbf{S}$ e $\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_k$ seus autovetores correspondentes e, sabendo que $\sum_j l_j = n$, os estimadores de máxima verossimilhança

(EMV) de $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, ..., \boldsymbol{\gamma}_k$ e $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_k)$ são dados, respectivamente, por

$$\widehat{\boldsymbol{\gamma}}_j = \mathbf{g}_j, \text{ para } j = 1, ..., k, \text{ e } \widehat{\lambda}_j \cong -n/l_j, \text{ para } j = 1, ..., k-1.$$

1.3.1 Simulação de dados com distribuição Bingham complexa

Para simular os dados que seguem distribuição Bingham complexa foi utilizado o método de truncamento pelo Simplex (KENT; CONSTABLE; ER, 2004). Em seu artigo, os autores mostram como simular dados com distribuição Bingham complexa utilizando a distribuição exponencial multivariada truncada. Para tanto, faz-se necessário gerar amostras aleatórias de $k-1$ exponenciais truncadas $T\exp(\lambda_j)$, como mostra o Algoritmo 1.

Algoritmo 1: Simulação da distribuição exponencial truncada

Result: Retorna um vetor de amostras aleatórias $\boldsymbol{S}$ que seguem distribuição exponencial truncada

1: Simule uma variável aleatória uniforme $U_j \sim U[0,1]$, para $j = 1, ..., k$;

2: Seja $S\prime_j = -(1/\lambda_j)\log(1 - U_j(1 - e^{-\lambda_j}))$, $\boldsymbol{S\prime} = (S\prime_1, ..., S\prime_k)^\top$;

3: Se $\sum_{j=1}^{k} S\prime_j < 1$, então $\boldsymbol{S} = \boldsymbol{S\prime}$. Caso contrário, rejeita-se $\boldsymbol{S\prime}$ e retorna ao passo 1:

O Algoritmo 2 descreve o procedimento para a geração de dados.

Algoritmo 2: Simulação da distribuição exponencial truncada

Result: Retorna um vetor $\mathbf{z} = (z_1, ..., z_k)^\top$ de amostras aleatórias que seguem distribuição Bingham complexa

1: Gere $\boldsymbol{S} = (S_1, ..., S_k)^\top$, em que $S_j \sim T\exp(\lambda_j)$, para $j = 1, ..., k$ (Algoritmo 1);

2: Se $\sum_{j=1}^{k} S\prime_j < 1$, escreva $S_k = 1 - \sum_{j=1}^{k} S\prime_j$. Caso contrário, volte ao passo 1:;

3: Gere ângulos independentes $\theta_j \sim U[0, 2\pi)$, $j = 1, ..., k$;

4: Calcule $z_j = S_j^{1/2}\exp\{i\theta_j\}$, $j = 1, ..., k$

1.3.2 Distâncias baseadas na distribuição Bingham

Dentre outros resultados de sua tese, Félix (2019) deriva distâncias entre dados que seguem a distribuição de probabilidade Bingham, baseadas nas divergências de Rényi, Hellinger e Battacharyya.

Nesta tese, chamaremos de "divergência" qualquer função $D(\cdot)$ aplicada entre dois vetores aleatórios ($\mathbf{x}$ e $\mathbf{y}$, por exemplo), para $\mathbf{x}, \mathbf{y} \in \mathcal{Z} \subseteq \mathbb{C}^p$ e $p \in \mathbb{Z}_+$, que satisfaça as propriedades:

$\Rightarrow$ Positividade - $D(\mathbf{x}||\mathbf{y}) \geq 0$, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{Z} \subseteq \mathbb{C}^p$;

$\Rightarrow$ Definitude - $D(\mathbf{x}||\mathbf{y}) = 0$, se, e somente se,$\mathbf{x} = \mathbf{y}$.

Quando, além dessas propriedades, a função possui também:

$\Rightarrow$ Simetria - $D(\mathbf{x}||\mathbf{y}) = D(\mathbf{y}||\mathbf{x})$, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{Z} \subseteq \mathbb{C}^p$,

dizemos que tal medida é uma "distância" (BARTLE; SHERBERT, 2000).

Dentre outros resultados de sua tese, Félix (2019) utiliza a divergência de Rényi de ordem β, $\beta \in (0, 1)$, para derivar uma distância baseada na Bingham complexa, além das distâncias de Bhattacharyya e Hellinger, também com base na distribuição de probabilidade Bingham complexa, como apresenta a Definição 1.3.1.

Definição 1.3.1. *Sejam $f_{\boldsymbol{x}_1}(\mathbf{z}; \boldsymbol{\theta}_1)$ e $f_{\boldsymbol{x}_2}(\mathbf{z}; \boldsymbol{\theta}_2)$ funções densidade de probabilidade de $\mathbf{X}_1$e $\mathbf{X}_2$, e, seus respectivos parâmetros $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$. Segue as definições das divergências de Rényi de ordem β, Hellinger e Bhattacharyya:*

A divergência que resulta na distância de Rényi de ordem β é dada por:

$$D_R^{\beta}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2) = \frac{1}{\beta - 1} \log \mathbb{E}_{\boldsymbol{\theta}_1} \left[\left(\frac{f_{\boldsymbol{x}_1}(\mathbf{z}; \boldsymbol{\theta}_1)}{f_{\boldsymbol{x}_2}(\mathbf{z}; \boldsymbol{\theta}_2)} \right)^{\beta - 1} \right]. \tag{1.10}$$

O fato da divergência de Rényi $D_R^{\beta}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2)$ não ser simétrica traz a necessidade de realizar alguma transformação, para que assim, possa resultar em uma medida de distância. Chung et al. (1989) propôs a simetrização, que será utilizada a seguir, e que resulta na distância de Rényi de ordem β, para $\beta \in (0, 1)$:

$$d_R^{\beta}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{\beta - 1} \log \left\{ \frac{\exp[(\beta - 1) D_R^{\beta}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2)] + \exp[(\beta - 1) D_R^{\beta}(\boldsymbol{\theta}_2||\boldsymbol{\theta}_1)]}{2} \right\}. \tag{1.11}$$

As distâncias de Bhattacharyya e Hellinger, por sua vez, são dadas respectivamente pelas equações a seguir:

$$d_B(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -\log \mathbb{E}_{\boldsymbol{\theta}_1} \left[\sqrt{\frac{f_{\boldsymbol{x}_2}(\mathbf{z}; \boldsymbol{\theta}_2)}{f_{\boldsymbol{x}_1}(\mathbf{z}; \boldsymbol{\theta}_1)}} \right], \tag{1.12}$$

$$d_H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 1 - \mathbb{E}_{\boldsymbol{\theta}_1} \left[\sqrt{\frac{f_{\boldsymbol{x}_2}(\mathbf{z}; \boldsymbol{\theta}_2)}{f_{\boldsymbol{x}_1}(\mathbf{z}; \boldsymbol{\theta}_1)}} \right]. \tag{1.13}$$

A partir das distâncias pré-estabelecidas, Félix (2019) considerou as funções densidade de probabilidade da Bingham complexa e derivou as distâncias para tal contexto. As provas do teorema a seguir podem ser consultadas em sua tese.

Teorema: Seja $\boldsymbol{x}_1 \sim \mathbb{CB}_{k-1}(\mathbf{A}_1)$ e $\boldsymbol{x}_2 \sim \mathbb{CB}_{k-1}(\mathbf{A}_2)$ com f.d.p. $f_{\boldsymbol{x}_1}(\mathbf{z};\mathbf{A}_1)$ e $f_{\boldsymbol{x}_2}(\mathbf{z};\mathbf{A}_2)$, respectivamente. Sob tais condições, as distâncias Rényi (R), com parâmetro de ordem $\beta \in (0,1)$, Hellinger (H) e Battacharyya (B) são dadas por:

$$d_R^{\beta}(\mathbf{A}_1,\mathbf{A}_2) = \frac{1}{\beta-1}\log\left\{\frac{c^{-\beta}(\boldsymbol{\Lambda}_1)c^{(\beta-1)}(\boldsymbol{\Lambda}_2)c(\boldsymbol{\Lambda}_{R12}) + c^{-\beta}(\boldsymbol{\Lambda}_2)c^{(\beta-1)}(\boldsymbol{\Lambda}_1)c(\boldsymbol{\Lambda}_{R21})}{2}\right\}, \tag{1.14}$$

$$d_H(\mathbf{A}_1,\mathbf{A}_2) = 1 - [c(\boldsymbol{\Lambda}_1)c(\boldsymbol{\Lambda}_2)^{-1/2}c(\boldsymbol{\Delta}_3)], \tag{1.15}$$

$$d_B(\mathbf{A}_1,\mathbf{A}_2) = \frac{1}{2}\log[c(\boldsymbol{\Lambda}_1)c(\boldsymbol{\Lambda}_2)] - \log[c(\boldsymbol{\Delta}_3)], \tag{1.16}$$

em que $\boldsymbol{\Lambda}_i$, $i \in \{1,2\}$ é a matriz de autovalores de $\mathbf{A}_i$, $\boldsymbol{\Lambda}_{R12}$ e $\boldsymbol{\Lambda}_{R21}$ são as matrizes de autovalores de $[\beta\mathbf{A}_1 + (1-\beta)\mathbf{A}_2]$ e $[\beta\mathbf{A}_1 + (1-\beta)\mathbf{A}_2]$, respectivamente, e $\boldsymbol{\Delta}_3$ corresponde à matriz de autovalores de $\frac{1}{2}(\mathbf{A}_1 + \mathbf{A}_2)$.

Tanto $\mathbf{A}$ quanto $\mathbf{A} + \Upsilon\mathbf{I}_k$, $\Upsilon \in \mathbb{R}$, indicam uma mesma distribuição $\mathbb{CB}$, de acordo com Kent, Constable e Er (2004). Isso significa, que sob tais condições é possível realizar um deslocamento nos autovalores sem que haja perda de generalidade. Assim, com a finalidade de obter apenas autovalores positivos, na etapa de simulação considerou-se $-\mathbf{A}$ e seus autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k = 0$.

A próxima seção mostra a metodologia utilizada com o intuito de encontrar os vizinhos mais próximos de cada ponto oriundo de uma amostra que segue distribuição Bingham complexa. O objetivo é definir subgrupos, tendo como base cada ponto da amostra, e calcular as distâncias entre eles por meio das equações (1.14), (1.15) ou (1.16).

1.4 K VIZINHOS MAIS PRÓXIMOS - KNN

Considerando que as distâncias aqui utilizadas são medidas construídas para calcular o nível de divergências entre distribuições de amostras, as quais originam matrizes complexas, não é possível calcular tais distâncias dispondo apenas de dois vetores distintos de pré-formas, pois há cômputo de decomposição em autovalores nas divergências. Portanto, propomos a utilização de um método de busca de K vizinhos mais próximos (KNN), utilizando a distância de procrustes completo, que se adequa a tal situação. Tal método foi incorporado dentro do

algoritmo de k-médias para construção das amostras no processo iterativo de agrupamento (tal metodologia será melhor detalhada na Seção 1.6).

Um algoritmo KNN considera as similaridades entre objetos que possuem maior probabilidade de pertencimento ao mesmo grupo, e assim, definindo o valor de K, para K $\leq n$, é possível captar os K objetos mais próximos do que foi tomado como referência.

Seja $\mathcal{N}_{\mathbf{x}}$ o conjunto das K observações de uma amostra de tamanho n que estão, por algum critério de similaridade, mais próximas de $\mathbf{x}$, ou seja,

$$\mathcal{N}_{\mathbf{x}} = \left\{ \mathbf{x}_j : j \in \{1, ..., n\}, d(\mathbf{x}_j, \mathbf{x}) \leq d^{\mathsf{K}}(\mathbf{x}_j, \mathbf{x}) \right\}, \tag{1.17}$$

em que $d^{\mathsf{K}}(\mathbf{x}_j, \mathbf{x})$ é a distância do K-ésimo vizinho mais próximo de $\mathbf{x}_j$ a $\mathbf{x}$ (IZBICKI; SANTOS, 2020).

Para melhor exemplificar o método, considere uma amostra com n observações, e $\mathbf{x}$ pertence à amostra. A distância de procrustes completo entre $\mathbf{x}$ e todos os outros elementos ($\mathbf{x}_j$) é calculada. Organizando as distâncias em uma lista crescente de valores, os K vizinhos mais próximos serão os K objetos relacionados às primeiras K distâncias da lista de valores.

Considerando o contexto das pré-formas explanado na Seção 1.2, o conjunto dos K vizinhos mais próximos é dado por:

$$\mathcal{N}_{\mathbf{z}_i} = \left\{ \mathbf{z}_i, \mathbf{z}_j : i, j \in \{1, ..., n\}, i \neq j, d(\mathbf{z}_j, \mathbf{z}_i) \leq d^{\mathsf{K}}_{FP}(\mathbf{z}_j, \mathbf{z}_i) \right\}, \tag{1.18}$$

em que $d^{\mathsf{K}}_{FP}(\cdot)$ corresponde à distância de procrustes completo do K-ésimo vizinho mais próximo de $\mathbf{z}_i$, e, $\mathbf{z}_j, \mathbf{z}_i$, para $i \neq j$, são os vetores de pré-formas da amostra. Como $\mathbf{z}_i$, para $i = 1, ..., n$, provém de distribuição Bingham, e $n \geq k$, então, $k \leq \mathsf{K} \leq n$.

1.5 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA NÃO SUPERVISIONADO

A inteligência artificial (IA) é uma importante área que se dedica ao desenvolvimento de sistemas computacionais capazes de executar tarefas que normalmente exigiriam inteligência humana (HASTIE et al., 2009). Existem várias abordagens e técnicas utilizadas nesse campo, incluindo aprendizado de máquina, processamento de linguagem natural, redes neurais, algoritmos evolutivos, entre outros. O aprendizado de máquina é uma subárea importante da IA, na qual os sistemas são treinados com conjuntos de dados para melhorar seu desempenho em tarefas específicas. Bishop e Nasrabadi (2006) mencionam que o reconhecimento de padrões

visa descobrir regularidades automaticamente nos dados, por meio do uso de algoritmos, e utilizar essas regularidades para ações como categorização dos dados.

Existem dois tipos distintos de aprendizado de máquina: o supervisionado e o não-supervisionado. Imagine um estudante fazendo uma prova em companhia de seu professor, o qual pode indicar se as respostas da prova estão corretas ou não. Esta é a ideia do aprendizado supervisionado, onde os algoritmos são treinados com conjuntos de dados rotulados, nos quais as respostas desejadas são conhecidas. Já o caso contrário seria o aprendizado não supervisionado, onde o estudante faria a prova sozinho, sem acesso às respostas corretas para comparação. No aprendizado de máquina não supervisionado ocorre a análise dos conjuntos de dados não rotulados, nos quais as respostas de interesse são desconhecidas. O objetivo é encontrar padrões, estruturas ou agrupamentos intrínsecos nos dados. Algoritmos de agrupamento, ou clustering, são exemplos de ferramentas de aprendizado de máquina não supervisionadas (HASTIE et al., 2009; KAUFMAN; ROUSSEEUW, 2009; BISHOP; NASRABADI, 2006).

1.5.1 Análise de agrupamentos

A área de análise de agrupamento engloba os métodos para reconhecimento de padrões e classificação de dados não categorizados. Um dos mais conhecidos é o k-médias, que separa a amostra em k grupos de maneira iterativa, minimizando a variância total dentro dos grupos (IZBICKI; SANTOS, 2020; FRIEDMAN et al., 2001).

Um outro método que comumente apresenta bons resultados para encontrar os agrupamentos que minimizam a variância total dentro dos grupos é o Algoritmo de Lloyd (IZBICKI; SANTOS, 2020; FRIEDMAN et al., 2001). A partir dele, é possível obter outras variações, as quais podem ser encontradas na literatura, a exemplo do k-médias$++$, apresentado por Arthur e Vassilvitskii (2006), o k-medoids, discutido como tópico no livro de Friedman et al. (2001), ou até mesmo o kernel k-médias, apresentado por Jayasumana et al. (2013). Nesta tese também propomos adaptações do algoritmo padrão de Lloyd (ver Algoritmo 3, implementando as distâncias propostas por Félix (2019) em conjunto com o método KNN.

Considere $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ vetores amostrais de tamanho k. Sejam $C_1, C_2, ..., C_\mathrm{k}$ partições dos n elementos amostrais, para k $\in \mathbb{N}^*$, e $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_\mathrm{k}$ seus respectivos centroides. Nesse contexto, o Algoritmo de Lloyd é dado por:

Algoritmo 3: Algoritmo de Lloyd

Result: Retorna $C_1, ..., C_{\mathsf{k}}$

1: Escolha aleatoriamente k centroides $\mathbf{c}_1, ..., \mathbf{c}_{\mathsf{k}}$;

2: **repeat**

3: Calcule a distância $d(\mathbf{x}_i, \mathbf{c}_j)$, para $i = 1, ..., n$ e $j = 1, ..., \mathsf{k}$;

4: Defina o cluster $C_j = \{\mathbf{x}_i : \arg\min_r d(\mathbf{x}_i, \mathbf{c}_r) = r\}$, para $r \in \{1, 2, ..., \mathsf{k}\}$;

5: Calcule os novos centroides a partir de C_j;

until *Convergência*;

Note que o Algoritmo de Lloyd é um método que dá espaço para adaptações e pode gerar inúmeros outros algoritmos. É possível definir qual distância será utilizada, como serão calculados os centroides e qual o critério de convergência será considerado.

Um dos critérios disponíveis na literatura é baseado na estatística de similaridade proposta por Goodall (1991), que pode ser calculada como se segue:

$$J_{\mathsf{G}}((\mathbf{z}_1, ..., \mathbf{z}_n), (\mathbf{c}_1, ..., \mathbf{c}_{\mathsf{k}})) = g(n, k)\frac{\sum_{j=1}^{\mathsf{k}} d_{FP}(\mathbf{c}_j, \hat{\mu})}{\sum_{j=1}^{\mathsf{k}} \sum_{i \in C_j} d_{FP}(\mathbf{z}_i, \mathbf{c}_j)}, \tag{1.19}$$

em que $g(n, \mathsf{k}) = \frac{n(n-1)\mathsf{k}}{\mathsf{k}-1}$, $\hat{\mu}$ é a forma média de procrustes completo de toda a amostra.

Considerando que o algoritmo trabalha de maneira iterativa, é necessário estabelecer, além do critério de convergência por meio da estatística de similaridade, o número máximo de iterações permitidas, para o caso de não haver convergência pela estatística.

No contexto de formas bidimensionais, um método de agrupamento existente na literatura é o K-médias para formas planas (KMPS), apresentado por Assis, Souza e Amaral (2021). O algoritmo consiste em uma adaptação do que foi explanado por Macqueen (1967), onde a distância euclidiana é substituída pela de procrustes, procrustes completo ou parcial (ver Algoritmo 4). Além disso, existem quatro possíveis critérios de agrupamento apresentados no mesmo artigo, dos quais escolhe-se um para seguir com o agrupamento – para mais detalhes, ver Assis, Souza e Amaral (2021).

Algoritmo 4: Algoritmo KMPS

Result: Retorna k grupos

1: Escolha aleatoriamente k distintos centroides e gere uma partição inicial;

2: Calcule a forma média de procrustes completo para cada grupo;

3: **repeat**

4: Aloque cada objeto no grupo em que a forma média for mais próxima, baseado nas distâncias de procrustes, procrustes completo ou procrustes parcial;

5: Calcule o critério de agrupamento escolhido;

6: Calcule a forma média dos novos centroides;

until *Convergência*;

A próxima seção detalha as variações do novo algoritmo que utiliza as distâncias baseadas na distribuição Bingham, apresentadas na subseção 1.3.2.

1.6 KNN k-MÉDIAS BASEADO EM DISTÂNCIAS PARA DISTRIBUIÇÃO BINGHAM

Nesta seção será apresentado o algoritmo baseado nas distâncias para distribuições Bingham, o qual propusemos para o contexto de agrupamento de formas. O objetivo do mesmo é realizar o agrupamento de vetores de pré-forma, formando subgrupos para cada um deles, juntamente com seus K vizinhos, e utilizar as distâncias propostas por Félix (2019) para formar os grupos propriamente ditos, bem como no cálculo das estatísticas dos critérios de convergência.

Iniciaremos estabelecendo os seguintes critérios de convergência: critério de parada 1 e critério de parada 2.

Seja t o número de iterações decorridas e T o número máximo de iterações pré-estabelecidas. Foram utilizados dois tipos de critérios de parada nos algoritmos:

- **Critério de parada 1** - quando se alcança o número máximo de iterações pré-estabelecidas $(t = T)$ ou quando o valor absoluto da diferença entre o critério de convergência da iteração anterior $(J^{(t-1)})$ e o atual $(J^{(t)})$ é menor que um ε muito pequeno, i.e. $|J^{(t-1)} - J^{(t)}| < \varepsilon$;

- **Critério de parada 2** - quando se alcança o número máximo de iterações pré-estabelecidas $(t = T)$ ou quando atingi-se uma iteração em que não há mudança do número de elementos nos grupos da iteração anterior e da atual, i.e. $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-2)} \cap C_j^{(t-1)}) = \sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-1)} \cap C_j^{(t)})$.

Nesta pesquisa, a estatística de Goodall foi utilizada como base e adaptada com as distâncias de Hellinger, Rényi e Bhattacharyya, denotadas aqui por $J_H(\cdot)$, $J_R^\beta(\cdot)$ e $J_B(\cdot)$, respectivamente. Para tanto, fez-se necessário utilizar o método dos K vizinhos no cômputo de tais distâncias, as quais originaram as seguintes estatísticas:

$$J_H((\mathbf{z}_1,...,\mathbf{z}_n),(\mathbf{c}_1,...,\mathbf{c}_\mathsf{k})) = g(n,\mathsf{k})\frac{\sum_{j=1}^{\mathsf{k}} d_H(\mathcal{N}_{\mathbf{c}_j},\mathcal{N}_{\hat{\mu}})}{\sum_{j=1}^{\mathsf{k}}\sum_{i\in C_j} d_H(\mathcal{N}_{\mathbf{z}_i},\mathcal{N}_{\mathbf{c}_j})}, \tag{1.20}$$

$$J_R^\beta((\mathbf{z}_1,...,\mathbf{z}_n),(\mathbf{c}_1,...,\mathbf{c}_\mathsf{k})) = g(n,\mathsf{k})\frac{\sum_{j=1}^{\mathsf{k}} d_R^\beta(\mathcal{N}_{\mathbf{c}_j},\mathcal{N}_{\hat{\mu}})}{\sum_{j=1}^{\mathsf{k}}\sum_{i\in C_j} d_R^\beta(\mathcal{N}_{\mathbf{z}_i},\mathcal{N}_{\mathbf{c}_j})}, \tag{1.21}$$

$$J_B((\mathbf{z}_1,...,\mathbf{z}_n),(\mathbf{c}_1,...,\mathbf{c}_\mathsf{k})) = g(n,\mathsf{k})\frac{\sum_{j=1}^{\mathsf{k}} d_B(\mathcal{N}_{\mathbf{c}_j},\mathcal{N}_{\hat{\mu}})}{\sum_{j=1}^{\mathsf{k}}\sum_{i\in C_j} d_B(\mathcal{N}_{\mathbf{z}_i},\mathcal{N}_{\mathbf{c}_j})}, \tag{1.22}$$

em que $g(n,\mathsf{k}) = \frac{n(n-1)\mathsf{k}}{\mathsf{k}-1}$, $\mathbf{z}_1,...,\mathbf{z}_n$ são os vetores de pré-forma e $\mathbf{c}_1,...,\mathbf{c}_\mathsf{k}$ os centroides dos grupos $C_1,...,C_\mathsf{k}$, para $\mathsf{k} < n$. As distâncias $d_R^\beta(\cdot)$, $d_B(\cdot)$ e $d_H(\cdot)$ estão definidas nas equações (1.11), (1.12) e (1.13), respectivamente.

Os subgrupos formados, $\mathcal{N}_{\mathbf{z}_i} = \left\{\mathbf{z}_i, \mathbf{z}_j : i,j \in \{1,...,n\}, i \neq j, d_{FP}(\mathbf{z}_j,\mathbf{z}_i) \leq d_{\mathbf{z}_i}^{\mathsf{K}}\right\}$ e $\mathcal{N}_{\mathbf{c}_j} = \left\{\mathbf{c}_j, \mathbf{z}_i : i \in \{1,...,n\}, j \in \{1,...,\mathsf{k}\}, d_{FP}(\mathbf{z}_i,\mathbf{c}_j) \leq d_{\mathbf{c}_j}^{\mathsf{K}}\right\}$, são os grupos dos vetores de pré-formas e dos centroides com seus respectivos K vizinhos.

O algoritmo KNN k-médias baseado nas distâncias para Bingham proposto está detalhado no Algoritmo 5.

Considerando que temos dois critérios de parada distintos, podendo serem utilizados com uma estatística de convergência de escolha, e além disso dispomos de três distâncias, foram definidos nesta tese nove variações do Algoritmo 5.

Algoritmo 5: Algoritmo KNN k-médias baseado nas distâncias para Bingham

Result: Retorna $C_1, ..., C_{\mathsf{k}}$

1: Escolha aleatoriamente k centroides $\mathbf{c}_1, ..., \mathbf{c}_{\mathsf{k}}$ dentre os vetores de pré-formas $\mathbf{z}_1, ..., \mathbf{z}_n$;

for $i \leftarrow 1$ **to** *n* **do**

  Obtenha o conjunto $\mathcal{N}_{\mathbf{z}_i}$ utilizando $d_{FP}(\cdot)$;

end

for $j \leftarrow 1$ **to** *k* **do**

  Obtenha o conjunto $\mathcal{N}_{\mathbf{c}_j}$ utilizando $d_{FP}(\cdot)$;

end

4: Considere $d(\cdot)$ igual à $d_R^{\beta}(\cdot)$, $d_H(\cdot)$ ou $d_B(\cdot)$;

if *Critério de parada 2* **then**

  faça $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(-1)} \cap C_j^{(0)}) = 0$ e $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(0)} \cap C_j^{(1)}) = 1$;

end

repeat

  for $i \leftarrow 1$ **to** n **do**

    for $j \leftarrow 1$ **to** *k* **do**

      Calcule a distância $d(\mathcal{N}_{\mathbf{z}_i}, \mathcal{N}_{\mathbf{c}_j})$;

    end

  end

  for $j \leftarrow 1$ **to** *k* **do**

    Defina o grupo $C_j = \{\mathbf{z}_i : \arg\min_r d(\mathcal{N}_{\mathbf{z}_i}, \mathcal{N}_{\mathbf{c}_j}) = r\}$;

  end

  for $j \leftarrow 1$ **to** *k* **do**

    Calcule $\hat{\mu}_j$ a partir de C_j e as considere como novos centroides;

  end

  if *Critério de parada 1* **then**

    Considere $J(\cdot)$ igual à $J_G(\cdot)$, $J_R^{\beta}(\cdot)$, $J_H(\cdot)$ ou $J_B(\cdot)$ e calcule $|J^{(t-1)} - J^{(t)}|$;

  end

  if *Critério de parada 2* **then**

    Calcule $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-2)} \cap C_j^{(t-1)})$ e $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-1)} \cap C_j^{(t)})$;

  end

until $|J^{(t-1)} - J^{(t)}| < \varepsilon$ *ou* $\sum_{j=1}^{k} \#(C_j^{(t-2)} \cap C_j^{(t-1)}) = \sum_{j=1}^{k} \#(C_j^{(t-1)} \cap C_j^{(t)})$;

Para nomenclatura de cada variação do Algoritmo 5 foi considerada: a distribuição na qual se baseiam as distâncias, neste caso, Bingham; a distância utilizada; critério de parada pré-definido; estatística usada no critério de convergência, quando for o caso do critério de parada 1 (Tabela 1).

Tabela 1 – Definições e nomenclatura dos algoritmos propostos.

Distribuição base	Distância	Critério de parada	Estatística/critério de convergência	Nomenclatura
Bingham	Battacharyya	1	J_B	BB1J_B
Bingham	Battacharyya	1	J_G	BB1J_G
Bingham	Battacharyya	2	-	BB2
Bingham	Hellinger	1	J_H	BH1J_H
Bingham	Hellinger	1	J_G	BH1J_G
Bingham	Hellinger	2	-	BH2
Bingham	Rényi de ordem β	1	J_R^β	BR$^\beta$1J_R^β
Bingham	Rényi de ordem β	1	J_G	BR$^\beta$1J_G
Bingham	Rényi de ordem β	2	-	BR$^\beta$2

1.7 METODOLOGIA

Como já mencionado na parte introdutória deste trabalho, o algoritmo proposto por Assis, Souza e Amaral (2021) foi tomado para comparação, considerando cenários de baixa concentração dos dados. Para tanto, foram utilizados como medidas comparativas: índice de Rand (HUBERT; ARABIE, 1985) e o coeficiente Kappa (COHEN, 1960).

O índice de Rand é um critério externo para validação de agrupamento muito conhecido, e o coeficiente Kappa é uma medida de concordância das alocações nos grupos. Ambas são medidas de similaridade e confiabilidade, as quais podem variar no intervalo $(-\infty, 1]$. Sendo que assumir valor 1 indica concordância perfeita entre as partições.

Nossas análises foram inicialmente focadas nos estudos gráficos e simulações de Monte Carlo, tomando cenários de baixa concentração dos dados, e, posteriormente, foram realizadas implementações em três bases de dados reais: dois deles de marcos craniofaciais de crianças da cidade de Ann Arbor, localizada no estado americano de Michigan (BOOKSTEIN, 1997); e a terceira, base *DrivFace*, de marcos anatômicos de faces de motoristas enquanto dirigem (DIAZ-CHITO; HERNÁNDEZ-SABATÉ; LÓPEZ, 2016).

Utilizou-se o ambiente de programação R (R Core Team, 2020), em um notebook com processador Intel(R) Core(TM) i5-8265U, CPU 1.60GHz+1.80 GHz, e memória RAM de 8.00GB.

As variações do algoritmo aqui descrito foram implementadas em ambiente de programação R, e os códigos poderão ser consultados em <https://github.com/adneferreira/tese>. Foi utilizado o pacote shapes para uso das bases de dados dos crânios de gorilas e chimpanzés (O'HIGGINS; DRYDEN, 1993; DRYDEN; MARDIA, 2016), na construção do respectivo gráfico de pré-formas. No cálculo do índice de Rand (HUBERT; ARABIE, 1985) utilizou-se o pacote `mclust`, e, para obtenção do coeficiente Kappa, o pacote `caret` (KUHN, 2008). Para facilitar a construção das tabelas, foi utilizado o pacote `xtable`, no qual, por meio da função de mesmo nome é possível transcrever as tabelas em modelo LaTeX. Para o gráfico da esfera complexa foi utilizada a função `sm.sphere`, disponível no pacote `sm` (BOWMAN; AZZALINI, 1997). Na construção dos gráficos da base *DrivFace* foi utilizado o `ggplot2`.

Na etapa de simulação, foi definido um contexto para avaliar os resultados, aqui denominado "cenário base". Tal cenário consiste em: $n = n_1 + n_2 = 100$ pré-formas, sendo $n_1 = 40$ e $n_2 = 60$ os tamanhos amostrais do grupo 1 e grupo 2, respectivamente; cada amostra gerada foi proveniente de uma distribuição Bingham complexa com função de densidade $f(\mathbf{z}) = c(\mathbf{A})^{-1} \exp\{\mathbf{z}^{\star}\mathbf{A}\mathbf{z}\}$, na qual $\lambda_1 < ... < \lambda_k = 0$ são os autovalores de $\mathbf{A}$, e $k = 3$; considerou-se baixa concentração dos dados, com $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$.

Além disso, para distinguir os grupos entre si, foi utilizada a técnica de rotação de vetores complexos descrita por Amaral, Dryden e Wood (2007), na qual para cada valor dentro do vetor encontra-se uma rotação que gira o vetor original para $\mathbf{v}$ (vetor base para rotação). Isso é feito de tal forma que todos os vetores ortogonais ao plano formado pelo vetor original e $\mathbf{v}$ são deixados inalterados. Nesta tese, foram utilizados como vetores para rotação da amostra $\mathbf{v}_1 = (0, 0, 1)$ no grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ no grupo 2.

A Figura 3 traz um exemplo da disposição espacial na esfera e no plano dos baricentros dos triângulos originados pelas pré-formas simuladas sob as condições do cenário base. A visualização de dados complexos é um grande desafio, entretanto, a ideia de representar os triângulos por seus baricentros ajuda a compreender melhor a diferença entre os grupos simulados. Ao invés de apresentar todos os vértices dos triângulos, gerando um gráfico saturado, podemos considerar apenas uma medida que represente cada triângulo, o baricentro, que é o centro de gravidade do mesmo. Note que é possível distinguir o grupo 1 no centro e o grupo 2 ao derredor.

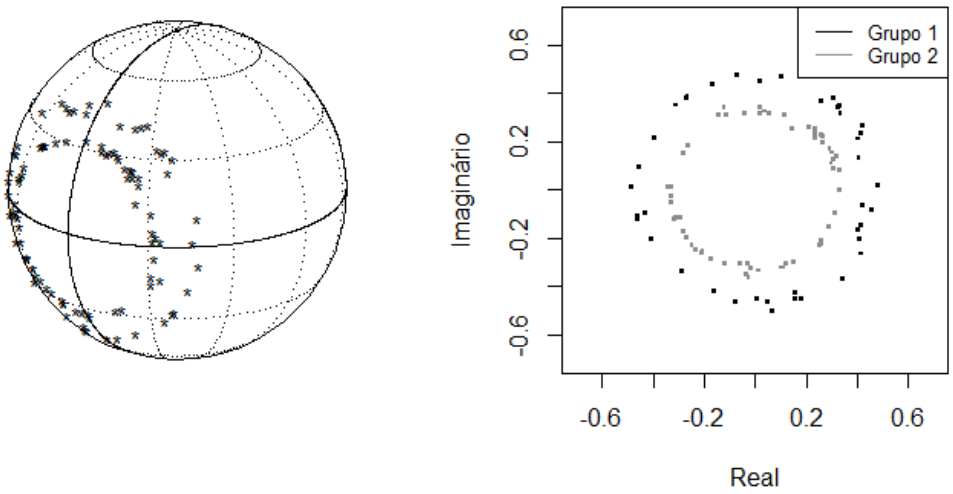


Figura 3 – Baricentros das pré-formas.

Na análise dos dados quantitativos dos índices de Rand ajustados, inicialmente, foram realizados testes de Shapiro-Wilk (ROYSTON, 1995), para verificar presença de normalidade. Posteriormente, foram aplicados testes de Kruskal-Wallis (KRUSKAL; WALLIS, 1952; HOLLANDER; WOLFE; CHICKEN, 2013) e Wilcoxon-Mann-Whitney (MANN; WHITNEY, 1947) para avaliar se as medianas dos índices de Rand ajustados dos algoritmos possuem diferença estatisticamente significativas.

Nesta etapa, utilizamos a correção de Bonferroni no nível de significância α, considerando que tais comparações em pares superestimam a probabilidade de erro tipo I, que corresponde ao nível de significância (BUSSAB; MORETTIN, 2010). A Correção de Bonferroni é uma das mais utilizadas pela sua característica de conservação, no que diz respeito à rigorosidade no controle do nível de significância. No nosso contexto, considerando o teste Wilcoxon-Mann-Whitney, a significância é a probabilidade de rejeitar a igualdade entre as medianas dos índices de Rand gerados pelos algoritmos, comparados dois a dois, quando elas são iguais.

A técnica de Bonferroni define um novo nível de significância, dividindo-se o valor original pelo número total de comparações. Para 10 variações de algoritmos, o número total de comparações duas a duas é dado por:

$$\binom{10}{2} = \frac{10!}{2! \times (10-2)!} = \frac{10 \times 9 \times 8!}{2 \times 1 \times 8!} = \frac{10 \times 9}{2} = 45.$$

Assim, para $\alpha = 1\%$, o nível de significância ajustado para cada teste individual é dado por:

$$\alpha^* = \frac{\alpha}{45} \approx 0.0002.$$

Cada valor p obtido nos testes de Wilcoxon Mann-Whitney é comparado com este novo nível de significância. Se o valor p for menor que 0.0002, a diferença é considerada significativa após a correção de Bonferroni, e podemos afirmar, com 99% de confiança, que as medianas dos índices de Rand ajustados gerados pelos algoritmos comparados são diferentes.

Os resultados foram apresentados e discutidos na subseção a seguir.

1.8 RESULTADOS NUMÉRICOS

Em nossas simulações, inicialmente verificamos o comportamento dos índices de Rand, os quais estão ajustados às variações do algoritmo. O resultado apresentado na Figura 4 sugere que, para o número de marcos anatômicos $k = 3$, os valores iniciais de K já apresentam índices de Rand ajustados elevados, tanto nas variações do algoritmo com distâncias de Battacharyya quanto nas de Hellinger. Já as variações com base na distância de Rényi não demonstraram comportamento semelhante. Tal resultado é similar quando considerado $k = 6$. Motivados por esta análise, decidimos considerar K$= k$, visando minimizar o custo computacional em simulações.

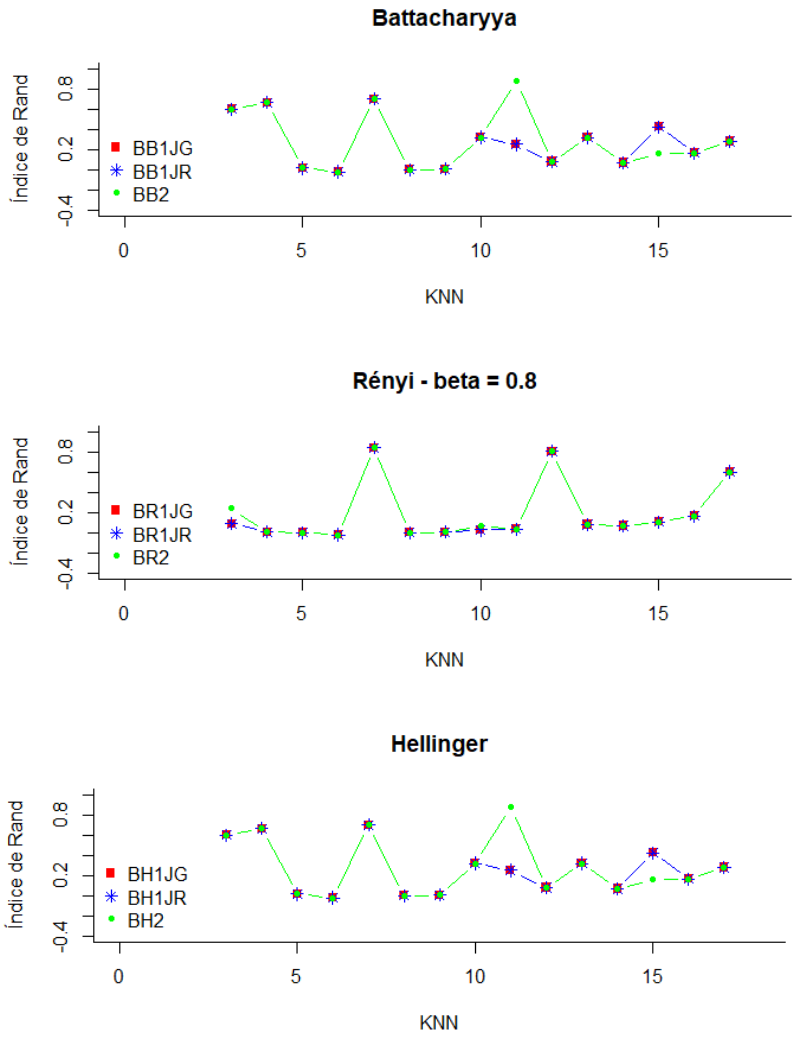


Figura 4 – Gráficos dos índices de Rand ajustados *versus* números de K vizinhos. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899.900)$, vetores de rotação $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2.

Posteriormente, foi simulado um conjunto de dados sob o cenário base, com a finalidade de aplicar uma busca em *grid* pelos melhores valores de acurácia e medidas de similaridade entre os dois grupos. Foi considerado $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 07, 0.8, 0.9\}$, para cada variação do algoritmo de agrupamento baseada na distância de Rényi.

Como apresentado na Tabela 2, para o algoritmo BR$^{\beta}$2, quando $\beta \in \{0.2, 0.8\}$ obtivemos os mesmos valores de coeficiente Kappa (0.8963) e índice de Rand (0.8079). Entretanto, quando se consideram, as outras duas variações do algoritmo (BR$^{\beta}1J_G$ e BR$^{\beta}1J_R^{\beta}$), apenas o $\beta = 0.8$ apresenta os maiores valores das medidas de desempenho nas três variações (coeficiente Kappa igual a 0.1748 e índice de Rand igual a 0.0716). Tal resultado baseou a decisão em seguir para as etapas de simulações de Monte Carlo considerando o valor de $\beta = 0.8$.

Tabela 2 – Índice de Rand e coeficiente Kappa das variações do algoritmo com base na distância de Rényi. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2.

β	Coeficiente Kappa			Índice de Rand ajustado		
	BR$^\beta$2	BR$^\beta$1J_G	BR$^\beta$1J_R^β	BR$^\beta$2	BR$^\beta$1J_G	BR$^\beta$1J_R^β
0.1	-0.1592	0.1546	0.1546	0.0900	0.0598	0.0598
0.2	**0.8963**	-0.1224	-0.1224	**0.8079**	0.0716	0.0716
0.3	0.2584	0.1748	0.1748	0.1179	0.0716	0.0716
0.4	0.4700	0.1748	0.1748	0.2750	0.0716	0.0716
0.5	0.4700	-0.1224	-0.1224	0.2750	0.0716	0.0716
0.6	-0.3604	0.1748	0.1748	0.2750	0.0716	0.0716
0.7	-0.2500	0.1748	0.1748	0.0690	0.0716	0.0716
0.8	**0.8963**	**0.1748**	**0.1748**	**0.8079**	**0.0716**	**0.0716**
0.9	-0.1592	-0.1092	-0.1092	0.0900	0.0598	0.0598

Como já mencionado na seção introdutória desta tese, para fins de comparação, tanto em simulação quanto em dados reais, consideramos o método de agrupamento K-médias para formas planas (KMPS), proposto por Assis, Souza e Amaral (2021). Tal ferramenta é uma adaptação do k-médias original, substituindo a distância euclidiana pela de procrustes completo.

A Tabela 3 apresenta os resultados de média, mediana e desvio padrão (DP) dos índice de Rand ajustados e coeficiente Kappa na simulação de Monte Carlo ($MC = 100$). Note que o maior valor médio de coeficiente Kappa (0.3190) e índice de Rand ajustado (0.1965) foi obtido pelo algoritmo de agrupamento BB1J_G, enquanto que o algoritmo KMPS apresentou a menor média de índice de Rand ajustado (0.0310).

Tabela 3 – Média, mediana e Desvio Padrão (DP) das acurácias, índice de Rand e coeficiente Kappa obtidos nas $MC = 100$ simulações de Monte Carlo, sob as seguintes condições: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2.

Método	Coeficiente Kappa			Índice de Rand ajustado		
	Média	Mediana	DP	Média	Mediana	DP
KMPS	0.2872	0.3011	0.0682	0.0310	0.0291	0.0230
BB1J_G	**0.3190**	**0.3128**	0.2632	**0.1965**	**0.1526**	0.1959
BB1J_B	0.2864	0.2584	0.2528	0.1711	0.1098	0.1861
BB2	0.2876	0.2481	0.2753	0.1793	0.1107	0.2044
BR$^\beta$1J_G	0.2691	0.2284	0.2701	0.1611	0.0867	0.1955
BR$^\beta$1J_R^β	0.2683	0.2144	0.2782	0.1645	0.0789	0.2105
BR$^\beta$2	0.2669	0.1975	0.2729	0.1594	0.0716	0.2017
BH1J_G	0.2746	0.2325	0.2525	0.1633	0.1017	0.1814
BH1J_H	0.2693	0.2308	0.2553	0.1602	0.1017	0.1770
BH2	0.2778	0.2308	0.2685	0.1701	0.1017	0.1957

Para verificar se houve diferença estatística entre as médias das medidas de desempenho obtidas pelos algoritmos, aplicamos um teste de normalidade, verificando a viabilidade de aplicar o teste T de diferença de médias.

A 1% de significância, há rejeição da hipótese de normalidade dos índices de Rand, os quais são advindos de cada algoritmo de agrupamento, com exceção apenas do KMPS, como exibido na Tabela 4. Como consequência dessas observações, escolhemos o teste não paramétrico de diferença de distribuições, Kruskal-Wallis (KRUSKAL; WALLIS, 1952).

Tabela 4 – Índice de Rand ajustado - Estatísticas e valores p dos testes de normalidade Shapiro-Wilk - H_0: admite distribuição normal dos dados. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2.

Algoritmo	Estatística	Valor p
KMPS	0.9675	0.0143
BB1J_G	0.9066	<0.001*
BB1J_B	0.8811	<0.001*
BB2	0.8684	<0.001*
BR$^\beta$1J_G	0.8543	<0.001*
BR$^\beta$1J_R^β	0.8265	<0.001*
BR$^\beta$2	0.8326	<0.001*
BH1J_G	0.8728	<0.001*
BH1J_H	0.8802	<0.001*
BH2	0.8617	<0.001*

*Rejeita normalidade dos dados a 1% de significância.

O teste de postos Kruskal-Wallis apresentou estatística $\chi^2 = 36.168$, com 9 graus de liberdade e valor p < 0.01. É possível afirmar, sob o nível de 1% de significância, que pelo menos um dos algoritmos de agrupamento apresentam distribuição dos índices de Rand ajustados diferente dos demais.

Aplicando o teste não paramétrico de Wilcoxon-Mann-Whitney (MANN; WHITNEY, 1947) nas distribuições de índices de Rand ajustados gerados pelos algoritmos de agrupamento, dois a dois, observa-se que o KMPS apresenta diferença estatística, ao nível ajustado de 0.02%, quando comparado com as variações do algoritmo contendo as distâncias de Battacharyya e Hellinger (Tabela 5).

Tabela 5 – Índices de Rand ajustados - Estatísticas e valores p de testes Wilcoxon-Man-Whitney - H_1 two-sided. Cenário: $(\lambda_1, \lambda_2, \lambda_3) = (898, 899, 900)$, vetores de rotações $\mathbf{v}_1 = (0, 0, 1)$ para o grupo 1 e $\mathbf{v}_2 = (1, 1, 0)$ para o grupo 2.

Algoritmos	KMPS	BB1J_G	BB1J_B	BB2	BR$^\beta$1J_G	BR$^\beta$1J_R^β	BR$^\beta$2	BH1J_G	BH1J_H
BB1J_G	<**0.01***	-	-	-	-	-	-	-	-
	(W=2744.5)	-	-	-	-	-	-	-	-
BB1J_B	<**0.01***	0.4074	-	-	-	-	-	-	-
	(W=2973.5)	(W=5339.5)	-	-	-	-	-	-	-
BB2	<**0.01***	0.379	0.9173	-	-	-	-	-	-
	(W=3242.5)	(W=5360.5)	(W=5043)	-	-	-	-	-	-
BR$^\beta$1J_G	0.0005	0.1222	0.4371	0.5364	-	-	-	-	-
	(W=3581)	(W=5633)	(W=5318.5)	(W=5253.5)	-	-	-	-	-
BR$^\beta$1J_R^β	0.0023	0.1128	0.3931	0.5149	0.9591	-	-	-	-
	(W=3754.5)	(W=5649.5)	(W=5350)	(W=5267)	(W=5021.5)	-	-	-	-
BR$^\beta$2	0.0016	0.1013	0.3633	0.4747	0.9309	0.9727	-	-	-
	(W=3713)	(W=5671)	(W=5372.5)	(W=5293)	(W=5036)	(W=5014.5)	-	-	-
BH1J_G	<**0.01***	0.279	0.7956	0.895	0.5699	0.4992	0.4695	-	-
	(W=3087.5)	(W=5443.5)	(W=5106.5)	(W=5054.5)	(W=4767)	(W=4723)	(W=4703.5)	-	-
BH1J_H	<**0.01***	0.1886	0.6267	0.7094	0.7909	0.7628	0.704	0.8069	-
	(W=3339.5)	(W=5538.5)	(W=5199.5)	(W=5153)	(W=4891)	(W=4876)	(W=4844)	(W=5100.5)	-
BH2	<**0.01***	0.2543	0.7433	0.8431	0.6442	0.6207	0.5774	0.9406	0.895
	(W=3241.5)	(W=5467)	(W=5134.5)	(W=5081.5)	(W=4810.5)	(W=4797)	(W=4771.5)	(W=5031)	(W=4945.5)

*Ao nível ajustado α^* igual a 0.02% de significância, rejeita igualdade das medianas dos grupos.

1.9 APLICAÇÕES EM BASES DE DADOS REAIS

Nesta seção são apresentados resultados das aplicações dos Algoritmos em três bases de dados reais, sendo duas delas correspondentes a mesma amostra em períodos diferentes.

1.9.1 Marcos craniofaciais de crianças

Das duas bases de dados reais aqui utilizados contêm as coordenadas de $k+1 = 8$ marcos craniofaciais de crianças da cidade de Ann Arbor, localizada no estado americano de Michigan,

sendo 26 meninas e 36 meninos (BOOKSTEIN, 1997, pg. 400). A primeira base corresponde a tais dados das crianças com 8 anos de idade, e a segunda foi coletada quando as mesmas tinham 14 anos de idade.

As informações são de coordenadas cartesianas, em polegadas, dos marcos anatômicos cranianos Sella, Nasion, Basion, Gonion, Menton, registro etmoidal, espinha nasal anterior e posterior. Tais coordenadas foram digitalizadas diretamente de exames de raios-x (BOOKSTEIN, 1997). O objetivo do agrupamento foi particionar os indivíduos em dois grupos: meninas e meninos. A Figura 5 exemplifica a representação gráfica dos marcos anatômicos.

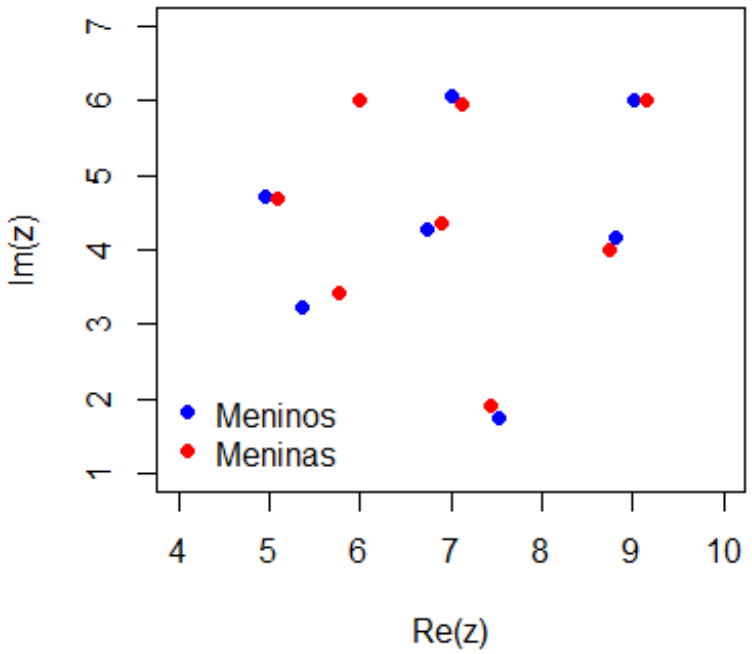


Figura 5 – Representação gráfica dos marcos anatômicos de duas crianças da cidade de Ann Arbor, ambas com 8 anos de idade.

Como mostra a Tabela 6, as variações do algoritmo proposto apresentaram performances muito semelhantes e medidas de desempenho maiores que o do método KMPS. Os algoritmos que obtiveram índice de Rand negativo foram KMPS, BR$^{\beta}1J_R^{\beta}$ e BH1J_G.

Na base de dados das mesmas crianças, com 14 anos de idade, o BB1J_B (Tabela 6) apresentou maiores valores de índice de Rand (0.1878) e coeficiente Kappa (0.0348).

Tabela 6 – Base de dados das crianças com 8 e 14 anos de idade. Índice de Rand ajustado e coeficiente Kappa dos algoritmos de agrupamento.

Algoritmo	8 anos de idade		14 anos de idade	
	Coeficiente Kappa	Índice de Rand	Coeficiente Kappa	Índice de Rand
KMPS	0.0169	-0.0153	0.0132	-0.0131
BB1J_G	0.0903	0.0078	0.0048	-0.0055
BB1J_B	0.0903	0.0078	0.1878	0.0348
BB2	0.0903	0.0078	0.0390	-0.0028
BR$^\beta$1J_G	0.0903	0.0078	0.0272	-0.0152
BR$^\beta$1J_R^β	0.0419	-0.0080	0.0103	-0.0163
BR$^\beta$2	0.0903	0.0078	-0.0105	-0.0163
BH1J_G	0.0419	-0.0080	0.1199	0.0007
BH1J_H	0.0903	0.0078	-0.0204	-0.0148
BH2	0.0903	0.0078	0.0390	-0.0028

1.9.2 Base de dados *DrivFace*

Outra base utilizada nas aplicações foi o *DrivFace* (DIAZ-CHITO; HERNÁNDEZ-SABATÉ; LÓPEZ, 2016), disponível para donwload no repositório da UCI *Machine Learning* por meio do sítio eletrônico: <https://archive.ics.uci.edu/ml/datasets/DrivFace>.

A base de dados *DrivFace* contém marcos anatômicos (nariz, canto direito e esquerdo da boca, posição dos olhos e do rosto) de quatro sujeitos, ao longo de dias diferentes, durante a condução de um veículo em cenários reais. A câmera fixa na parte frontal e no centro do carro captou três posições para onde se dirigiam as cabeças dos motoristas: frente, direita e esquerda. Na Figura 6 segue um exemplo gráfico dos landmarks registrados.

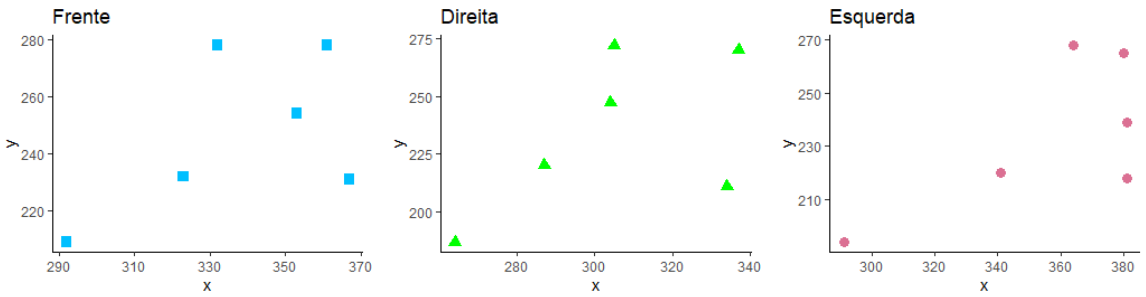


Figura 6 – Exemplo gráfico dos *landmarks* dos rostos de motoristas captados com a posição da cabeça para frente, direita ou esquerda. Base de dados *DrivFace*, 2016.

O objetivo do agrupamento foi discriminar entre tais posições, comparando-as duas a duas, e para isso foram definidos 4 cenários: No cenário 1, tem-se motoristas olhando para direita

(grupo 1) e motoristas olhando para frente (grupo 2); o cenário 2 possui indivíduos olhando para esquerda (grupo 1) e indivíduos olhando para frente (grupo 2); no cenário 3 tem-se motoristas olhando para direita (grupo 1) e motoristas olhando para esquerda (grupo 2); e o cenário 4, sujeitos olhando para frente (grupo 1) e sujeitos olhando para direita ou esquerda (grupo 2), foi escolhido propositalmente pela relevância da alegação mencionada por Diaz-Chito, Hernández-Sabaté e López (2016), de que os maiores números de acidentes ocorrem quando os motoristas não estão atentos ao trânsito, ou seja, olhando para a direita ou para a esquerda.

A Tabelas 7 apresenta os resultados das aplicações nos cenários 1, 2, 3 e 4. Note que, no cenário 1 o algoritmo com melhor desempenho foi o BR$^{\beta}$2, com coeficiente Kappa de 0.2971 e índice de Rand igual a 0.2491. No cenário 2 todos os algoritmos apresentaram índices de Rand ajustados negativos, e no cenário 3 destaca-se o algoritmo KMPS, com coeficiente Kappa e índice de Rand iguais a 1.00. No cenário 4, o melhor desempenho foi do algoritmo BB1J_B, com coeficiente Kappa de 0.3255 e índice de Rand igual a 0.2107.

Tabela 7 – Coeficiente Kappa e índice de Rand ajustado dos algoritmos aplicados na base *DrivFace* - Cenários 1, 2, 3 e 4.

Algoritmo	Cenário 1		Cenário 2		Cenário 3		Cenário 4	
	C. Kappa	I. Rand	C. Kappa	I. Rand	C. Kappa	I. Rand	C. Kappa	I. Rand
KMPS	0.2726	0.2081	-0.0998	-0.0627	**1.0000**	**1.0000**	0.1452	0.0942
BB1J_G	0.1208	0.0599	-0.0702	-0.0159	0.4718	0.2387	0.0721	0.0097
BB1J_B	0.1208	0.0599	-0.0180	-0.0072	0.4718	0.2387	**0.3255**	**0.2107**
BB2	0.1208	0.0599	-0.0180	-0.0072	0.5548	0.3099	-0.1418	-0.0638
BR$^{\beta}$1J_G	0.1887	0.1454	-0.0175	-0.0071	0.4718	0.2387	0.1616	0.0648
BR$^{\beta}$1J_R^{β}	0.0855	-0.0005	0.0153	-0.0006	0.4718	0.2387	-0.0986	-0.0623
BR$^{\beta}$2	**0.2971**	**0.2491**	0.0153	-0.0006	0.4718	0.2387	0.1616	0.0648
BH1J_G	0.1208	0.0599	-0.0702	-0.0159	0.4718	0.2387	0.0721	0.0097
BH1J_H	0.1208	0.0599	0.0170	-0.0001	0.4718	0.2387	-0.1027	-0.0531
BH2	0.1208	0.0599	-0.0702	-0.0159	0.4718	0.2387	-0.1027	-0.0531

As Figuras 7, 8 e 9 mostram uma representação gráfica dos marcos anatômicos dos grupos nos cenários 1, 2 e 3, definidos na base *DrivFace*, respectivamente. Tanto no cenário 2 (Figura 8) quanto no 3 (Figura 9), nos quais o algoritmo KMPS mostrou melhor desempenho no agrupamento, o grupo 1 e 2 contém uma diferença visualmente perceptível na disposição dos marcos no quadrante superior direito dos gráficos. Já no cenário 1 (Figura 9), os marcos dos grupos aparentam ser semelhantes, e, nestas condições, o KMPS não apresentou o melhor resultado.

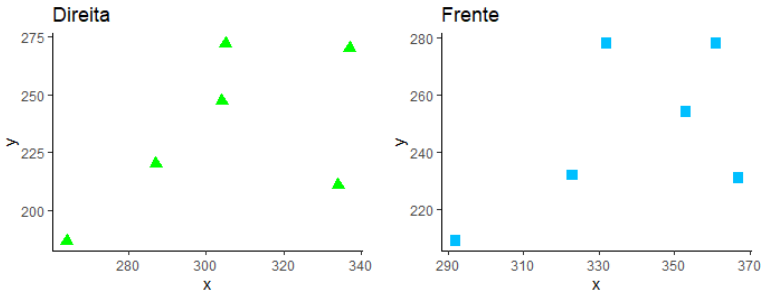


Figura 7 – Exemplo gráfico dos *landmarks* no cenário 1: motoristas olhando para direita (grupo 1) e motoristas olhando para frente (grupo 2). Base de dados *DrivFace*, 2016.

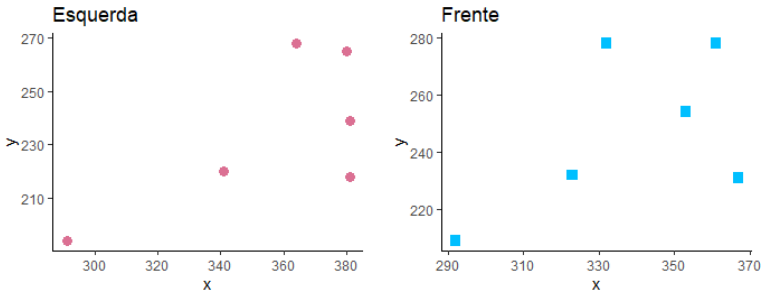


Figura 8 – Exemplo gráfico dos *landmarks* no cenário 2: motoristas olhando para esquerda (grupo 1) e motoristas olhando para frente (grupo 2). Base de dados *DrivFace*, 2016.

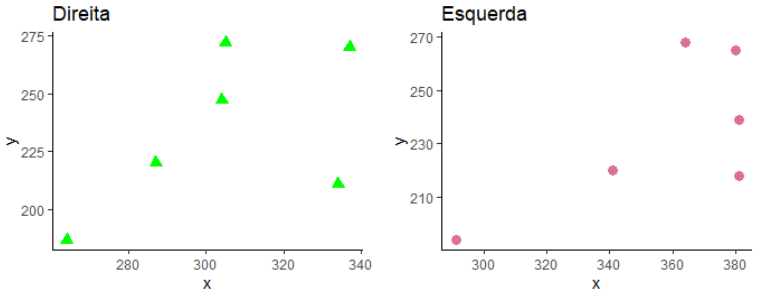


Figura 9 – Exemplo gráfico dos *landmarks* no cenário 3: motoristas olhando para direita (grupo 1) e motoristas olhando para esquerda (grupo 2). Base de dados *DrivFace*, 2016.

1.10 CONSIDERAÇÕES FINAIS

A análise de agrupamentos é uma área importante no campo do reconhecimento de padrões dentro das bases de dados (KAUFMAN; ROUSSEEUW, 2009). Algoritmos têm sido utilizados na área de análise de formas com a finalidade de agrupar objetos com características semelhantes, de maneira não supervisionada.

Uma das abordagens mais comuns é a que utiliza os pontos de referências nas imagens, os *landmarks* (DRYDEN; MARDIA, 2016). No contexto de formas planas, existe toda uma teoria bem fundamentada, e que abre caminho para exploração de novas técnicas.

Neste primeiro capítulo da tese foi apresentada uma proposta de algoritmo de agrupamento com nove possíveis variações (BB1J_G, BB1J_B, BB2, BR$^\beta$1J_G, BR$^\beta$1J_R^β, BR$^\beta$2, BH1J_G, BH1J_H, BH2), bem como seus resultados em simulações e bases de dados reais. O objetivo principal foi encontrar algoritmos que retornassem melhores resultados em cenários de baixa concentração dos dados, considerando que foram exatamente nesses cenários onde o algoritmo pré-existente na literatura, aqui denominado KMPS (ASSIS; SOUZA; AMARAL, 2021), não desempenhou bem o seu papel.

Na análise do numero de K vizinhos, observou-se maior acurácia quando K$= k$, e, dentre os valores de β que foram considerados, $\beta = 0.8$ obteve melhores medidas de desempenho. Dentre as nove variações do algoritmo de agrupamento proposto, o BB1J_G mostrou melhores resultados em simulações.

Quando nos cenários de dados reais: o BB1J_B apresentou melhor acurácia na base de dados craniofaciais das crianças com 14 anos de idade; na base de marcos anatômicos da face de motoristas, os algoritmos com melhores desempenhos foram BR$^\beta$2, BB1j_B e o KMPS.

Dadas as dificuldades encontradas por Assis, Souza e Amaral (2021) no contexto de agrupamento de dados com baixa concentração, cremos ter contribuído no acréscimo de outras opções de algoritmos para agrupamento de formas. Considerar as distâncias entre distribuições Bingham e aliá-las ao método KNN promoveu bons resultados.

2 DISTÂNCIAS PARA DADOS DIRECIONAIS EM ESPAÇO MULTIDIMENSIONAL BASEADAS NA DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL

2.1 INTRODUÇÃO

A análise de dados direcionais é um ramo de estudo importante na Estatística, muitos estudos nessa temática têm sido desenvolvidos. Dados direcionais são aqueles que possuem informações sobre direções distinguíveis, s quais originam orientações no espaço em que estão inseridos. Além disso, sua distribuição de probabilidade está intimamente relacionada com a normal multivariada (DOWNS, 1972).

Em seu artigo de revisão, Mardia (1988) faz uma revisão na literatura das principais técnicas e descobertas que foram desenvolvidas até o ano de 1988, contando inclusive com a menção dos dois modelos básicos para o caso multidimensional, conhecidos como distribuição de Von Mises-Fisher e distribuição de Bingham (MARDIA, 1975; KHATRI; MARDIA, 1977; JUPP; MARDIA, 1979).

Alguns autores apresentaram diversas aplicações de análises estatísticas em dados direcionais, dentre eles: Downs (1972) trabalha com dados de vetorcardiograma clínico; Prentice (1986) também utiliza dados de vetorcardiograma para demonstrar a aplicabilidade da alternativa para estimativa de máxima verossimilhança que propôs; Kunze e Schaeben (2005) implementam técnicas para determinar padrões de orientações cristalográficas; Souza, Nascimento e Amaral (2021) fizeram uma aplicação em dados paleomagnéticos e ilustraram como seu modelo pode ser empregado para analisar a concentração de dados esféricos. Outro resultado relevante é descrito por Diethe (2015). Em seu artigo, o autor apresenta uma derivação da divergência de Kullback-Leibler para a distribuição de Von Mises-Fisher em p dimensões. Partindo dessa ideia, e somando-se àquelas apresentadas por Félix (2019) em sua tese sobre testes de hipóteses e divergências, nesta pesquisa fomos capazes de contribuir para literatura sobre dados direcionais, adicionando novas distâncias e algoritmos de agrupamento.

O objetivo principal foi propor distâncias de Bhattacharyya, Hellinger e Rényi de ordem β, no contexto multidimensional, tendo-se com base a distribuição Von Mises-Fisher matricial. Sendo assim, capazes de apresentar novos algoritmos de agrupamento para dados direcionais.

O Capítulo 2 desta tese está organizado da seguinte forma: a primeira seção (2.2) explana sobre a distribuição Von Mises-Fisher matricial, sua função hipergeométrica e Função Geradora de Momentos; a seção subsequente (2.3) apresenta as três distâncias aqui propostas para dados

direcionais, com base na distribuição Von Mises-Fisher matricial; posteriormente, a Seção 2.4 detalha o algoritmo de agrupamento com base nas distâncias propostas; a Seção 2.5 apresenta a metodologia utilizada e a Seção 2.6, os resultados numéricos; aplicações em bases de dados reais são analisadas na Seção 2.7; por fim, seguem as considerações finais deste Capítulo na Seção 2.8. As demonstrações e cálculo das distâncias encontram-se no Anexo A, após as referências bibliográficas.

2.2 DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL

Em p dimensões, assumimos a orientação de um objeto por n direções $\mathbf{x}_1, ..., \mathbf{x}_n$ em $\mathbb{R}^p$, as quais compõem as linhas da matriz $\mathbf{X}$ com dimensão $n \times p$ e $n \leq p$. Além dessa matriz satisfazer

$$\mathbf{X}\mathbf{X}^\top = \mathbf{I}_n, \tag{2.1}$$

em que $\mathbf{I}_n$ é a matriz identidade com dimensão $n \times n$ e $\mathbf{X}^\top$ é a matriz transposta de $\mathbf{X}$. O espaço cujo os elementos têm as características de $\mathbf{X}$ é chamado de Stiefel múltiplo, e aqui será denotado por $O(n, p)$. Para $n = p$, por exemplo, o espaço de Stiefel múltiplo se torna o espaço ortogonal (KHATRI; MARDIA, 1977).

Turaga, Chellappa e Srivastava (2013) apresentam algumas relações entre espaços em múltiplas dimensões. A hiperesfera $(p-1)$-dimensional, por exemplo, denotada por S^{p-1}, pode ser transformada em um subconjunto de $\mathbb{R}^p$. As matrizes ortogonais com determinantes iguais a 1 formam um grupo ortogonal especial, sendo o espaço múltiplo de Stiefel o conjunto de todas as bases ortogonais n-dimensionais em $\mathbb{R}^p$.

Nesse contexto, a distribuição Von Mises-Fisher matricial é escrita da seguinte maneira

$$a(\mathbf{F}) \exp\{\text{tr}\mathbf{F}\mathbf{X}^\top\}[d\mathbf{X}], \quad \mathbf{X} \in O(n, p), \tag{2.2}$$

em que $[d\mathbf{X}]$, $\mathbf{X} \in O(n, p)$ corresponde à unidade de massa da medida Haar[1] em $O(n, p)$, $\mathbf{F}$ é uma matriz de parâmetros com dimensão $n \times p$, e $a(\mathbf{F})$ é uma função normalizadora definida pela relação

$$\frac{1}{a(\mathbf{F})} = {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}\mathbf{F}\mathbf{F}^\top\right), \tag{2.3}$$

onde ${}_0F_1(\cdot)$ é a função hipergeométrica com argumento matricial (DOWNS, 1972; KOEV; EDELMAN, 2006).

[1] A medida é simplesmente a área de superfície normalizada da superfície dimensional $np - \frac{1}{2}n(n+1)$ no espaço np definido por (2.1) e, portanto, pode ser considerado a distribuição de probabilidade uniforme em $O(n, p)$.

Uma variável $\mathbf{X}$ que segue distribuição Von Mises-Fisher matricial com matriz de parâmetros $\mathbf{F}$ de dimensão $n \times p$ será denotada por $\mathbf{X} \sim M(p, n, \mathbf{F})$. Na equação (**??**), o caso em que n é igual a 1 se reduz à distribuição Von Mises-Fisher p variada (MARDIA, 1975).

Considere a estimativa de máxima verossimilhança no modelo com densidade Von Mises-Fisher matricial, para amostras independentes e identicamente distribuídas $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_\mathsf{N}$. De acordo com Mardia, Jupp e Mardia (2000), sendo este um modelo exponencial, o estimador de máxima verossimilhança do parâmetro canônico $\mathbf{F}$ é dado implicitamente por:

$$\overline{\mathbf{X}} = \sum_{i=1}^{\mathsf{N}} \frac{\mathbf{X}_i}{\mathsf{N}}, \tag{2.4}$$

em que $\mathbf{X}_i/\mathsf{N}$ denota a divisão de cada elemento de $\mathbf{X}_i$ pelo numero natural N, o qual representa o tamanho da amostra.

Seja $\mathbf{F}$ uma matriz de posto completo r, para $0 \leq r \leq n \leq p$. Existem matrizes $\mathbf{\Delta}$, $\mathbf{\Gamma}$ e $\mathbf{D}_\phi = \mathsf{diag}(\phi_1, \phi_2, ..., \phi_r)$, para $\phi_i > 0 \ \forall i$, tais que

$$\mathbf{F} = \mathbf{\Delta}\mathbf{D}_\phi\mathbf{\Gamma}, \quad \mathbf{\Gamma}\mathbf{\Gamma}^\top = \mathbf{I}_r, \quad \mathbf{\Delta}^\top\mathbf{\Delta} = \mathbf{I}_r, \tag{2.5}$$

em que $\mathbf{I}_r$ é a matriz identidade com dimensão $r \times r$.

Para garantir que $\mathbf{F}$ seja não singular, assume-se que $\phi_1 > \phi_2 > ... > \phi_r > 0$. Também assumimos que os elementos da primeira coluna de $\mathbf{\Gamma}$ são positivos. Caso algum elemento da primeira coluna de $\mathbf{\Gamma}$ ou da primeira linha de $\mathbf{\Delta}$ sejam iguais a zero, então pegamos o primeiro elemento não nulo de cada linha de $\mathbf{\Gamma}$ (DOWNS, 1972).

Uma representação para $\mathbf{F}$ é apresentada por Mardia, Jupp e Mardia (2000), Khatri e Mardia (1977) e Downs (1972). Denotando-se

$$\mathbf{F} = \mathbf{K}\mathbf{M}, \quad \mathbf{M} = \mathbf{\Delta}\mathbf{\Gamma}, \quad \mathbf{K} = \mathbf{\Delta}\mathbf{D}_\phi\mathbf{\Delta}^\top, \tag{2.6}$$

em que $\mathbf{M}$ é a componente de orientação (ou polar) e $\mathbf{K}$ é o componente de concentração (ou elíptico) de $\mathbf{F}$. Além disso, os autores também provam que

$${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}\mathbf{F}\mathbf{F}^\top\right) = {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}\mathbf{D}_\phi^2\right), \tag{2.7}$$

ou seja, a constante de normalização $a(\mathbf{F})$ depende apenas de $\mathbf{D}_\phi$.

Para $n \geq 2$, Downs (1972) mostra que existe uma relação entre a normal multivariada e as distribuições da família a qual a Von Mises-Fisher matricial pertence. Nesta direção, suponha que $\mathbf{X}$ é uma matriz $n \times p$ que possui colunas independentes entre si, cujas observações seguem distribuição normal n-variada com matriz de covariância comum $\mathbf{K}^{-1}$ e vetores de médias $\boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, ..., \mu_{nj})^\top$, para $j = 1, ..., p$. No caso em que $\boldsymbol{\mu}_j$ correspondem às colunas de $\mathbf{M}$ e $\mathbf{M}\mathbf{M}^\top = \mathbf{I}_n$, então $\mathbf{X} \sim M(p, n, \mathbf{K}\mathbf{M})$, ou seja, $\mathbf{X} \sim M(p, n, \mathbf{F})$.

2.2.1 Função Hipergeometrica e Função Geradora de Momentos

Khatri e Mardia (1977) apresentam aproximações para a função hipergeométrica, sendo a primeira delas dada por

$$ {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}\mathbf{F}\mathbf{F}^\top\right) \approx \frac{\left\{2^{-\frac{1}{4}n(n+5)+\frac{1}{2}np}\pi^{-\frac{1}{2}n}\prod_{j=1}^{n}\Gamma\left(\frac{p-j+1}{2}\right)\right\}\exp\{\mathrm{tr}\mathbf{D}_\phi\}}{\left[\left\{\prod_{j=2}^{n}\prod_{i=1}^{j-1}(\phi_i+\phi_j)^{\frac{1}{2}}\right\}\prod_{i=1}^{n}\phi_i^{\frac{1}{2}(p-n)}\right]}, \tag{2.8}$$

para $\phi_1, ..., \phi_n$ grande, em que Γ é a função gama. Já para valores pequenos de $\phi_1, ..., \phi_n$, tem-se

$$ {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}\mathbf{F}\mathbf{F}^\top\right) \approx 1 + (2p)^{-1}\sum_{i=1}^{n}\phi_i^2. \tag{2.9}$$

A Função Geradora de Momentos (F.G.M.) de $\mathbf{X} \sim M(p, n, \mathbf{F})$, definida por Khatri e Mardia (1977), é apresentada a seguir.

Seja $\mathbf{X} \sim M(p, n, \mathbf{F})$, sua F.G.M é dada por:

$$\begin{aligned} \mathbb{M}(\mathbf{T}) &= \mathbb{E}[\exp\{\mathrm{tr}\mathbf{T}\mathbf{X}^\top\}] && (2.10)\\ &= \frac{{}_0F_1\left(\frac{p}{2}; \frac{1}{4}(\mathbf{F}+\mathbf{T})(\mathbf{F}+\mathbf{T})^\top\right)}{{}_0F_1\left(\frac{p}{2}; \frac{1}{4}\mathbf{F}\mathbf{F}^\top\right)}, && (2.11)\end{aligned}$$

em que $\mathbf{T}$ é uma matriz $n \times p$.

Tais conceitos serão utilizados na próxima seção para construção de novas distâncias com base em distribuição von Mises-Fisher matricial.

A fim de denotar as fórmulas de maneira mais sucinta, a partir da próxima seção, nós consideraremos

$$ {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}\mathbf{F}\mathbf{F}^\top\right) = {}_0F_1\left(\mathbf{F}\right). $$

2.3 DISTÂNCIAS BASEADAS NA DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL

Como mencionado no Capítulo 1, para uma função ser considerada distância, precisa satisfazer as propriedades de positividade, definitude e simetria. Assim como Félix (2019) utiliza as distâncias de Bhattacharyya, Hellinger e Rényi de ordem β, $\beta \in (0, 1)$, para derivar distâncias baseadas na distribuição Bingham complexa, também derivamos distâncias, mas para o espaço multidimensional, com base na distribuição matricial von Mises-Fisher.

A seguir apresentamos expressões analíticas exatas para as distâncias de Rényi, Bhattacharyya e Hellinger entre medidas de probabilidade com distribuição von Mises-Fisher matricial.

Definição 2.3.1. *Sejam $f_{\boldsymbol{X}_1}(\mathbf{X};\boldsymbol{\theta}_1)$ e $f_{\boldsymbol{X}_2}(\mathbf{X};\boldsymbol{\theta}_2)$ funções densidade de probabilidade de $\mathbf{X}_1$ e $\mathbf{X}_2$ e $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, seus respectivos parâmetros . Considere ainda a divergência de Rényi de ordem β:*

$$D_R^{\beta}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2) = \frac{1}{\beta-1}\log \mathbb{E}_{\boldsymbol{\theta}_1}\left[\left(\frac{f_{\boldsymbol{X}_1}(\mathbf{X};\boldsymbol{\theta}_1)}{f_{\boldsymbol{X}_2}(\mathbf{X};\boldsymbol{\theta}_2)}\right)^{\beta-1}\right]. \tag{2.12}$$

As distâncias de Rényi de ordem β, para $\beta \in (0,1)$, Hellinger e Bhattacharyya são respectivamente dadas por:

$$d_R^{\beta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \frac{1}{\beta-1}\log\left\{\frac{\exp[(\beta-1)D_R^{\beta}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2)] + \exp[(\beta-1)D_R^{\beta}(\boldsymbol{\theta}_2||\boldsymbol{\theta}_1)]}{2}\right\}. \tag{2.13}$$

$$d_B(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = -\log\left\{\mathbb{E}_{\boldsymbol{\theta}_1}\left[\sqrt{\frac{f_{\boldsymbol{X}_2}(\mathbf{X};\boldsymbol{\theta}_2)}{f_{\boldsymbol{X}_1}(\mathbf{X};\boldsymbol{\theta}_1)}}\right]\right\}, \tag{2.14}$$

$$d_H(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = 1 - \mathbb{E}_{\boldsymbol{\theta}_1}\left[\sqrt{\frac{f_{\boldsymbol{X}_2}(\mathbf{X};\boldsymbol{\theta}_2)}{f_{\boldsymbol{X}_1}(\mathbf{X};\boldsymbol{\theta}_1)}}\right]. \tag{2.15}$$

Note que há uma relação entre $d_B(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ e $d_H(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, na qual $d_B(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = -\log\{d_H(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\}$.

Utilizando essas distâncias, tomamos as funções densidade de probabilidade da von Mises-Fisher matricial e derivamos as distâncias para o contexto multidimensional.

Teorema: Sejam $\mathbf{X_1} \sim M(p,n,\mathbf{F_1})$, $\mathbf{X_2} \sim M(p,n,\mathbf{F_2})$ e suas funções densidades de probabilidades:

$$f_{\mathbf{X_j}}(\mathbf{X};\mathbf{F_j}) = \exp\left\{\text{tr}(\mathbf{F_j}\mathbf{X}^{\top}) - \log\left[{}_0F_1\left(\mathbf{F_j}\right)\right]\right\}, \tag{2.16}$$

para $j \in \{1,2\}$.

Substituindo $f_{\boldsymbol{X}_1}(\mathbf{X};\boldsymbol{\theta}_1)$ e $f_{\boldsymbol{X}_2}(\mathbf{X};\boldsymbol{\theta}_2)$ por $f_{\mathbf{X_1}}(\mathbf{X};\mathbf{F_1})$ e $f_{\mathbf{X_2}}(\mathbf{X};\mathbf{F_2})$, em $D_R^{\beta}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2)$, $D_R^{\beta}(\boldsymbol{\theta}_2||\boldsymbol{\theta}_1)$, $d_B(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ e $d_H(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, temos, respectivamente:

$$D_R^{\beta}(\mathbf{F_1}||\mathbf{F_2}) = \frac{1}{\beta-1}\log\left\{\frac{{}_0F_1(\mathbf{F_2})^{\beta-1}\,{}_0F_1(\beta\mathbf{F_1}-(\beta-1)\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})^{\beta}}\right\},$$

$$D_R^{\beta}(\mathbf{F_2}||\mathbf{F_1}) = \frac{1}{\beta-1}\log\left\{\frac{{}_0F_1(\mathbf{F_1})^{\beta-1}\,{}_0F_1(\beta\mathbf{F_2}-(\beta-1)\mathbf{F_1})}{{}_0F_1(\mathbf{F_2})^{\beta}}\right\},$$

$$d_H(\mathbf{F_1},\mathbf{F_2}) = 1 - \frac{{}_0F_1(\frac{1}{2}(\mathbf{F_1}+\mathbf{F_2}))}{[{}_0F_1(\mathbf{F_2})\,{}_0F_1(\mathbf{F_1})]^{\frac{1}{2}}} \tag{2.17}$$

e

$$d_B(\mathbf{F_1},\mathbf{F_2}) = \log\left\{\frac{[{}_0F_1(\mathbf{F_2})\,{}_0F_1(\mathbf{F_1})]^{\frac{1}{2}}}{{}_0F_1(\frac{1}{2}(\mathbf{F_1}+\mathbf{F_2}))}\right\}. \tag{2.18}$$

Agora, substituindo $D_R^\beta(\mathbf{F_1}||\mathbf{F_2})$ e $D_R^\beta(\mathbf{F_2}||\mathbf{F_1})$ em $d_R^\beta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, temos:

$$d_R^\beta(\mathbf{F_1}, \mathbf{F_2}) = \frac{1}{\beta - 1} \log \left\{ \frac{1}{2} \left[G(1,2) + G(2,1) \right] \right\}, \tag{2.19}$$

em que $G(1,2)$ e $G(2,1)$ correspondem respectivamente a

$$G(1,2) = \frac{{}_0F_1(\mathbf{F_2})^{\beta-1} {}_0F_1(\beta \mathbf{F_1} - (\beta - 1)\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})^\beta},$$

$$G(2,1) = \frac{{}_0F_1(\mathbf{F_1})^{\beta-1} {}_0F_1(\beta \mathbf{F_2} - (\beta - 1)\mathbf{F_1})}{{}_0F_1(\mathbf{F_2})^\beta}.$$

As demonstrações e provas das propriedades de definição de distância podem ser encontradas no Apêndice A desta tese.

2.4 ALGORITMO DE AGRUPAMENTO: NOSSA PROPOSTA

Na área de análise de agrupamento, os algoritmos têm como objetivo reconhecer padrões e subdividir os dados não categorizados em grupos (IZBICKI; SANTOS, 2020; FRIEDMAN et al., 2001). Um famoso algoritmo é o k-médias, definido por Lloyd (IZBICKI; SANTOS, 2020; FRIEDMAN et al., 2001).

Neste Capítulo adaptamos o algoritmo k-médias utilizando as distâncias de Bhattacharyya, Hellinger e Rényi para dados com distribuição von Mises-Fisher matricial, descritas na seção anterior. Também modificamos a estatística de similaridade proposta por Goodall (1991), por meio das distâncias d_B, d_H e d_R^β. Assim, considerando $C_1, C_2, ..., C_\mathsf{k}$ partições dos N elementos amostrais, para $\mathsf{k} \in \mathbb{N}^*$, e $\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_\mathsf{k}$ suas respectivas matrizes centroides, as estatísticas de similaridade adaptadas são dadas por:

$$J_B((\mathbf{X}_1, ..., \mathbf{X}_n), (\mathbf{C}_1, ..., \mathbf{C}_\mathsf{k})) = g(\mathsf{N}, \mathsf{k}) \frac{\sum_{j=1}^\mathsf{k} d_B(\mathbf{C}_j, \overline{\mathbf{X}})}{\sum_{j=1}^\mathsf{k} \sum_{i \in C_j} d_B(\mathbf{X}_i, \mathbf{C}_j)}, \tag{2.20}$$

$$J_H((\mathbf{X}_1, ..., \mathbf{X}_n), (\mathbf{C}_1, ..., \mathbf{C}_\mathsf{k})) = g(\mathsf{N}, \mathsf{k}) \frac{sum_{j=1}^\mathsf{k} d_H(\mathbf{C}_j, \overline{\mathbf{X}})}{\sum_{j=1}^\mathsf{k} \sum_{i \in C_j} d_H(\mathbf{X}_i, \mathbf{C}_j)}, \tag{2.21}$$

$$J_R^\beta((\mathbf{X}_1, ..., \mathbf{X}_n), (\mathbf{C}_1, ..., \mathbf{C}_\mathsf{k})) = g(\mathsf{N}, \mathsf{k}) \frac{\sum_{j=1}^\mathsf{k} d_R^\beta(\mathbf{C}_j, \overline{\mathbf{X}})}{\sum_{j=1}^\mathsf{k} \sum_{i \in C_j} d_R^\beta(\mathbf{X}_i, \mathbf{C}_j)}, \tag{2.22}$$

em que $g(\mathsf{N}, \mathsf{k}) = \frac{\mathsf{N}(\mathsf{N}-1)\mathsf{k}}{\mathsf{k}-1}$, $\overline{\mathbf{X}}$ é a média das matrizes de direções em $\mathbb{R}^p$, dada por $\sum_{i=1}^N \mathbf{X_i}/\mathsf{N}$.

Quanto aos critérios de convergências adotados, foram utilizados dois tipos. Seja t o número de iterações decorridas e T o número de máximo de iterações pré-estabelecidas:

- **Critério de parada 1** - quando se alcança o número máximo de iterações pré-estabelecidas $(t = T)$ ou quando o valor absoluto da diferença entre o critério de convergência da iteração anterior $(J^{(t-1)})$ e o atual $(J^{(t)})$ é menor que um ε muito pequeno, i.e. $|J^{(t-1)} - J^{(t)}| < \varepsilon$;

- **Critério de parada 2** - quando se alcança o número máximo de iterações pré-estabelecidas $(t = T)$ ou quando atingi-se uma iteração em que não há mudança do número de elementos nos grupos da iteração anterior e da atual, i.e. $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-2)} \cap C_j^{(t-1)}) = \sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-1)} \cap C_j^{(t)})$.

Perceba que temos dois critérios de parada distintos, podendo ser utilizado com uma estatística de convergência de escolha. Além disso, dispomos de três distâncias. Assim, foram definidas seis variações do algoritmo proposto, descrito sumariamente no Algoritmo 6.

Algoritmo 6: Algoritmo baseado nas distâncias para Von Mises-Fisher matricial

Result: Grupos $C_1, ..., C_{\mathsf{k}}$.

1: Escolha aleatoriamente k centroides $\mathbf{C}_1, ..., \mathbf{C}_{\mathsf{k}}$ dentre as matrizes $\mathbf{X}_1, ..., \mathbf{X}_{\mathsf{N}}$;

if *Critério de parada 2* **then**

faça $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(-1)} \cap C_j^{(0)}) = 0$ e $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(0)} \cap C_j^{(1)}) = 1$;

end

repeat

for $i \leftarrow 1$ **to** *N* **do**

for $j \leftarrow 1$ **to** *k* **do**

Calcule a distância $d(\mathbf{X}_i, \mathbf{C}_j)$;

end

end

for $j \leftarrow 1$ **to** *k* **do**

Defina o grupo $C_j = \{\mathbf{X}_i : \arg\min_r d(\mathbf{X}_i, \mathbf{C}_j) = r\}$;

end

for $j \leftarrow 1$ **to** *k* **do**

Calcule $\overline{\mathbf{X}}_j$ a partir de C_j e as considere como novos centroides;

end

if *Critério de parada 1* **then**

Faça $J(\cdot)$ igual à $J_R^{\beta}(\cdot)$, $J_H(\cdot)$ ou $J_B(\cdot)$ e calcule $|J^{(t-1)} - J^{(t)}|$;

end

if *Critério de parada 2* **then**

Calcule $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-2)} \cap C_j^{(t-1)})$ e $\sum_{j=1}^{\mathsf{k}} \#(C_j^{(t-1)} \cap C_j^{(t)})$;

end

until $|J^{(t-1)} - J^{(t)}| < \varepsilon$ *ou* $\sum_{j=1}^{k} \#(C_j^{(t-2)} \cap C_j^{(t-1)}) = \sum_{j=1}^{k} \#(C_j^{(t-1)} \cap C_j^{(t)})$;

Para a nomenclatura das variações do algoritmo proposto, considerou-se: a distribuição na qual se baseiam as distâncias, neste caso, a von Mises-Fisher matricial; as distâncias utilizadas; critérios de parada predefinidos (Tabela 8).

Tabela 8 – Definições e nomenclaturas das variações do algoritmo proposto.

Distribuição base	Distância	Critério de parada	Estatística	Nomenclatura
vMF	Battacharyya	1	J_B	vMFB1
vMF	Battacharyya	2	-	vMFB2
vMF	Hellinger	1	J_H	vMFH1
vMF	Hellinger	2	-	vMFH2
vMF	Rényi (ordem β)	1	J_R^β	vMFR1
vMF	Rényi (ordem β)	2	-	vMFR2

2.5 METODOLOGIA

Os códigos que geraram os resultados apresentados neste capítulo foram construídos para uso em ambiente de programação R (R Core Team, 2020), e podem ser consultados em <https://github.com/adneferreira/tese>.

O pacote `Rcpp` foi utilizado juntamente com o pacote `Hypergeomat`. Em conjunto, permitem calcular o valor da função confluente hipergeométrica com argumento matricial. As matrizes utilizadas para gerar dados com distribuição Von Mises-Fisher matricial foram obtidas por meio da função `rmf.matrix`, do pacote `rstiefel` (HOFF, 2009), e na construção dos gráficos, foi utilizado o `ggplot2`.

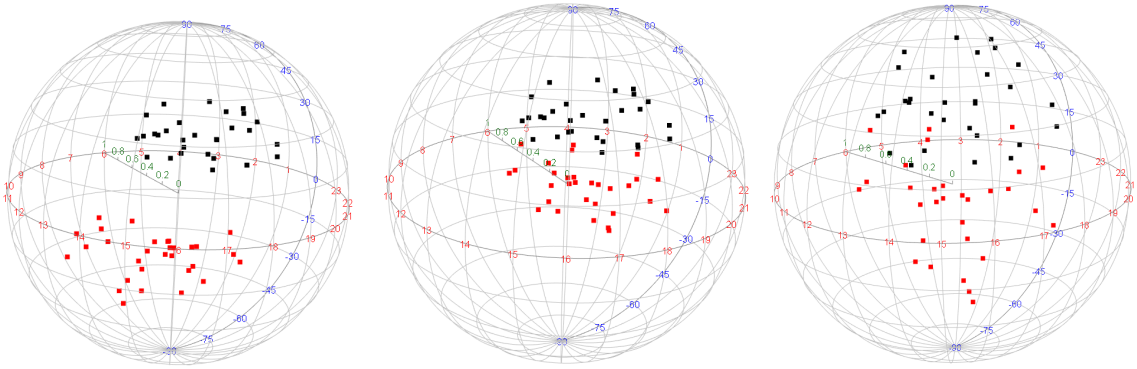
Nas análises, foi verificado o comportamento gráfico das distâncias ao modificar gradativamente um elemento de uma matriz pré-definida. Posteriormente, foram definidos três cenários base para as simulações, como exemplificado na Figura 10.

- **Cenário 1** - neste cenário, o objetivo foi manter os grupos bem definidos e separados, para facilitar o agrupamento. Espera-se que as medidas de desempenho alcancem os maiores valores neste cenário.

- **Cenário 2** - aqui os grupos são justapostos, mas não se misturam. Espera-se que os valores das medidas de desempenho apresentem um leve decaimento neste cenário.

- **Cenário 3** - neste último cenário, os grupos são gerados de maneira a se misturarem, e de forma esparsa, dificultando o agrupamento. Espera-se que as medidas de desempenho decaiam ainda mais neste cenário, quando comparadas com as dos outros dois.

Os parâmetros utilizados para gerar os dados sintéticos estão definidos na subseção 2.5.1, a qual trata, especificamente, da simulação de dados Von Mises-Fisher matriciais.



(a) (b) (c)

Figura 10 – Representação gráfica dos cenários 1 (a), 2 (b) e 3 (c).

Em seguida, foram realizadas 100 simulações de Monte Carlo para definir o valor de β mais adequado para algoritmos com distâncias de Rényi. Nesta etapa, $k = 2$, $p = 3$, $n \in \{1, 2, 3\}$ e os tamanhos amostrais dos grupos 1 e 2 são dados por $n_1 = n_2 = 30$, respectivamente. Consideramos o cenário 3 e um conjunto de valores para $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

Depois de selecionar um valor para β, a análise continuou com a avaliação do impacto do tamanho amostral nos valores dos índices de Rand ajustados, tomando amostras com $n_1 = n_2 \in \{5, 10, 15, 20, 30, 50, 100, 150\}$ no cenário 3. Os resultados foram apresentados graficamente e discutidos. Posteriormente, considerando os cenários 1, 2 e 3, foram realizadas 100 simulações de Monte Carlo em cada. Variando o número de direções $n \in \{1, 2, 3\}$ e dimensão fixada em $p = 3$, médias, medianas e desvios padrões dos coeficientes Kappa e índices Rand foram estimados, bem como o tempo em segundos do cômputo das variações do algoritmo. Por fim, é apresentada uma aplicação em base de dados reais, avaliando a performance e o custo computacional das variações do algoritmo. Os agrupamentos foram realizados considerando dois, três e quatro grupos, $k \in \{2, 3, 4\}$. Nesta etapa, além do cálculo do coeficiente Kappa e índice de Rand ajustado, também foram construídas tabelas com matrizes de confusão. Apesar de tal artifício ser bastante utilizado nas análises de classificadores, em aprendizado de máquinas, como a base dados reais possui a categoria de cada grupo, é possível verificar os grupos que os algoritmos estão tendo melhor desempenho.

A matriz de confusão em nosso contexto consiste numa tabela cruzada dos grupos de referência (verdadeiros) com os grupos preditos resultantes do agrupamento, como segue o exemplo dado na Tabela 9. Os valores F21 e F31 equivalem à quantidade de elementos que pertencem ao grupo 1, mas foram alocados erroneamente nos grupos 2 e 3, respectivamente.

F12 e F32 são as quantidades de objetos do grupo 2 que foram alocados nos grupos 1 e 3, respectivamente. Já F13 e F23 são os elementos que foram erroneamente alocados nos grupos 1 e 2, respectivamente, mas que pertencem em sua origem ao grupo 3. Os valores da diagonal, são os objetos que foram alocados de maneira correta, sendo **V11** para o grupo 1, **V22** para o grupo 2 e **V33** para o grupo 3.

Tabela 9 – Matriz de Confusão - Exemplo para $k = 3$.

Grupo predito	Grupo de referência		
	1	2	3
1	**V11**	F12	F13
2	F21	**V22**	F23
3	F31	F32	**V33**

A interpretação da matriz de confusão é de que quanto menor a quantidade de observações fora da diagonal principal, melhor está sendo o agrupamento. Desta forma, um agrupamento perfeito origina uma matriz de confusão onde os elementos fora da diagonal principal são iguais a zero (GLAROS; KLINE, 1988).

Para rotular os grupos previstos, comparamos os grupos gerados pelas variações do algoritmo com os conjuntos originais. Os grupos que apresentaram maior número de elementos em comum foram considerados pertencentes à mesma categoria. Por exemplo: considere que o algoritmo utilizado gerou dois grupos, inicialmente sem rótulos. Comparando um dos conjuntos previstos com as categorias dos dados originais (anteriormente conhecidos como grupos 1 e 2) constata-se que, majoritariamente, os elementos correspondem aos do grupo 1. Dessa forma, tais dados serão rotulado como grupo 1, apenas para construir a matriz de confusão. O objetivo é identificar em quais condições os algoritmos apresentam melhor desempenho.

2.5.1 Simulação de dados Von Mises-Fisher matriciais

Como já mencionado na seção sobre a distribuição Von Mises-Fisher matricial, existe uma relação entre tal distribuição e normais multivariadas.

Definição 2.5.1. *Considere uma matriz* $\mathbf{M}$*, de dimensão* $n \times p$*, em que* $\mathbf{M}\mathbf{M}^\top = \mathbf{I}_n$ *e* $\mathbf{X}$*, também uma matriz* $n \times p$*, que possui colunas independentes entre si. Considere ainda uma matriz simetrica* $\mathbf{K}$*, de dimensão* $n \times n$ *e que* $\mathbf{F} = \mathbf{K}\mathbf{M}$*. Se as colunas de* $\mathbf{X}$ *seguem distribuição normal* n*-variada com matriz de covariância comum* $\mathbf{K}^{-1}$ *e vetores de médias*

$\boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, ..., \mu_{nj})^\top$, *para* $j = 1, ..., p$, *e* $\boldsymbol{\mu}_j$ *correspondem às colunas de* $\mathbf{M}$, *então* $\mathbf{X} \sim M(p, n, \mathbf{KM})$, *ou seja,* $\mathbf{X} \sim M(p, n, \mathbf{F})$.

Assim, para simular os dados utilizados nas análises, considerando k$= 2$ (dois grupos), os parâmetros definidos foram $\mathbf{F1}$, $\mathbf{M1}$, $\mathbf{K}$ para o grupo 1 e $\mathbf{F2}$, $\mathbf{M2}$, $\mathbf{K}$ para o grupo 2.

Para cada $n \in \{1, 2, 3\}$ foram gerados $\mathbf{M1}$ e $\mathbf{M2}$ por meio da função `rmf.matrix`, do pacote `rstiefel` (HOFF, 2009). Tal função gera uma matriz de dimensão $n \times p$ no espaço Stiefel múltiplo. Os valores definidos são apresentados a seguir.

Para $n = 1$:

$$\mathbf{M1} = \begin{bmatrix} -0.8090170 & 0.3496281 & 0.4724952 \end{bmatrix},$$

$$\mathbf{M2} = \begin{bmatrix} 0.4383711 & 0.7537921 & -0.4895183 \end{bmatrix}.$$

Para $n = 2$:

$$\mathbf{M1} = \begin{bmatrix} -0.2109817 & 0.9539778 & 0.2131033 \\ 0.8557073 & 0.2856346 & -0.4314833 \end{bmatrix},$$

$$\mathbf{M2} = \begin{bmatrix} 0.99089648 & -0.1280309 & -0.04162031 \\ 0.08931486 & 0.8565073 & -0.50834845 \end{bmatrix}.$$

Para $n = 3$:

$$\mathbf{M1} = \begin{bmatrix} -0.4871300 & 0.66999588 & 0.5601873 \\ 0.7957333 & 0.07618675 & 0.6008362 \\ 0.3598789 & 0.73844506 & -0.5702509 \end{bmatrix},$$

$$\mathbf{M2} = \begin{bmatrix} 0.358640829 & 0.93320429 & -0.02250581 \\ 0.933450919 & -0.35835165 & 0.01592096 \\ 0.006792515 & -0.02671798 & -0.99961993 \end{bmatrix}.$$

Em cada cenário (1, 2 e 3), a matriz $\mathbf{K}$ foi modificada com a finalidade de dispersar mais os dados de cada grupo, promovendo uma sobreposição entre eles.

Os cenários são definidos de acordo com os valores que a matriz $\mathbf{K}$ assume. Estabelecemos aqui $\mathbf{K} = \text{diag}(1/\sigma, n)$ como uma matriz de dimensão $n \times n$ cuja diagonal é composta por n repetições de $1/\sigma$ e qualquer valor fora da diagonal é nulo. A matriz inversa de $\mathbf{K}$ é a matriz simétrica $\mathbf{K}^{-1} = \text{diag}(\sigma, n)$. Assim, os cenários foram definidos:

- **Cenário 1** - $\sigma = 0.2$, $\mathbf{K} = \text{diag}(5, n)$ e $\mathbf{K}^{-1} = \text{diag}(0.2, n)$;

- **Cenário 2** - $\sigma = 0.3$, $\mathbf{K} = \text{diag}(3.33, n)$ e $\mathbf{K}^{-1} = \text{diag}(0.3, n)$;
- **Cenário 3** - $\sigma = 0.4$, $\mathbf{K} = \text{diag}(2.5, n)$ e $\mathbf{K}^{-1} = \text{diag}(0.4, n)$.

2.6 RESULTADOS NUMÉRICOS

Buscando-se entender o comportamento das distâncias ao modificar um elemento de uma dada matriz, apresentamos na Figura 11 um gráfico resultante de nosso estudo computacional. Nele, a matriz tomada como referência foi

$$\mathbf{X} = \begin{bmatrix} -1.2529076 & 3.1905616 & 0.9748581 \\ 0.3672866 & 0.6590155 & 1.4766494 \\ -1.6712572 & -1.6409368 & 1.1515627 \end{bmatrix}.$$

Variando apenas o elemento x_{11} de -15 até $+15$, com incrementos de 1 unidade por vez, obtemos o gráfico disposto na Figura 11. Note que as distâncias se aproximam de zero quando o valor de x_{11} se aproxima de -1.2529076. A distância de Hellinger varia no intervalo $[0, 1]$, como esperado, e a de Bhattacharyya aparenta ser mais sensível às mudanças, apresentando distâncias maiores quando x_{11} assume valores mais afastados de -1.2529076.

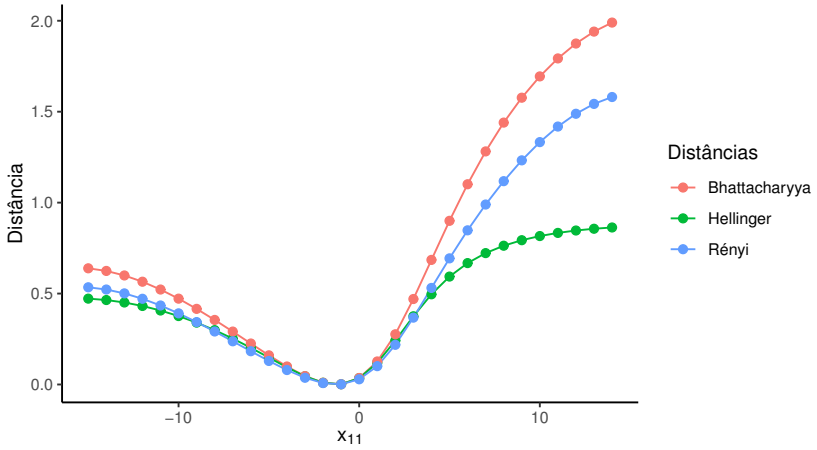


Figura 11 – Distâncias mudando x_{11}.

Seguimos no estudo computacional para verificar qual o β ótimo em vMFR1 e vMFR2. A Figura 12 mostra o comportamento do índice Rand médio de tais algoritmos quando aplicado ao conjunto de valores de $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5,\ 0.6, 0.7, 0.8, 0.9\}$. Observe que médias de índices de Rand ajustados são muito próximas quando comparados os valores de β em cada

número de direções. Assim, $\beta = 0.1$ foi escolhido arbitrariamente, apenas para prosseguir com as simulações.

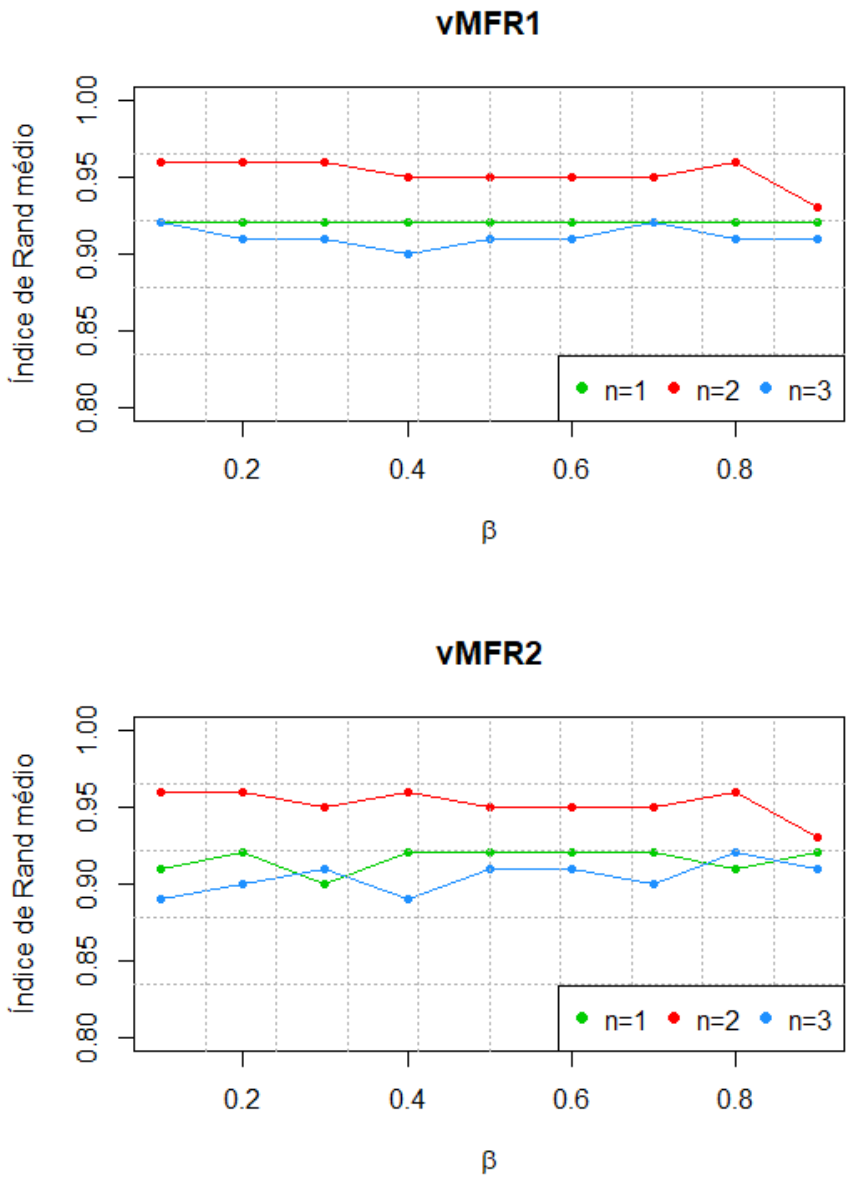


Figura 12 – Comparação entre as médias dos índices de Rand ajustados para os valores de $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ em vMFR1 e vMFR2, por n números de direções.

A Figura 13 mostra o impacto do tamanho da amostra nos valores médios dos índices de Rand ajustados das variações do algoritmo. É perceptível que, à medida que o tamanho da amostra aumenta, os valores médios dos índices Rand ajustados também aumentam.

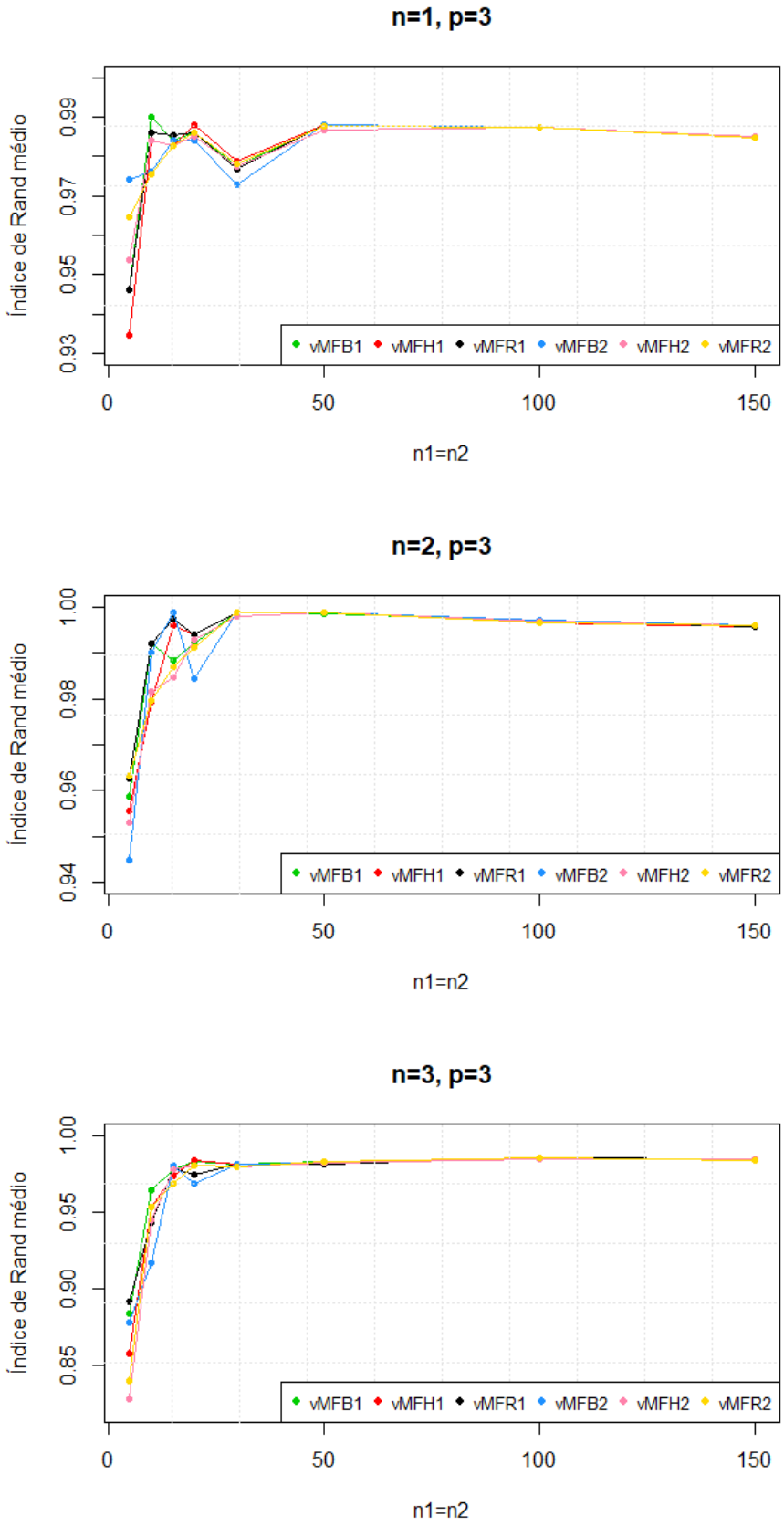


Figura 13 – Comparação das médias de índices de Rand ajustados por tamanhos amostrais $n_1 = n_2 \in \{5, 10, 15, 20, 30, 50, 100, 150\}$.

As Tabelas 10, 11 and 12 apresentam as médias, medianas e desvios padrões das medidas de desempenho: índice de Rand ajustado, Coeficiente Kappa e tempo de processamento do algoritmo. Observe que no cenário 1, independentemente do número de direções n, as medições mostram um excelente desempenho das variações do algoritmo. Nessas condições, o índice Rand ajustado médio variou no intervalo [0.9980;1.0000], e o Coeficiente Kappa médio variou entre [0.9990;1.0000]. Isso era esperado, pois o cenário 1 tem os dois grupos bem definidos e distantes entre si.

Quando partimos para o cenário 2, as médias decaem um pouco. Neste cenário os grupos estão justapostos, portanto há uma dificuldade a mais em identificá-los, diferentemente do cenário 1. Ainda assim, as médias de índice de Rand ajustado e Coeficiente Kappa variaram nos intervalos [0.9768;0.9973] e [0.9883;0.9987], respectivamente.

Nas condições do cenário 3, em que os grupos gerados encontram-se parcialmente sobrepostos e σ é maior, as médias de índice de Rand ajustado variaram entre [0.9433;0.9833], e as de Coeficiente Kappa entre [0.8959;0.9669].

Tabela 10 – Medidas de desempenho nos cenários 1, 2 e 3, para $n = 1$.

Cen.	Algorit.	Coeficiente Kappa			Índ. de Rand Ajust.			Tempo (s)		
		Média	Mediana	DP	Média	Mediana	DP	Média	Mediana	DP
1	vMFB1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	0.5506	0.5650	0.1766
	vMFH1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	0.5230	0.4400	0.1604
	vMFR1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	1.0681	0.9550	0.3024
	vMFB2	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	0.4394	0.4100	0.0893
	vMFH2	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	0.4685	0.4800	0.1061
	vMFR2	0.9993	1.0000	0.0067	0.9987	1.0000	0.0131	0.9447	0.8200	0.3158
2	vMFB1	0.9893	1.0000	0.0195	0.9788	1.0000	0.0384	0.6555	0.6350	0.2028
	vMFH1	0.9890	1.0000	0.0201	0.9782	1.0000	0.0397	0.6483	0.6100	0.2264
	vMFR1	0.9883	1.0000	0.0209	0.9768	1.0000	0.0411	1.3767	1.3000	0.4062
	vMFB2	0.9883	1.0000	0.0209	0.9768	1.0000	0.0411	0.5301	0.5100	0.1777
	vMFH2	0.9883	1.0000	0.0214	0.9769	1.0000	0.0422	0.5417	0.5250	0.1734
	vMFR2	0.9893	1.0000	0.0195	0.9788	1.0000	0.0384	1.0869	1.0400	0.2851
3	vMFB1	0.9520	0.9667	0.0394	0.9063	0.9333	0.0751	0.7556	0.7550	0.2472
	vMFH1	0.9507	0.9667	0.0403	0.9038	0.9333	0.0767	0.7400	0.6450	0.2569
	vMFR1	0.9503	0.9667	0.0386	0.9030	0.9333	0.0734	1.4776	1.4550	0.4528
	vMFB2	0.9507	0.9667	0.0392	0.9037	0.9333	0.0745	0.5757	0.5600	0.1419
	vMFH2	0.9433	0.9667	0.0884	0.8959	0.9333	0.1182	0.5947	0.5850	0.1639
	vMFR2	0.9500	0.9667	0.0398	0.9024	0.9333	0.0755	1.1610	1.0900	0.2614

Tabela 11 – Medidas de desempenho nos cenários 1, 2 e 3, para $n = 2$.

Cen.	Algorit.	Coeficiente Kappa			Índ. de Rand Ajust.			Tempo (s)		
		Média	Mediana	DP	Média	Mediana	DP	Média	Mediana	DP
1	vMFB1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	3.2815	3.2650	0.9542
	vMFH1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	3.3264	3.1250	1.0741
	vMFR1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	6.6835	6.9750	1.7926
	vMFB2	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	3.0159	2.9050	0.7436
	vMFH2	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	3.0877	3.0950	0.5850
	vMFR2	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	5.8958	5.8500	1.1390
2	vMFB1	0.9987	1.0000	0.0066	0.9973	1.0000	0.0131	2.5958	2.4150	0.9159
	vMFH1	0.9987	1.0000	0.0066	0.9973	1.0000	0.0131	2.6978	2.4100	1.0709
	vMFR1	0.9987	1.0000	0.0066	0.9973	1.0000	0.0131	5.3809	4.7800	2.0547
	vMFB2	0.9980	1.0000	0.0093	0.9960	1.0000	0.0184	2.2947	2.1200	0.7777
	vMFH2	0.9987	1.0000	0.0066	0.9973	1.0000	0.0131	2.4008	2.1500	0.8276
	vMFR2	0.9987	1.0000	0.0066	0.9973	1.0000	0.0131	4.6576	4.2650	1.4225
3	vMFB1	0.9830	1.0000	0.0225	0.9662	1.0000	0.0443	2.9980	3.1800	0.8527
	vMFH1	0.9833	1.0000	0.0225	0.9669	1.0000	0.0444	3.0385	3.1900	0.8967
	vMFR1	0.9827	1.0000	0.0225	0.9656	1.0000	0.0442	6.1502	6.3600	1.6678
	vMFB2	0.9793	1.0000	0.0334	0.9595	1.0000	0.0629	2.4519	2.6200	0.5301
	vMFH2	0.9827	1.0000	0.0244	0.9656	1.0000	0.0479	2.5372	2.6400	0.6328
	vMFR2	0.9743	1.0000	0.1008	0.9588	1.0000	0.1064	4.9380	5.2300	1.0344

Tabela 12 – Medidas de desempenho nos cenários 1, 2 e 3, para $n = 3$.

Cen.	Algorit.	Coeficiente Kappa			Índ. de Rand Ajust.			Tempo (s)		
		Média	Mediana	DP	Média	Mediana	DP	Média	Mediana	DP
1	vMFB1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	11.7894	10.8000	4.8880
	vMFH1	0.9997	1.0000	0.0033	0.9993	1.0000	0.0067	11.6018	10.7750	4.3346
	vMFR1	0.9997	1.0000	0.0033	0.9993	1.0000	0.0067	22.1591	21.4100	7.9226
	vMFB2	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	10.2087	9.5200	3.3955
	vMFH2	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000	10.2354	9.4700	3.6209
	vMFR2	0.9990	1.0000	0.0074	0.9980	1.0000	0.0146	19.9937	19.1950	6.2556
2	vMFB1	0.9907	1.0000	0.0184	0.9814	1.0000	0.0363	10.9149	10.2650	4.6695
	vMFH1	0.9903	1.0000	0.0185	0.9808	1.0000	0.0366	10.5775	8.7450	4.6744
	vMFR1	0.9907	1.0000	0.0184	0.9814	1.0000	0.0363	21.0566	17.6500	9.7401
	vMFB2	0.9907	1.0000	0.0184	0.9814	1.0000	0.0363	9.1259	7.1950	4.3142
	vMFH2	0.9897	1.0000	0.0199	0.9795	1.0000	0.0392	8.8609	8.0150	3.6184
	vMFR2	0.9897	1.0000	0.0216	0.9796	1.0000	0.0422	17.8563	15.7050	7.2261
3	vMFB1	0.9503	0.9667	0.0443	0.9035	0.9333	0.0841	14.6315	14.3450	5.7132
	vMFH1	0.9507	0.9667	0.0420	0.9039	0.9333	0.0798	13.8344	12.4250	5.1851
	vMFR1	0.9460	0.9667	0.0464	0.8953	0.9333	0.0876	26.8636	24.0550	9.8963
	vMFB2	0.9483	0.9667	0.0427	0.8994	0.9333	0.0813	11.2667	11.2950	3.6808
	vMFH2	0.9477	0.9667	0.0472	0.8986	0.9333	0.0893	10.7164	10.4600	3.5676
	vMFR2	0.9467	0.9667	0.0457	0.8965	0.9333	0.0864	22.5141	21.9850	7.6388

Quanto ao tempo de processamento, as variações do algoritmo que apresentaram maiores custos computacionais foram as que utilizam a distância de Rényi: vMFR1 e vMFR2. Tais algoritmos necessitam de maior quantidade de operações matemáticas para seu cômputo, considerando que a distância de Rényi é uma equação mais extensa que a de Bhattacharyya e Hellinger. Note também que, ao passo que a dimensão das matrizes aumentam – $n \in \{1, 2, 3\}$ – o tempo de processamento também se estende, pois o cálculo das distâncias e agrupamento das matrizes é mais custoso quando $n \times p$ é maior.

Faz-se necessário ainda pontuar que, quanto mais sobrepostos estiverem os grupos, haverão mais iterações até que haja convergência dos algoritmos, de acordo com os critérios de parada definidos, acrescendo assim, o tempo de processamento. E quanto aos critérios de convergência, quando comparados entre os algoritmos, é possível evidenciar que o critério de parada 2 é menos custoso, pois independe do cálculo das estatísticas de similaridade adaptadas $J_B(\cdot)$, $J_H(\cdot)$ e $J_R^{\beta}(\cdot)$.

2.7 BASES DE DADOS REAIS

As bases de dados reais utilizadas estão disponíveis no livro de Fisher, Lewis e Embleton (1993) (ver Apêndice B). Os quatro conjuntos de dados escolhidos – denotados por B2, B6, B7 e B9 – tratam-se de medidas de remanências magnéticas de diferentes tipos de rochas.

A remanência magnética das rochas se refere à preservação do magnetismo adquirido pela rocha durante sua formação em um campo magnético externo. Quando as rochas são formadas, os minerais magnéticos presentes nelas podem se alinhar com o campo magnético da Terra existente na época. Após a solidificação da rocha, os minerais magnéticos conservam sua orientação magnética. Isso significa que eles preservam uma "impressão" do campo magnético existente naquele momento específico da história da Terra. Essa preservação é conhecida como remanência magnética das rochas. Tal propriedade é importante para o estudo do paleomagnetismo[2] pois fornece informações valiosas sobre a história geológica da Terra (DUNLOP; ÖZDEMIR, 1997).

A remanência magnética das rochas está intimamente relacionada às coordenadas de declinação e inclinação no campo do paleomagnetismo. A declinação e a inclinação são parâmetros que descrevem a direção e a orientação do vetor de magnetização resultante das partículas

[2] Paleomagnetismo é o estudo da evolução do campo magnético e polaridade terrestre em épocas geológicas passadas.

magnéticas presentes nas rochas.

A declinação magnética é a medida angular do desvio entre o norte geográfico e o norte magnético. Essa diferença é causada pela variação do campo magnético da Terra ao longo do tempo e da localização geográfica. A inclinação magnética, por sua vez, é a medida angular da orientação vertical do vetor de magnetização (DUNLOP; ÖZDEMIR, 1997).

As bases B2, B6, B7 e B9 são conjuntos de dados contendo medidas de declinação e inclinação, dadas em graus, de rochas em localidades diferentes, como elencado abaixo:

- B2 - medidas de remanências magnéticas em espécimes de *red-beds* (rochas sedimentares avermelhadas) paleozóicas da Argentina (Tabela 13);
- B6 - medidas de remanências magnéticas em espécimes de rochas vulcânicas da era Pré-Cambriana (Tabela 15);
- B7 - medições de remanências magnéticas em espécimes coletados do arenito Tumbaglooda, na Austrália Ocidental (Tabela 14);
- B9 - medições da direção da magnetização em espécimes do *Great Whin Sill* (Tabela 16).

Tabela 13 – Base de dados B2.

	Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)
1	122.5	55.5	10	147.5	54.5	19	141.0	57.0
2	130.5	58.0	11	142.0	51.0	20	143.5	67.5
3	132.5	44.0	12	163.5	56.0	21	131.5	62.5
4	148.5	56.0	13	141.0	59.5	22	147.5	63.5
5	140.0	63.0	14	156.0	56.5	23	147.0	55.5
6	133.0	64.5	15	139.5	54.0	24	149.0	62.0
7	157.5	53.0	16	153.5	47.5	25	144.0	53.5
8	153.0	44.5	17	151.5	61.0	26	139.5	58.0
9	140.0	61.5	18	147.5	58.5	-	-	-

Tabela 14 – Base de dados B7.

	Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)
1	66	-11	7	25	-14	13	63	-21
2	73	-31	8	42	-36	14	63	-24
3	76	-38	9	51	-36	15	64	-33
4	78	-26	10	52	-24	16	81	-30
5	83	-58	11	56	-34	17	81	-43
6	122	-50	12	61	-35	-	-	-

Tabela 15 – Base de dados B6.

	Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)
1	30.7	319.1	37	62.6	323.2	73	22.5	89.4
2	-49.2	347.4	38	62.3	210.5	74	70.3	339.7
3	-37.3	344.5	39	-16.3	316.0	75	71.9	359.9
4	-11.4	313.6	40	-19.2	306.9	76	-45.1	314.1
5	31.3	177.0	41	-31.0	331.0	77	-24.4	325.4
6	-31.3	3.0	42	-12.3	298.1	78	-17.6	1.9
7	4.6	308.0	43	-21.7	296.2	79	-7.1	177.3
8	42.8	165.0	44	-14.5	10.9	80	-68.3	324.5
9	-22.2	352.6	45	28.8	249.0	81	-58.1	346.9
10	8.8	307.0	46	38.2	262.3	82	-57.5	350
11	19.3	190.3	47	31.8	270.0	83	-1.6	321.4
12	-28.2	3.6	48	-6.0	338.5	84	-5.7	314.6
13	-2.3	306.1	49	-2.8	332.2	85	-22.3	320.6
14	32.4	180.2	50	3.4	222.3	86	-17.3	304.7
15	-29.7	9.1	51	-20.6	3.3	87	-22.6	314.2
16	-39.3	355.7	52	-5.7	325.9	88	-16.5	306.3
17	21.2	286.8	53	4.6	345.6	89	-18.4	271.6
18	19.8	207.7	54	-34.4	294.4	90	-36.7	316
19	18.9	192.1	55	-8.2	358.7	91	-20.3	309.8
20	35.8	202.3	56	-44.3	23.8	92	-20.5	313
21	41.3	44.9	57	-43.5	20.8	93	-32.3	329.4
22	-41.4	23.8	58	8.7	280.5	94	-29.3	334.9
23	-25.9	9.3	59	-47.2	347.4	95	-15.7	312.3
24	-15.8	18.3	60	-48.3	341.9	96	5.9	269.8
25	27.3	186.2	61	-17.4	10.2	97	-24.9	318.2
26	-0.8	31.1	62	-2.7	12.4	98	-40	0.4
27	12.7	165.5	63	-17.6	321.0	99	-15.2	309
28	37.2	202.3	64	-45.5	16.7	100	-21.6	310.4
29	59.7	191.5	65	-22.7	353.5	101	-43.3	348
30	-16.4	297.2	66	26.8	264.8	102	41	291.9
31	6.3	248.7	67	39.1	258.5	103	51.4	272.2
32	-57.5	155.8	68	15.3	310.6	104	43.9	273
33	11.8	190.1	69	-12.0	281.2	105	-25.1	338.2
34	-11.4	341.3	70	-15.4	7.0	106	-4.7	318.5
35	-3.4	346.3	71	23.2	342.8	107	19.2	298.2
36	3.0	184.6	72	47.2	243.6	-	-	-

Tabela 16 – Base de dados B9.

	Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)		Decl. (°)	Incl. (°)
1	176	-10	13	187	-15	25	195	11
2	185	-6	14	203	-26	26	182	11
3	196	-5	15	196	-19	27	182	7
4	179	-4	16	205	6	28	190	-15
5	186	-5	17	180	9	29	183	-20
6	192	3	18	190	-13	30	194	-7
7	183	-1	19	183	1	31	186	-6
8	203	-11	20	190	15	32	192	-9
9	190	-28	21	183	7	33	192	0
10	171	-14	22	179	11	34	180	0
11	191	-14	123	90	13	-	-	-
12	190	-20	24	177	-10	-	-	-

As medidas de declinação e inclinação são dadas em graus e, para que sejam descritas no espaço Stiefel múltiplo precisam ser submetidas a uma transformação, resultando em coordenadas no espaço $\mathbb{R}^3$, como se segue.

Seja $\mathbf{v} = (v_1, v_2)^\top$ um vetor contendo medidas de declinação (v_1) e inclinação (v_2) de remanência magnética de uma rocha, e sejam x, y e z suas respectivas coordenadas em $\mathbb{R}^3$.

Primeiro, deve-se transformar $\mathbf{v} = (v_1, v_2)^\top$, que está em graus, para radianos. Assim, tem-se o vetor $\mathbf{v}^\pi = (v_1^\pi, v_2^\pi)^\top$, dado em radianos, em que

$$v_i^\pi = v_i \frac{\pi}{180},$$

para i=1,2.

Agora, considere uma amostra aleatória de N vetores de remanências magnéticas, dadas pelo conjunto de $\mathbf{v}^\pi{}_j = (v_{1j}^\pi, v_{2j}^\pi)^\top$, para $j = 1, 2, ...$,N. A transformação para coordenadas em $\mathbb{R}^3$ correspondente a cada vetor $\mathbf{v}^\pi{}_j$ é dada por

$$\boldsymbol{\eta}_j = (cos(v_{1j}^\pi), \sin(v_{1j}^\pi)\cos(v_{2j}^\pi), \sin(v_{1j}^\pi)\sin(v_{2j}^\pi)^\top. \tag{2.23}$$

A partir daqui, a amostra pode ser representada por matrizes $\mathbf{X}_j$, para $j = 1, 2, ...$N, de dimensões $n \times p$, em que $n = 1$, $p = 3$ e $\mathbf{X}_j\mathbf{X}_j^\top = 1$. As Tabelas 17, 19, 18 e 20 apresentam as coordenadas em $\mathbb{R}^3$ resultantes para cada vetor de remanência magnética das bases de dados B2, B6, B7 e B9, respectivamente.

Tabela 17 – Base de dados B2 transformados para $\mathbb{R}^3$.

	x	y	z		x	y	z		x	y	z
1	-0.54	0.48	0.7	10	-0.84	0.31	0.44	19	-0.78	0.34	0.53
2	-0.65	0.4	0.64	11	-0.79	0.39	0.48	20	-0.8	0.23	0.55
3	-0.68	0.53	0.51	12	-0.96	0.16	0.24	21	-0.66	0.35	0.66
4	-0.85	0.29	0.43	13	-0.78	0.32	0.54	22	-0.84	0.24	0.48
5	-0.77	0.29	0.57	14	-0.91	0.22	0.34	23	-0.84	0.31	0.45
6	-0.68	0.31	0.66	15	-0.76	0.38	0.53	24	-0.86	0.24	0.45
7	-0.92	0.23	0.31	16	-0.89	0.3	0.33	25	-0.81	0.35	0.47
8	-0.89	0.32	0.32	17	-0.88	0.23	0.42	26	-0.76	0.34	0.55
9	-0.77	0.31	0.56	18	-0.84	0.28	0.46	-	-	-	-

Tabela 18 – Base de dados B7 transformados para $\mathbb{R}^3$.

	x	y	z		x	y	z		x	y	z
1	0.41	0.9	-0.17	7	0.91	0.41	-0.1	13	0.45	0.83	-0.32
2	0.29	0.82	-0.49	8	0.74	0.54	-0.39	14	0.45	0.81	-0.36
3	0.24	0.76	-0.6	9	0.63	0.63	-0.46	15	0.44	0.75	-0.49
4	0.21	0.88	-0.43	10	0.62	0.72	-0.32	16	0.16	0.86	-0.49
5	0.12	0.53	-0.84	11	0.56	0.69	-0.46	17	0.16	0.72	-0.67
6	-0.53	0.55	-0.65	12	0.48	0.72	-0.5	-	-	-	-

Tabela 19 – Base de dados B6 transformados para $\mathbb{R}^3$.

	x	y	z		x	y	z		x	y	z
1	0.86	0.39	-0.33	37	0.46	0.71	-0.53	73	0.92	0	0.38
2	0.65	-0.74	0.17	38	0.46	-0.76	-0.45	74	0.34	0.88	-0.33
3	0.8	-0.58	0.16	39	0.96	-0.2	0.19	75	0.31	0.95	0
4	0.98	-0.14	0.14	40	0.94	-0.2	0.26	76	0.71	-0.49	0.51
5	0.85	-0.52	0.03	41	0.86	-0.45	0.25	77	0.91	-0.34	0.23
6	0.85	-0.52	-0.03	42	0.98	-0.1	0.19	78	0.95	-0.3	-0.01
7	1	0.05	-0.06	43	0.93	-0.16	0.33	79	0.99	0.12	-0.01
8	0.73	-0.66	0.18	44	0.97	-0.25	-0.05	80	0.37	-0.76	0.54
9	0.93	-0.37	0.05	45	0.88	-0.17	-0.45	81	0.53	-0.83	0.19
10	0.99	0.09	-0.12	46	0.79	-0.08	-0.61	82	0.54	-0.83	0.15
11	0.94	-0.33	-0.06	47	0.85	0	-0.53	83	1	-0.02	0.02
12	0.88	-0.47	-0.03	48	0.99	-0.1	0.04	84	1	-0.07	0.07
13	1	-0.02	0.03	49	1	-0.04	0.02	85	0.93	-0.29	0.24
14	0.84	-0.54	0	50	1	-0.04	-0.04	86	0.95	-0.17	0.24
15	0.87	-0.49	-0.08	51	0.94	-0.35	-0.02	87	0.92	-0.27	0.28
16	0.77	-0.63	0.05	52	1	-0.08	0.06	88	0.96	-0.17	0.23
17	0.93	0.1	-0.35	53	1	0.08	-0.02	89	0.95	-0.01	0.32
18	0.94	-0.3	-0.16	54	0.83	-0.23	0.51	90	0.8	-0.43	0.42
19	0.95	-0.32	-0.07	55	0.99	-0.14	0	91	0.94	-0.22	0.27
20	0.81	-0.54	-0.22	56	0.72	-0.64	-0.28	92	0.94	-0.24	0.26
21	0.75	0.47	0.47	57	0.73	-0.64	-0.24	93	0.85	-0.46	0.27
22	0.75	-0.61	-0.27	58	0.99	0.03	-0.15	94	0.87	-0.44	0.21
23	0.9	-0.43	-0.07	59	0.68	-0.72	0.16	95	0.96	-0.18	0.2
24	0.96	-0.26	-0.09	60	0.67	-0.71	0.23	96	0.99	0	-0.1
25	0.89	-0.46	-0.05	61	0.95	-0.29	-0.05	97	0.91	-0.31	0.28
26	1	-0.01	-0.01	62	1	-0.05	-0.01	98	0.77	-0.64	0
27	0.98	-0.21	0.06	63	0.95	-0.23	0.19	99	0.97	-0.17	0.2
28	0.8	-0.56	-0.23	64	0.7	-0.68	-0.2	100	0.93	-0.24	0.28
29	0.5	-0.85	-0.17	65	0.92	-0.38	0.04	101	0.73	-0.67	0.14
30	0.96	-0.13	0.25	66	0.89	-0.04	-0.45	102	0.75	0.24	-0.61
31	0.99	-0.04	-0.1	67	0.78	-0.13	-0.62	103	0.62	0.03	-0.78
32	0.54	0.77	-0.35	68	0.96	0.17	-0.2	104	0.72	0.04	-0.69
33	0.98	-0.2	-0.04	69	0.98	-0.04	0.2	105	0.91	-0.39	0.16
34	0.98	-0.19	0.06	70	0.96	-0.26	-0.03	106	1	-0.06	0.05
35	1	-0.06	0.01	71	0.92	0.38	-0.12	107	0.94	0.16	-0.29
36	1	-0.05	0	72	0.68	-0.33	-0.66	-	-	-	-

Tabela 20 – Base de dados B9 transformados para $\mathbb{R}^3$.

	x	y	z		x	y	z		x	y	z
1	-1	0.07	-0.01	13	-0.99	-0.12	0.03	25	-0.97	-0.25	-0.05
2	-1	-0.09	0.01	14	-0.92	-0.35	0.17	26	-1	-0.03	-0.01
3	-0.96	-0.27	0.02	15	-0.96	-0.26	0.09	27	-1	-0.03	0
4	-1	0.02	0	16	-0.91	-0.42	-0.04	28	-0.98	-0.17	0.04
5	-0.99	-0.1	0.01	17	-1	0	0	29	-1	-0.05	0.02
6	-0.98	-0.21	-0.01	18	-0.98	-0.17	0.04	30	-0.97	-0.24	0.03
7	-1	-0.05	0	19	-1	-0.05	0	31	-0.99	-0.1	0.01
8	-0.92	-0.38	0.07	20	-0.98	-0.17	-0.04	32	-0.98	-0.21	0.03
9	-0.98	-0.15	0.08	21	-1	-0.05	-0.01	33	-0.98	-0.21	0
10	-0.99	0.15	-0.04	22	-1	0.02	0	34	-1	0	0
11	-0.98	-0.19	0.05	23	-0.98	-0.17	-0.04	-	-	-	-
12	-0.98	-0.16	0.06	24	-1	0.05	-0.01	-	-	-	-

A Tabela 21 traz as principais medidas descritivas dos dados de remanência magnética das rochas, estratificadas por bases de dados (B2, B6, B7 e B9). As medianas das coordenadas em $\mathbb{R}^3$ são próximas das suas respectivas médias, o que sugere que aproximadamente metade dos valores em cada coordenada encontram-se abaixo da média, se a média for maior ou igual à mediana, ou acima dela, se a média for menor que a mediana.

Os desvios padrões variaram entre 0.04 e 0.35, e as base que apresentaram os maiores desvios foram B6 e B7. Tal medida de dispersão retrata a distância de um caonjunto de valores em relação a sua própria média e ajuda a compreender o quão "espalhados" estão os dados. Quanto mais dispersos, maior a chance dos grupos se sobreporem, e, consequentemente, pior o desempenho do algoritmo de agrupamento.

Tabela 21 – Estatísticas descritivas das bases B2, B6, B7 e B9 em $\mathbb{R}^3$.

Base	Coordenadas em $\mathbb{R}^3$	Média	Desvio Padrão	Mediana	Mínimo	Máximo
B2	x	-0.80	0.10	-0.81	-0.96	-0.54
	y	0.31	0.08	0.31	0.16	0.53
	z	0.49	0.12	0.48	0.24	0.70
B6	x	0.86	0.16	0.92	0.31	1.00
	y	-0.22	0.35	-0.21	-0.85	0.95
	z	-0.01	0.27	-0.00	-0.78	0.54
B7	x	0.37	0.32	0.44	-0.53	0.91
	y	0.71	0.14	0.72	0.41	0.90
	z	-0.46	0.18	-0.46	-0.84	-0.10
B9	x	-0.98	0.02	-0.99	-1.00	-0.91
	y	-0.13	0.13	-0.14	-0.42	0.15
	z	0.01	0.04	0.00	-0.05	0.17

A Figura 14 mostra a representação gráfica bidimensional e trimensional das bases de dados B2, B6, B7 e B9. Note que, no contexto tridimensional, os grupos B2, B7 e B9 encontram-se bem distintos, enquanto o grupo B6 se sobrepõe, principalmente, sobre B7.

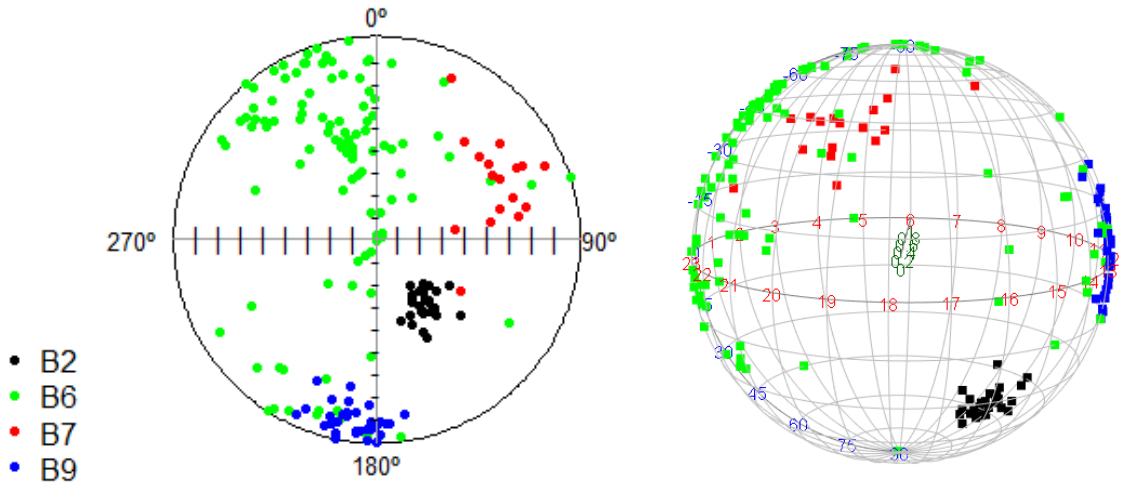


Figura 14 – Representação gráfica bidimensional e tridimensional das bases de dados reais B2, B6, B7 e B9 (FISHER; LEWIS; EMBLETON, 1993).

A Tabela 22 apresenta os resultados do índice de Rand ajustado e coeficiente Kappa para cada variação do algoritmo proposto, quando os grupos são comparados dois a dois. Na maioria dos cenários, as variações do algoritmo obteve excelente desempenho, com índices de Rand ajustado e coeficientes Kappa iguais a 1.00, com exceção apenas do agrupamento

considerando as bases B6 *versus* B7, a qual obteve índice de Rand ajustado igual a 0.6716 e coeficiente Kappa igual a 0.7491. Vale ressaltar que os grupos B6 e B7 foram os que apresentaram maiores desvios padrões em suas coordenadas, e a dificuldade em agrupar tais dados de maneira perfeita já era esperada.

Quanto ao custo computacional, semelhante aos resultados obtidos em simulação, as variações do algoritmo com maior tempo de processamento foram as baseadas nas distâncias de Rényi, variando entre 0.44 e 2.32 segundos (Tabela 22).

Tabela 22 – Desempenho do agrupamento com k=2.

Grupos	Variações do algoritmo	Coeficiente Kappa	Índice de Rand	Tempo (s)
B2 x B6	vMFB1	1.0000	1.0000	1.1900
	vMFH1	1.0000	1.0000	0.9200
	vMFR1	1.0000	1.0000	1.7900
	vMFB2	1.0000	1.0000	0.8400
	vMFH2	1.0000	1.0000	0.7900
	vMFR2	1.0000	1.0000	1.5400
B2 x B7	vMFB1	1.0000	1.0000	0.3000
	vMFH1	1.0000	1.0000	0.3100
	vMFR1	1.0000	1.0000	0.5800
	vMFB2	1.0000	1.0000	0.2700
	vMFH2	1.0000	1.0000	0.2700
	vMFR2	1.0000	1.0000	0.5200
B2 x B9	vMFB1	1.0000	1.0000	0.2800
	vMFH1	1.0000	1.0000	0.5000
	vMFR1	1.0000	1.0000	0.5500
	vMFB2	1.0000	1.0000	0.2800
	vMFH2	1.0000	1.0000	0.3200
	vMFR2	1.0000	1.0000	0.5600
B6 x B7	vMFB1	0.7491	0.6716	1.0400
	vMFH1	0.7491	0.6716	0.8500
	vMFR1	0.7491	0.6716	1.6700
	vMFB2	0.7491	0.6716	1.0300
	vMFH2	0.7491	0.6716	0.7000
	vMFR2	0.7491	0.6716	1.3900
B6 x B9	vMFB1	1.0000	1.0000	1.2600
	vMFH1	1.0000	1.0000	1.2500
	vMFR1	1.0000	1.0000	2.3200
	vMFB2	1.0000	1.0000	1.0400
	vMFH2	1.0000	1.0000	1.0400
	vMFR2	1.0000	1.0000	1.9900
B7 x B9	vMFB1	1.0000	1.0000	0.2300
	vMFH1	1.0000	1.0000	0.4400
	vMFR1	1.0000	1.0000	0.4300
	vMFB2	1.0000	1.0000	0.2200
	vMFH2	1.0000	1.0000	0.2300
	vMFR2	1.0000	1.0000	0.4400

As Tabelas 23, 24, 25 e 26 apresentam os desempenhos das variações do algoritmo nos cenário em que se consideram os grupos de três bases de dados: B2, B6 e B7; B2, B7 e B9; B2, B6 e B9; B6, B7 e B9, respectivamente.

As variações do algoritmo apresentaram excelente agrupamento – índice de Rand ajustado e coeficiente Kappa iguais a 1.00 – apenas no cenário em que havia ausência da base B6, como mostra a Tabela 24.

O cenário menos favorável ao agrupamento foi o que se considera B2, B6 e B9. Nesse contexto, os índices de Rand ajustados e os coeficientes Kappas decaíram para 0.6177 e 0.5205, respectivamente. Pela matriz de confusão, percebe-se que, na presença de B6, os algoritmos alocam todos os elementos de B2 em B9, e 20 elementos de B6 são alocados erroneamente em B2.

As variações do algoritmo que apresentaram maior tempo de processamento foram vMFR1 e vMFR2, com valores variando entre 2.00 e 7.96 segundos.

Tabela 23 – Tempos de processamento, Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B2, B6 e B7.

Variações do Algoritmo	Grupo predito	Grupo real B2	Grupo real B6	Grupo real B7	Coeficiente Kappa	Índice de Rand	Tempo (s)
vMFB1	B2	**26**	0	0	0.8761	0.8154	4.21
	B6	0	**98**	0			
	B7	0	9	**17**			
vMFH1	B2	**26**	0	0	0.8761	0.8154	3.95
	B6	0	**98**	0			
	B7	0	9	**17**			
vMFR1	B2	**26**	0	0	0.8761	0.8154	7.96
	B6	0	**98**	0			
	B7	0	9	**17**			
vMFB2	B2	**26**	0	0	0.8761	0.8154	3.06
	B6	0	**98**	0			
	B7	0	9	**17**			
vMFH2	B2	**26**	0	0	0.8761	0.8154	3.48
	B6	0	**98**	0			
	B7	0	9	**17**			
vMFR2	B2	**26**	0	0	0.8761	0.8154	6.42
	B6	0	**98**	0			
	B7	0	9	**17**			

Tabela 24 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B2, B7 e B9.

Variações do Algoritmo	Grupo predito	Grupo real B2	B7	B9	Coeficiente Kappa	Índice de Rand	Tempo (s)
vMFB1	B2	**26**	0	0	1.0000	1.0000	1.31
	B7	0	**17**	0			
	B9	0	0	**34**			
vMFH1	B2	**26**	0	0	1.0000	1.0000	1.17
	B7	0	**17**	0			
	B9	0	0	**34**			
vMFR1	B2	**26**	0	0	1.0000	1.0000	2.05
	B7	0	**17**	0			
	B9	0	0	**34**			
vMFB2	B2	**26**	0	0	1.0000	1.0000	1.24
	B7	0	**17**	0			
	B9	0	0	**34**			
vMFH2	B2	**26**	0	0	1.0000	1.0000	1.03
	B7	0	**17**	0			
	B9	0	0	**34**			
vMFR2	B2	**26**	0	0	1.0000	1.0000	2.00
	B7	0	**17**	0			
	B9	0	0	**34**			

Tabela 25 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B2, B6 e B9.

Variações do Algoritmo	Grupo predito	Grupo real B2	Grupo real B6	Grupo real B9	Coeficiente Kappa	Índice de Rand	Tempo (s)
vMFB1	B2	**0**	20	0	0.5205	0.6177	1.40
	B6	0	**87**	0			
	B9	26	0	**34**			
vMFH1	B2	**0**	20	0	0.5205	0.6177	1.35
	B6	0	**87**	0			
	B9	26	0	**34**			
vMFR1	B2	**0**	20	0	0.5205	0.6177	2.94
	B6	0	**87**	0			
	B9	26	0	**34**			
vMFB2	B2	**0**	20	0	0.5205	0.6177	1.67
	B6	0	**87**	0			
	B9	26	0	**34**			
vMFH2	B2	**0**	20	0	0.5205	0.6177	1.50
	B6	0	**87**	0			
	B9	26	0	**34**			
vMFR2	B2	**0**	20	0	0.5205	0.6177	2.79
	B6	0	**87**	0			
	B9	26	0	**34**			

Tabela 26 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com três grupos: B6, B7 e B9.

Variações do Algoritmo	Grupo predito	Grupo real B6	Grupo real B7	Grupo real B9	Coeficiente Kappa	Índice de Rand	Tempo (s)
vMFB1	B6	**98**	0	0	0.8772	0.8269	3.16
	B7	9	**16**	0			
	B9	0	1	**34**			
vMFH1	B6	**98**	0	0	0.8772	0.8269	3.08
	B7	9	**16**	0			
	B9	0	1	**34**			
vMFR1	B6	**98**	0	0	0.8772	0.8269	7.03
	B7	9	**16**	0			
	B9	0	1	**34**			
vMFB2	B6	**95**	0	0	0.8436	0.7779	3.08
	B7	12	**16**	0			
	B9	0	1	**34**			
vMFH2	B6	**95**	0	0	0.8436	0.7779	2.75
	B7	12	**16**	0			
	B9	0	1	**34**			
vMFR2	B6	**95**	0	0	0.8436	0.7779	3.57
	B7	12	**16**	0			
	B9	0	1	**34**			

A Tabela 27 apresenta o desempenho das variações do algoritmo no cenário em que se consideram os quatro grupos (B2, B6, B7 e B9). Nesse contexto, os índices de Rand ajustados e os coeficientes Kappas decaem para 0.4659 e 0.4589, respectivamente. Pela matriz de confusão, percebe-se que o grupo B6 interfere no agrupamento e se mistura com B2 e B7, além disso, os grupos B2 e B9 foram alocados todos no mesmo grupo.

Quanto ao custo computacional, permanecem vMFR1 e vMFR2 como os algoritmos com maior tempo de processamento, sendo 8.36 e 12.10 segundos, respectivamente.

Tabela 27 – Coeficientes Kappas, Índices de Rand e Matrizes de confusão dos resultados para o cenário com os quatro grupos: B2, B6, B7 e B9.

Variações do Algoritmo	Grupo predito	Grupo real B2	B6	B7	B9	Coeficiente Kappa	Índice de Rand	Tempo (s)
vMFB1	B2	**0**	34	0	0	0.4589	0.4659	4.67
	B6	0	**65**	1	0			
	B7	0	8	**16**	0			
	B9	26	0	0	**34**			
vMFH1	B2	**0**	34	0	0	0.4589	0.4659	4.19
	B6	0	**65**	1	0			
	B7	0	8	**16**	0			
	B9	26	0	0	**34**			
vMFR1	B2	**0**	34	0	0	0.4589	0.4659	8.36
	B6	0	**65**	1	0			
	B7	0	8	**16**	0			
	B9	26	0	0	**34**			
vMFB2	B2	**0**	34	0	0	0.4589	0.4659	5.02
	B6	0	**65**	1	0			
	B7	0	8	**16**	0			
	B9	26	0	0	**34**			
vMFH2	B2	**0**	34	0	0	0.4589	0.4659	6.26
	B6	0	**65**	1	0			
	B7	0	8	**16**	0			
	B9	26	0	0	**34**			
vMFR2	B2	**0**	34	0	0	0.4589	0.4659	12.10
	B6	0	**65**	1	0			
	B7	0	8	**16**	0			
	B9	26	0	0	**34**			

2.8 CONSIDERAÇÕES FINAIS

Os dados direcionais podem ser encontrados em diversas aplicações, dentre elas: dados de vetorcardiograma clínico, dados paleomagnéticos ou dados cristalográficos.

Nas análises gráficas, observamos que as distâncias apresentaram o desempenho esperado. Quando modificado um elemento da matriz utilizada no cálculo, e, dentre as distâncias propostas, a de Bhattacharyya mostrou-se mais sensível às mudanças.

Foi proposto um algoritmo de agrupamento que utiliza tais distâncias e possui seis variações possíveis: vMFB1, vMFB2, vMFH1, vMFH2, vMFR1 e vMFR2. Simulações de Monte Carlo

foram realizadas com a finalidade de avaliar o desempenho de tais variações.

Quando nos cenários de dados sintéticos, o valor de $\beta = 0.1$ foi escolhido arbitrariamente dentre o *grid* porque as médias de índices de Rand ajustado foram muito semelhantes entre os diferentes β's.

Com relação ao tamanho amostral, amostras maiores resultaram em melhores agrupamentos. E quanto aos cenários pré-definidos, os melhores desempenhos foram verificados no cenário 1, em que os grupos estão bem separados entre si. Com relação ao custo computacional, matrizes com maiores dimensões demandam mais tempo, assim como os algoritmos com cômputo de estatísticas de similaridade adaptadas no critério de convergência.

Nos dados reais, as variações do algoritmo proposto obtiveram resultados semelhantes, com índices de Rand ajustados variando entre 0.6716 e 1.0000 nos cenários dos grupos comparados dois a dois. Quando aplicados à base de dados completa (quatro categorias) ou com três categorias de configurações, o desempenho se tornou mais realista, e o custo computacional foi mais elevado.

Esta pesquisa contribuiu com a adição de novas distâncias no contexto do espaço esférico, além de apresentar um novo algoritmo e suas variações. As simulações mostram resultados promissores dos algoritmos e a aplicação em dados reais evidencia a utilidade de tais ferramentas.

Em resumo, as novas distâncias aliadas à construção desse novo algoritmo de agrupamento representam um avanço importante no campo da análise de dados, oferecendo ferramentas valiosas para a extração de informações e a compreensão de padrões em conjuntos de dados complexos, como os tridimensionais.

SUBMISSÕES E TRABALHOS FUTUROS

A pesquisa e aplicações apresentadas no Capítulo 1 desta tese geraram dois trabalhos. Um deles foi submetido em resumo, foi aceito e apresentado no XLI Congresso Nacional de Matemática Aplicada e Computacional (CNMAC 2022). O outro foi redigido em formato de artigo e submetido para o *Brazilian Journal of Probability and Statistics*. No momento, segue aguardando avaliação dos revisores.

Uma pré-impressão proveniente da pesquisa que desenvolvemos no Capítulo 2, foi submetida à revista *Journal of classification*. No momento, segue aguardando avaliação dos revisores.

Nosso algoritmo de agrupamento é capaz de segregar em grupos dados matriciais que estejam no espaço múltiplo de Manifold, e dados de vetorcardiograma são dados direcionais que possuem essa característica. Certamente, tal algoritmo seja útil em situações de tomada de decisão em Saúde, com o retorno do uso do vetorcardiograma, em conjunto com o eletrocardiograma, para detectar cardiopatias.

A ideia é continuar pesquisas nessa área, solicitando e/ou coletando dados de vetorcardiograma em associação com alguma empresa que possua o maquinário necessário para obter tais informações, e implementar um estudo sobre esses bancos, identificando cardiopatas e não cardiopatas, além de verificar se é possível também identificar o tipo de cardiopatia.

Outra análise interessante poderia ser implementada no contexto do agrupamento de imagens tridimensionais ao considerar pontos que gerem vetores ortogonais.

REFERÊNCIAS

AMARAL, G. A.; DRYDEN, I.; WOOD, A. T. A. Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *Journal of the American Statistical Association*, Taylor & Francis, v. 102, n. 478, p. 695–707, 2007.

ARTHUR, D.; VASSILVITSKII, S. *k-means++: The Advantages of Careful Seeding*. [S.l.], 2006. Disponível em: <http://ilpubs.stanford.edu:8090/778/>.

ASSIS, E. C. D.; SOUZA, R. M. C. R. D.; AMARAL, G. J. A. D. Using bagging to enhance clustering procedures for planar shapes. *International Journal of Business Intelligence and Data Mining*, Inderscience Publishers (IEL), v. 18, n. 1, p. 30–48, 2021.

BARTLE, R. G.; SHERBERT, D. R. *Introduction to real analysis*. [S.l.]: Wiley New York, 2000. v. 2.

BISHOP, C. M.; NASRABADI, N. M. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006. v. 4.

BOOKSTEIN, F. L. *Morphometric tools for landmark data: geometry and biology*. [S.l.]: Cambridge University Press, 1997.

BOWMAN, A. W.; AZZALINI, A. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. [S.l.]: OUP Oxford, 1997. v. 18.

BUSSAB, W. d. O.; MORETTIN, P. A. *Estatística básica*. [S.l.]: São Paulo: Saraiva, 2010. 540 p.

CHUNG, J.; KANNAPPAN, P.; NG, C.; SAHOO, P. Measures of distance between probability distributions. *Journal of mathematical analysis and applications*, Elsevier, v. 138, n. 1, p. 280–292, 1989.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

DAVID, F. N. *Games, gods, and gambling: A history of probability and statistical ideas*. [S.l.]: Courier Corporation, 1998.

DIAZ-CHITO, K.; HERNÁNDEZ-SABATÉ, A.; LÓPEZ, A. M. A reduced feature set for driver head pose estimation. *Applied Soft Computing*, Elsevier, v. 45, p. 98–107, 2016.

DIETHE, T. A note on the kullback-leibler divergence for the von mises-fisher distribution. *arXiv preprint arXiv:1502.07104*, 2015.

DOWNS, T. D. Orientation statistics. *Biometrika*, Oxford University Press, v. 59, n. 3, p. 665–676, 1972.

DRYDEN, I. L.; MARDIA, K. V. *Statistical shape analysis*. [S.l.]: John Wiley & Sons, 1998. v. 1.

DRYDEN, I. L.; MARDIA, K. V. *Statistical shape analysis: with applications in R*. [S.l.]: John Wiley & Sons, 2016. v. 995.

DUNLOP, D. J.; ÖZDEMIR, Ö. *Rock magnetism: fundamentals and frontiers*. [S.l.]: Cambridge university press, 1997.

FÉLIX, W. V. Statistical inference based on information theory for pre-shape data. Universidade Federal de Pernambuco, 2019.

FISHER, N. I.; LEWIS, T.; EMBLETON, B. J. *Statistical analysis of spherical data*. [S.l.]: Cambridge university press, 1993.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. et al. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001. v. 1.

GLAROS, A. G.; KLINE, R. B. Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of clinical psychology*, Wiley Online Library, v. 44, n. 6, p. 1013–1023, 1988.

GOODALL, C. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 53, n. 2, p. 285–321, 1991.

GUPTA, A. K.; NAGAR, D. K. *Matrix variate distributions*. [S.l.]: Chapman and Hall/CRC, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009. v. 2.

HOFF, P. D. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 18, n. 2, p. 438–456, 2009.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. *Nonparametric statistical methods*. [S.l.]: John Wiley & Sons, 2013. v. 751.

HUANG, C.; STYNER, M.; ZHU, H. Clustering high-dimensional landmark-based two-dimensional shape data. *Journal of the American Statistical Association*, Taylor & Francis, v. 110, n. 511, p. 946–961, 2015.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020.

JAYASUMANA, S.; SALZMANN, M.; LI, H.; HARANDI, M. A framework for shape analysis via hilbert space embedding. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2013. p. 1249–1256.

JUPP, P. E.; MARDIA, K. V. Maximum likelihood estimators for the matrix von mises-fisher and bingham distributions. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 7, n. 3, p. 599–606, 1979.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 2009. v. 344.

KENT, J. T. The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 56, n. 2, p. 285–299, 1994.

KENT, J. T.; CONSTABLE, P. D.; ER, F. Simulation for the complex bingham distribution. *Statistics and Computing*, Springer, v. 14, n. 1, p. 53–57, 2004.

KHATRI, C.; MARDIA, K. V. The von mises–fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 95–106, 1977.

KOEV, P.; EDELMAN, A. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, v. 75, n. 254, p. 833–846, 2006.

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, Taylor & Francis, v. 47, n. 260, p. 583–621, 1952.

KUHN, M. Building predictive models in r using the caret package. *Journal of statistical software*, v. 28, p. 1–26, 2008.

KUNZE, K.; SCHAEBEN, H. Ideal patterns of crystallographic preferred orientation and their representation by the von mises-fisher matrix or bingham quaternion distribution. In: *Materials Science Forum*. [S.l.: s.n.], 2005. v. 495, p. 295–300.

LIMA, W. V. F. de; NASCIMENTO, A. D. C. do; AMARAL, G. J. A. do. Distance-based tests for planar shape. *Journal of Multivariate Analysis*, Elsevier, v. 184, p. 104743, 2021.

MACQUEEN, J. *Some methods for classification and analysis of multivariate observations*. [S.l.]: University of California Press, 1967.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, JSTOR, p. 50–60, 1947.

MARDIA, K. Statistical of directional data (with discussion). *Journal of the Royal Statistical Society*, v. 37, n. 3, p. 390, 1975.

MARDIA, K. Directional data analysis: an overview. *Journal of Applied Statistics*, Taylor & Francis, v. 15, n. 2, p. 115–122, 1988.

MARDIA, K. V. Statistics of directional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 37, n. 3, p. 349–371, 1975.

MARDIA, K. V.; JUPP, P. E.; MARDIA, K. *Directional statistics*. [S.l.]: Wiley Online Library, 2000. v. 2.

O'HIGGINS, P.; DRYDEN, I. L. Sexual dimorphism in hominoids: further studies of craniofacial shape differences in pan, gorilla and pongo. *Journal of Human Evolution*, Elsevier, v. 24, n. 3, p. 183–205, 1993.

PRENTICE, M. J. Orientation statistics without parametric assumptions. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 48, n. 2, p. 214–222, 1986.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>.

RODRÍGUEZ, W. D.; AMARAL, G. J.; NASCIMENTO, A. D.; FERREIRA, J. A. Information criteria in classification: new divergence-based classifiers. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 90, n. 17, p. 3200–3217, 2020.

ROYSTON, P. Remark as r94: A remark on algorithm as 181: The w-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, JSTOR, v. 44, n. 4, p. 547–551, 1995.

SHEN, W.; WANG, Y.; BAI, X.; WANG, H.; LATECKI, L. J. Shape clustering: Common structure discovery. *Pattern Recognition*, Elsevier, v. 46, n. 2, p. 539–550, 2013.

SOBHANI, F.; SALEMI, F.; MIRESMAEILI, A.; FARHADIAN, M. Morphometric analysis of the inter-mastoid triangle for sex determination: Application of statistical shape analysis. *Imaging Science in Dentistry*, v. 51, 2021.

SOUZA, S. d. A.; NASCIMENTO, A. D.; AMARAL, G. J. A. do. Estatística para dados direcionais. *Proceeding Series of the Brazilian Society of Computational and Applied Mathematics*, v. 8, n. 1, 2021.

SRIVASTAVA, A.; JOSHI, S. H.; MIO, W.; LIU, X. Statistical shape analysis: Clustering, learning, and testing. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 27, n. 4, p. 590–602, 2005.

TURAGA, P.; CHELLAPPA, R.; SRIVASTAVA, A. Statistical methods on special manifolds for image and video understanding. In: *Handbook of Statistics*. [S.l.]: Elsevier, 2013. v. 31, p. 178–201.

VINUÉ, G.; SIMÓ, A.; ALEMANY, S. The k-means algorithm for 3d shapes with an application to apparel design. *Advances in Data Analysis and Classification*, Springer, v. 10, n. 1, p. 103–132, 2016.

APÊNDICE - DEMONSTRAÇÕES DA CONSTRUÇÃO DAS DISTÂNCIAS

Sejam $\mathbf{X_1} \sim M(p, n, \mathbf{F_1})$, $\mathbf{X_2} \sim M(p, n, \mathbf{F_2})$ e suas funções densidades de probabilidades:

$$f_{\mathbf{X_j}}(\mathbf{X}; \mathbf{F_j}) = \exp\left\{\text{tr}(\mathbf{F_j}\mathbf{X}^\top) - \log\left[{}_0F_1\left(\mathbf{F_j}\right)\right]\right\}, \tag{24}$$

para $j \in \{1, 2\}$.

A seguir, mostramos a construção das distâncias de Rényi, Bhattacharyya e Hellinger. Com a finalidade de deixar nossas formulações menos densas, utilizaremos a seguinte notação:

$${}_0F_1\left(\mathbf{F}\right) = {}_0F_1\left(\frac{p}{2}; \frac{1}{4}\mathbf{F}\mathbf{F}^\top\right),$$

para $\mathbf{F}$ uma matriz $p \times n$ qualquer, considerando $\frac{p}{2}$ e $\frac{1}{4}$ constantes. Como descrito na página 44, nas Equações (2.2) e (2.3), temos

$$f_{\mathbf{X_j}}(\mathbf{X}; \mathbf{F_j}) = \exp\left\{\text{tr}\mathbf{F_j}\mathbf{X}^\top - \log\left[{}_0F_1\left(\mathbf{F_j}\right)\right]\right\} = \frac{\exp\{\text{tr}\mathbf{F_j}\mathbf{X}^\top\}}{{}_0F_1\left(\mathbf{F_j}\right)},$$

para $j \in \{1, 2\}$.

.1 DISTÂNCIA DE RÉNYI PARA DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL

Sejam $f_{\boldsymbol{X}_1}(\mathbf{X}; \boldsymbol{\theta}_1)$ e $f_{\boldsymbol{X}_2}(\mathbf{X}; \boldsymbol{\theta}_2)$ funções densidade de probabilidade de $\mathbf{X}_1$e $\mathbf{X}_2$, e $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$, seus respectivos parâmetros. Como descrito na página 47, Equações (2.12) e (2.13), a divergência de Rényi de ordem β é dada por

$$D_R^\beta(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2) = \frac{1}{\beta - 1}\log \mathbb{E}_{\boldsymbol{\theta}_1}\left[\left(\frac{f_{\boldsymbol{X}_1}(\mathbf{X}; \boldsymbol{\theta}_1)}{f_{\boldsymbol{X}_2}(\mathbf{X}; \boldsymbol{\theta}_2)}\right)^{\beta-1}\right], \tag{25}$$

e a distância de Rényi de ordem β, para $\beta \in (0, 1)$, é dada por

$$d_R^\beta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{\beta - 1}\log\left\{\frac{\exp[(\beta - 1)D_R^\beta(\boldsymbol{\theta}_1||\boldsymbol{\theta}_2)] + \exp[(\beta - 1)D_R^\beta(\boldsymbol{\theta}_2||\boldsymbol{\theta}_1)]}{2}\right\}. \tag{26}$$

Para descrevermos a distância de Rényi multidimensional, utilizaremos a função densidade de probabilidade da Von Mises-Fisher matricial.

Em (25) substituiremos os parâmetros $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_2$ por $\mathbf{F_1}$ e $\mathbf{F_2}$, repsectivamente. Além disso, faremos uso das propriedades de linearidade da função traço[3] tr (GUPTA; NAGAR, 2018), temos:

$$\begin{aligned}
D_R^{\beta}(\mathbf{F_1}||\mathbf{F_2}) &= \frac{1}{\beta-1}\log \mathbb{E}_{\boldsymbol{\theta}_1}\left[\left(\frac{\frac{\exp\{\text{tr}\mathbf{F_1}\mathbf{X}^\top\}}{{}_0F_1(\mathbf{F_1})}}{\frac{\exp\{\text{tr}\mathbf{F_2}\mathbf{X}^\top\}}{{}_0F_1(\mathbf{F_2})}}\right)^{\beta-1}\right] \\
&= \frac{1}{\beta-1}\log \mathbb{E}_{\boldsymbol{\theta}_1}\left[\left(\frac{\exp\{\text{tr}\mathbf{F_1}\mathbf{X}^\top\}}{{}_0F_1(\mathbf{F_1})}\frac{{}_0F_1(\mathbf{F_2})}{\exp\{\text{tr}\mathbf{F_2}\mathbf{X}^\top\}}\right)^{\beta-1}\right] \\
&= \frac{1}{\beta-1}\log\left\{\left(\frac{{}_0F_1(\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})}\right)^{\beta-1}\mathbb{E}_{\boldsymbol{\theta}_1}\left[\exp\{(\beta-1)(\text{tr}\mathbf{F_1}\mathbf{X}^\top-\text{tr}\mathbf{F_2}\mathbf{X}^\top)\}\right]\right\} \\
&= \frac{1}{\beta-1}\log\left\{\left(\frac{{}_0F_1(\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})}\right)^{\beta-1}\mathbb{E}_{\boldsymbol{\theta}_1}\left[\exp\{\text{tr}[(\beta-1)(\mathbf{F_1}-\mathbf{F_2})\mathbf{X}^\top]\}\right]\right\}
\end{aligned}$$

Para calcular $\mathbb{E}_{\boldsymbol{\theta}_1}[\cdot]$ foi utilizada a definição da função geradora de momentos (F.G.M) $\mathbb{M}_{\boldsymbol{\theta}_1}(\mathbf{T})$, detalhada em Khatri e Mardia (1977) e apresentada na Equação (2.11):

$$\mathbb{M}_{\boldsymbol{\theta}_1}(\mathbf{T}) = \mathbb{E}_{\boldsymbol{\theta}_1}[\exp\{\text{tr}\mathbf{T}\mathbf{X}^\top\}] = \frac{{}_0F_1(\mathbf{F_1}+\mathbf{T})}{{}_0F_1(\mathbf{F_1})},$$

em que $\mathbf{T}$ é uma matriz $n\times p$. Fazendo $\mathbf{T} = (\beta-1)(\mathbf{F_1}-\mathbf{F_2})$, temos:

$$\begin{aligned}
D_R^{\beta}(\mathbf{F_1}||\mathbf{F_2}) &= \frac{1}{\beta-1}\log\left\{\left(\frac{{}_0F_1(\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})}\right)^{\beta-1}\mathbb{M}_{\boldsymbol{\theta}_1}\left[(\beta-1)(\mathbf{F_1}-\mathbf{F_2})\right]\right\} \\
&= \frac{1}{\beta-1}\log\left\{\left(\frac{{}_0F_1(\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})}\right)^{\beta-1}\frac{{}_0F_1(\mathbf{F_1}+(\beta-1)(\mathbf{F_1}-\mathbf{F_2}))}{{}_0F_1(\mathbf{F_1})}\right\} \\
&= \frac{1}{\beta-1}\log\left\{\left(\frac{{}_0F_1(\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})}\right)^{\beta-1}\frac{{}_0F_1(\beta\mathbf{F_1}-(\beta-1)\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})}\right\} \\
&= \frac{1}{\beta-1}\log\left\{\frac{{}_0F_1(\mathbf{F_2})^{\beta-1}\,{}_0F_1(\beta\mathbf{F_1}-(\beta-1)\mathbf{F_2})}{{}_0F_1(\mathbf{F_1})^{\beta}}\right\}.
\end{aligned}$$

Analogamente, podemos obter

$$D_R^{\beta}(\mathbf{F_2}||\mathbf{F_1}) = \frac{1}{\beta-1}\log\left\{\frac{{}_0F_1(\mathbf{F_1})^{\beta-1}\,{}_0F_1(\beta\mathbf{F_2}-(\beta-1)\mathbf{F_1})}{{}_0F_1(\mathbf{F_2})^{\beta}}\right\}.$$

Substituindo $D_R^{\beta}(\mathbf{F_1}||\mathbf{F_2})$ e $D_R^{\beta}(\mathbf{F_2}||\mathbf{F_1})$ em (26), somos capazes de demonstrar que a distância $d_R^{\beta}(\mathbf{F_1},\mathbf{F_2})$ satisfaz a propriedade de definitude (ver página 23).

■

[3] Sejam $\mathbf{A}$ e $\mathbf{B}$ matrizes de mesma ordem. As propriedades de linearidade da função traço garantem que $\text{tr}(\mathbf{A}+\mathbf{B}) = \text{tr}(\mathbf{A})+\text{tr}(\mathbf{B})$ e $\text{tr}(c\mathbf{A}) = c\text{tr}(\mathbf{A})$, para $c\in\mathbb{R}$, constante.

Seja $\beta \in (0,1)$. Se $\mathbf{F_1} = \mathbf{F_2}$, então

$$\begin{aligned} d_R^{\beta}(\mathbf{F_1},\mathbf{F_2}) &= \frac{1}{\beta-1}\log\left\{\frac{1}{2}\left[\frac{{}_0F_1\left(\mathbf{F_1}\right)^{\beta-1}\,{}_0F_1\left(\beta\mathbf{F_1}-(\beta-1)\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_1}\right)^{\beta}}+\frac{{}_0F_1\left(\mathbf{F_1}\right)^{\beta-1}\,{}_0F_1\left(\beta\mathbf{F_1}-(\beta-1)\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_1}\right)^{\beta}}\right]\right\} \\ &= \frac{1}{\beta-1}\log\left\{\frac{1}{2}\left[\frac{{}_0F_1\left(\mathbf{F_1}\right)^{\beta}}{{}_0F_1\left(\mathbf{F_1}\right)^{\beta}}+\frac{{}_0F_1\left(\mathbf{F_1}\right)^{\beta}}{{}_0F_1\left(\mathbf{F_1}\right)^{\beta}}\right]\right\} \end{aligned}$$

■

DISTÂNCIA DE BHATTACHARYYA PARA DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL

Como pode ser visto na página 47, Equação (2.14), a distância de Bhattacharyya é definida por:

$$d_B(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = -\log \mathbb{E}_{\boldsymbol{\theta}_1}\left[\sqrt{\frac{f_{\boldsymbol{X}_2}(\mathbf{X};\boldsymbol{\theta}_2)}{f_{\boldsymbol{X}_1}(\mathbf{X};\boldsymbol{\theta}_1)}}\right], \tag{27}$$

em que $f_{\boldsymbol{X}_1}(\mathbf{X};\boldsymbol{\theta}_1)$ e $f_{\boldsymbol{X}_2}(\mathbf{X};\boldsymbol{\theta}_2)$ são as funções densidades de probabilidades da distribuições em comparação.

Agora, sejam $\mathbf{X_1} \sim M(p,n,\mathbf{F_1})$ e $\mathbf{X_2} \sim M(p,n,\mathbf{F_2})$, e, suas respectivas funções densidades de probabilidades $f_{\mathbf{X_1}}(\mathbf{X};\mathbf{F_1})$ e $f_{\mathbf{X_2}}(\mathbf{X};\mathbf{F_2})$. A distância de Bhattacharyya para distribuições Von Mises-Fisher matriciais é dada por:

$$\begin{aligned} d_B(\mathbf{F_1},\mathbf{F_2}) &= -\log \mathbb{E}_{\boldsymbol{\theta_1}}\left[\frac{f_{\mathbf{X_2}}(\mathbf{X};\mathbf{F_2})^{\frac{1}{2}}}{f_{\mathbf{X_1}}(\mathbf{X};\mathbf{F_1})^{\frac{1}{2}}}\right] \\ &= -\log \mathbb{E}_{\boldsymbol{\theta_1}}\left[f_{\mathbf{X_2}}(\mathbf{X};\mathbf{F_2})^{\frac{1}{2}} f_{\mathbf{X_1}}(\mathbf{X};\mathbf{F_1})^{-\frac{1}{2}}\right] \\ &= -\log \mathbb{E}_{\boldsymbol{\theta_1}}\left[\left(\frac{\exp\left\{\mathsf{tr}\mathbf{F_2}\mathbf{X}^{\top}\right\}}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}}\left(\frac{\exp\left\{\mathsf{tr}\mathbf{F_1}\mathbf{X}^{\top}\right\}}{{}_0F_1\left(\mathbf{F_1}\right)}\right)^{-\frac{1}{2}}\right] \\ &= -\log \mathbb{E}_{\boldsymbol{\theta_1}}\left[\exp\left\{\frac{1}{2}\mathsf{tr}\mathbf{F_2}\mathbf{X}^{\top}-\frac{1}{2}\mathsf{tr}\mathbf{F_1}\mathbf{X}^{\top}\right\}\left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}}\right] \\ &= -\log\left\{\left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}}\mathbb{E}_{\boldsymbol{\theta_1}}\left[\exp\left\{\mathsf{tr}\frac{1}{2}\mathbf{F_2}\mathbf{X}^{\top}-\mathsf{tr}\frac{1}{2}\mathbf{F_1}\mathbf{X}^{\top}\right\}\right]\right\} \\ &= -\log\left\{\left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}}\mathbb{E}_{\boldsymbol{\theta_1}}\left[\exp\left\{\mathsf{tr}\left(\frac{1}{2}(\mathbf{F_2}-\mathbf{F_1})\mathbf{X}^{\top}\right)\right\}\right]\right\} \end{aligned}$$

Utilizando a definição da F.G.M., $\mathbb{M}_{\boldsymbol{\theta}_1}(\mathbf{T}) = \mathbb{E}_{\boldsymbol{\theta}_1}[\exp\{\mathsf{tr}\mathbf{T}\mathbf{X}^{\top}\}] = \frac{{}_0F_1(\mathbf{F_1}+\mathbf{T})}{{}_0F_1(\mathbf{F_1})}$, apresentada

em (2.11), e fazendo $\mathbf{T} = \frac{1}{2}(\mathbf{F_2} - \mathbf{F_1})$, tem-se:

$$\begin{aligned} d_B(\mathbf{F_1}, \mathbf{F_2}) &= -\log\left\{\left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}} \mathbb{M}_{\boldsymbol{\theta}_1}\left(\frac{1}{2}(\mathbf{F_2} - \mathbf{F_1})\right)\right\} \\ &= -\log\left\{\left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}} \frac{{}_0F_1\left(\mathbf{F_1} + \frac{1}{2}(\mathbf{F_2} - \mathbf{F_1})\right)}{{}_0F_1\left(\mathbf{F_1}\right)}\right\} \\ &= -\log\left\{\frac{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1} + \mathbf{F_2})\right)}{\left[{}_0F_1\left(\mathbf{F_2}\right){}_0F_1\left(\mathbf{F_1}\right)\right]^{\frac{1}{2}}}\right\} \\ &= \log\left\{\frac{\left[{}_0F_1\left(\mathbf{F_2}\right){}_0F_1\left(\mathbf{F_1}\right)\right]^{\frac{1}{2}}}{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1} + \mathbf{F_2})\right)}\right\} \end{aligned}$$

■

Em particular, se $\mathbf{F_1} = \mathbf{F_2}$, teremos a propriedade da definitude satisfeita:

$$d_B(\mathbf{F_1}, \mathbf{F_2}) = \log\left\{\frac{\left[{}_0F_1\left(\mathbf{F_1}\right){}_0F_1\left(\mathbf{F_1}\right)\right]^{\frac{1}{2}}}{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1} + \mathbf{F_1})\right)}\right\} = \log\left\{\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_1}\right)}\right\}$$

■

DISTÂNCIA DE HELLINGER PARA DISTRIBUIÇÃO VON MISES-FISHER MATRICIAL

A distância de Hellinger, já apresentada na página 47, Equação (2.15), é definida por:

$$d_H(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 1 - \mathbb{E}_{\boldsymbol{\theta}_1}\left[\sqrt{\frac{f_{\boldsymbol{X}_2}(\mathbf{X}; \boldsymbol{\theta}_2)}{f_{\boldsymbol{X}_1}(\mathbf{X}; \boldsymbol{\theta}_1)}}\right]. \tag{28}$$

Sejam $f_{\mathbf{X_1}}(\mathbf{X}; \mathbf{F_1})$ e $f_{\mathbf{X_2}}(\mathbf{X}; \mathbf{F_2})$ as funções densidades de probabilidades de $\mathbf{X_1} \sim M(p, n, \mathbf{F_1})$ e $\mathbf{X_2} \sim M(p, n, \mathbf{F_2})$, respectivamente. A distância de Hellinger com base na distribuição de probabilidade Von Mises-Fisher matricial é dada por:

$$\begin{aligned} d_H(\mathbf{F_1}, \mathbf{F_2}) &= 1 - \mathbb{E}_{\boldsymbol{\theta_1}}\left[\sqrt{\frac{f_{\mathbf{X_2}}(\mathbf{X}; \mathbf{F_2})}{f_{\mathbf{X_1}}(\mathbf{X}; \mathbf{F_1})}}\right] \\ &= 1 - \mathbb{E}_{\boldsymbol{\theta_1}}\left[\sqrt{\frac{\exp\left\{\mathsf{tr}\mathbf{F_2}\mathbf{X}^\top\right\}}{{}_0F_1\left(\mathbf{F_2}\right)}\frac{{}_0F_1\left(\mathbf{F_1}\right)}{\exp\left\{\mathsf{tr}\mathbf{F_1}\mathbf{X}^\top\right\}}}\right] \\ &= 1 - \left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}} \mathbb{E}_{\boldsymbol{\theta_1}}\left[\exp\left\{\frac{1}{2}\mathsf{tr}\mathbf{F_2}\mathbf{X}^\top - \frac{1}{2}\mathsf{tr}\mathbf{F_1}\mathbf{X}^\top\right\}\right] \\ &= 1 - \left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}} \mathbb{E}_{\boldsymbol{\theta_1}}\left[\exp\left\{\mathsf{tr}\left(\frac{1}{2}(\mathbf{F_2} - \mathbf{F_1})\mathbf{X}^\top\right)\right\}\right]. \end{aligned}$$

Pela definição da F.G.M. apresentada na Equação (2.11), sabemos que $\mathbb{M}_{\boldsymbol{\theta}_1}\left(\frac{1}{2}(\mathbf{F_2}-\mathbf{F_1})\right) = \mathbb{E}_{\boldsymbol{\theta_1}}\left[\exp\left\{\text{tr}\left(\frac{1}{2}(\mathbf{F_2}-\mathbf{F_1})\mathbf{X}^\top\right)\right\}\right]$, e, que $\mathbb{M}_{\boldsymbol{\theta}_1}\left(\frac{1}{2}(\mathbf{F_2}-\mathbf{F_1})\right) = \frac{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1}+\mathbf{F_2})\right)}{{}_0F_1(\mathbf{F_1})}$. Portanto, substituindo $\mathbb{M}_{\boldsymbol{\theta}_1}\left(\frac{1}{2}(\mathbf{F_2}-\mathbf{F_1})\right)$ em $d_H(\mathbf{F_1},\mathbf{F_2})$, temos:

$$\begin{aligned} d_H(\mathbf{F_1},\mathbf{F_2}) &= 1-\left(\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_2}\right)}\right)^{\frac{1}{2}}\frac{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1}+\mathbf{F_2})\right)}{{}_0F_1\left(\mathbf{F_1}\right)} \\ &= 1-\frac{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1}+\mathbf{F_2})\right)}{{}_0F_1\left(\mathbf{F_2}\right)^{\frac{1}{2}}\,{}_0F_1\left(\mathbf{F_1}\right){}_0F_1\left(\mathbf{F_1}\right)^{-\frac{1}{2}}} \\ &= 1-\frac{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1}+\mathbf{F_2})\right)}{\left[{}_0F_1\left(\mathbf{F_2}\right){}_0F_1\left(\mathbf{F_1}\right)\right]^{\frac{1}{2}}} \end{aligned}$$

■

Quando $\mathbf{F_1}=\mathbf{F_2}$, teremos a propriedade da definitude satisfeita:

$$d_H(\mathbf{F_1},\mathbf{F_2}) = 1-\frac{{}_0F_1\left(\frac{1}{2}(\mathbf{F_1}+\mathbf{F_1})\right)}{\left[{}_0F_1\left(\mathbf{F_1}\right){}_0F_1\left(\mathbf{F_1}\right)\right]^{\frac{1}{2}}} = 1-\frac{{}_0F_1\left(\mathbf{F_1}\right)}{{}_0F_1\left(\mathbf{F_1}\right)}$$

■