Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística

FERNANDO LUIZ MAIA GOMES

MÉTODOS DE AGRUPAMENTO PARA FORMAS PLANAS

Recife

2020

FERNANDO LUIZ MAIA GOMES

MÉTODOS DE AGRUPAMENTO PARA FORMAS PLANAS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

Área de Concentração: Estatística aplicada.

Orientador: Prof. Dr. Getúlio José Amorim do Amaral
Coorientadora: Profa. Dra. Fernanda De Bastiani

Recife

2020

Catalogação na fonte
Bibliotecária Arabelly Ascoli CRB4-2068

G633m Gomes, Fernando Luiz Maia
Métodos de agrupamento para formas planas / Fernando Luiz Maia Gomes. – 2020.
65 f.: il., fig., tab.

Orientador: Getúlio José Amorim do Amaral
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN. Estatística. Recife, 2020.
Inclui referências.

1. Marcos anatômicos. 2. Espaços não-euclidianos. 3. Análise de agrupamento. 4. Fuzzy c-means. I. Amaral, Getúlio José Amorim do (orientador). II. Título.

310 CDD (22. ed.) UFPE-CCEN 2020-47

FERNANDO LUIZ MAIA GOMES

ALGUNS MÉTODOS DE AGRUPAMENTO PARA FORMAS PLANAS

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

Aprovada em: 11 de fevereiro de 2020.

BANCA EXAMINADORA

Prof.(º) Getúlio José Amorim do Amaral
UFPE

Prof.(ª) Audrey Helen Mariz de Aquino Cysneiros
UFPE

Prof.(º) Eufrásio de Andrade Lima Neto
UFPB

Dedico este trabalho à minha família, minha namorada, meus amigos e professores, por me ajudarem e me apoiarem em todos os momentos, além de serem compreensivos e me darem forças para continuar.

AGRADECIMENTOS

Agradeço aos meus pais, Eulália Maia e Vicente Gomes, por todo apoio que me deram. Mesmo longe, sempre sentia como se estivessem por perto, seja por ligações para saber como eu estava ou mensagens perguntando no que podiam ajudar.

Agradeço às minhas irmãs, Anna Cecília e Carolina Chaves, pelas palavras de sentimentos, carinhos e amor, sempre me incentivando a continuar e seguir em frente, nunca desistir.

Agradeço à minha namorada, Elaine, com todo amor e carinho, que mesmo morando distante sempre esteve presente na minha vida desde o começo do namoro e me ajudou várias vezes, lendo a dissertação comigo em busca de erros de português, me ouvindo explicar várias vezes sobre o trabalho aqui apresentado e estudando comigo as melhores palavras para se utilizar.

Agradeço a todos os meus professores, que me ajudaram não apenas no ensino escolar, mas também no ensino da vida como cidadão.

Sou grato também pela confiança depositada em mim e minha proposta de projeto pelos meus professores orientador Getúlio e minha coorientadora Fernanda, que me ajudaram no meu trabalho lado a lado algumas vezes, tentando superar as dificuldades em conjunto. Obrigado por me manterem motivado durante todo o processo.

Gostaria de agradecer também à CAPES pelo período de bolsa que recebi durante o curso para poder estudar com remuneração.

Por último, quero agradecer também à Universidade Federal de Pernambuco e todo o seu corpo docente.

RESUMO

A captura de imagens em duas e três dimensões tem demandado novas metodologias estatísticas para modelar esse tipo de dados. Nesse contexto, surge a morfometria, que permite a análise de imagens de objetos a partir de marcos anatômicos. Várias análises são de interesse no contexto de morfometria. Dentre estas análises, surge a análise de agrupamento que corresponde à obtenção de grupos que sejam internamente homogêneos e heterogêneos entre si. Deve-se destacar que o espaço onde são estudados os vetores que representam os objetos são não-euclideanos. Dessa forma, é necessário definir algoritmos de agrupamento com distâncias apropriadas. A distância geodésica, por exemplo, é uma boa alternativa. O presente trabalho considera dois algoritmos de análise de agrupamento, que são o k-medóide e o fuzzy c-means. Estes métodos são comparados ao algoritmo k-means que já é utilizado na literatura. Resultados numéricos, que são baseados no índice de Rand, indicam que o algoritmo fuzzy é uma boa opção dentre os três métodos considerados.

Palavras-chave: Marcos anatômicos. Espaços não-euclidianos. Análise de agrupamento. Fuzzy c-means.

ABSTRACT

Image capturing in two and three dimensions has been demanding new statistical methodologies to model this type of data. In this context, arises the morphometry, which allows the analysis of images of objects from anatomical landmarks. Several analysis are of interest in the context of morphometry. Among these analysis, arises the cluster analysis, which corresponds to obtaining groups that are internally homogeneous and heterogeneous among themselves. It should be noted that the space where the vectors representing the objects are studied are non-euclidean. Thus, it is necessary to define clustering algorithms with appropriate distances. Geodetic distance, for example, is a good alternative. The present work considers two cluster analysis algorithms, which are k-medoids and fuzzy c-means. These methods are compared to the k-means algorithm that is already used in the literature. Numerical results, which are based on the Rand index, indicate that the fuzzy algorithm is a good choice among the three methods considered.

Keywords: Landmarks. Non-euclidean spaces. Cluster analysis. Fuzzy c-means.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

Em toda parte do mundo existem objetos, sejam eles naturais ou feitos pelo homem. Imagens de objetos estão disponíveis em vários campos, como na engenharia, medicina, biologia, etc. É possível coletar dados geométricos rotineiramente destes objetos e as suas formas podem ser úteis para análises e tomadas de decisões importantes. Por exemplo, um médico pode diagnosticar um paciente com esquizofrenia ou não com base em sua ressonância magnética. Ou ainda, uma empresa pode identificar se há petróleo em certa área a partir de uma imagem de satélite. A análise estatística de forma pode ser usada para tomar decisões com base nessas informações obtidas. Portanto, o desenvolvimento de métodos de análises de tamanhos e formas de imagens é importante e necessário. Para melhorar o estudo da imagem uma análise estatística pode ser usada para tomadas de decisões.

Com os avanços na tecnologia, são cada vez mais fáceis as coletas de rotinas de informações geométricas e o estudo de formas de objeto é cada vez mais importante. As transformações mais comuns nos dados obtidos são em relação aos efeitos de locação, escala e rotação que são removidos do objeto. A análise estatística de formas é uma área da estatística que se propõe a estudar medidas de distâncias e comparações de formas de objetos, segundo Dryden e Mardia (1998). O primeiro trabalho nesta área foi feito por Kendall (1977), embora apenas no trabalho de Kendall (1984) é que a análise de formas foi formalizada e foram definidos os conceitos básicos da área. Um dos tópicos mais importantes no trabalho de Kendall foi a elaboração de um sistema de coordenadas que tinham a finalidade de obter a forma de um objeto através de pontos dispostos nele, chamados de marcos anatômicos. Essas coordenadas propostas estão dispostas em um espaço não-euclidiano e, portanto, são necessárias ferramentas adequadas para lidar com tais informações. A forma de um objeto é a informação restante quando os efeitos de locação, escala e rotação são removidos por meio de operações matemáticas. A análise estatística da forma é um tópico relativamente recente na estatística tendo como principais referências os trabalhos de Bookstein (1986) e Kendall (1984). Além disso, também existem livros didáticos, como os de Dryden e Mardia (2016) e Small (1996).

A maioria dos dados é armazenada digitalmente em mídia eletrônica, proporcionando um enorme potencial para o desenvolvimento de técnicas automáticas de análise, classificação e recuperação de dados. Além do crescimento da quantidade de dados, a variedade de dados disponíveis (texto, imagem e vídeo) também aumentou. Câmeras digitais e de vídeo baratas

disponibilizaram enormes arquivos de imagens e vídeos (Jain, 2010).

O desenvolvimento de técnicas de agrupamento tem sido um esforço verdadeiramente interdisciplinar. Taxonomistas, cientistas sociais, psicólogos, biólogos, estatísticos, matemáticos, engenheiros, cientistas da computação, pesquisadores médicos e outros que coletam e processam dados reais, todos contribuíram para a metodologia de agrupamento (Jain, 2009). Os algoritmos de agrupamento também foram extensivamente estudados em mineração de dados e aprendizado de máquina (Bishop, 2006).

O aumento no volume e na variedade de dados requer avanços na metodologia para entender, processar e resumir automaticamente os dados. As técnicas de análise de dados podem ser amplamente classificadas em dois tipos principais (Tukey, 1977):

1 Exploratório ou descritivo, o que significa que o pesquisador não possui modelos ou hipóteses pré-especificadas, mas deseja entender as características gerais ou a estrutura do dados dimensionais;
2 Confirmatórios ou inferenciais, significando que o investigador deseja confirmar a validade de uma hipótese / modelo ou de um conjunto de premissas, considerando os dados disponíveis.

Muitas técnicas estatísticas foram propostas para analisar os dados, como análise de variância, regressão linear, análise discriminante, análise de correlação canônica, escala multidimensional, análise fatorial, análise de componentes principais e análise de agrupamentos, entre outras (Jain, 2010).

Em geral, a análise de agrupamento refere-se a um amplo espectro de métodos que tentam subdividir um conjunto de dados X em subconjuntos c (grupos) que são separados por pares, todos não vazios, e sua união é o próprio X. Existem muitos algoritmos, cada um com seu próprio critério de agrupamento matemático para identificar grupos "ótimos", que são discutidos na literatura.

1.1 OBJETIVO

O objetivo dessa dissertação é fazer uma revisão literária dos métodos de agrupamentos existentes k-means e k-medóide além de contribuir para a literatura com o desenvolvimento do agrupamento c-means para formas planas.

1.2 ORGANIZAÇÃO DA DISSERTAÇÃO

As próximas etapas da dissertação estão organizadas da seguinte forma:

1.2.1 Capítulo 2

São apresentados alguns conceitos essenciais para o entedimento do trabalho como um todo. Esses conceitos são morfometria, formas, marcos anatômicos, matriz de Helmert, pré-formas e seu algoritmo de formação, assim como é demonstrado como gerar a distribuição Bingham complexa e seu algoritmo de formação outrossim como rotacionar os dados gerados pela distribuição.

1.2.2 Capítulo 3

Nesse capítulo são exploradas as várias interpretações do que é um grupo, suas razões e necessidade de formação e classificação. Também são apresentados alguns métodos de agrupamento, k-means, k-medóide e o c-means, bem como uma discussão sobre cada método e seus algoritmos de formação. Por último, é visto o índice de Rand, que serve para avaliar a qualidade do agrupamento obtido.

1.2.3 Capítulo 4

Neste capítulo visualizamos uma análise numérica dos métodos de agrupamentos, por meio de dados reais de crânios de gorilas e dados simulados com a distribuição Bingham.

1.2.4 Capítulo 5

Neste capítulo há o objeitvo de proporcionar uma visão geral do trabalho, suas contribuições e conclusões. Há também o objetivo de ilustrar o que ainda pode ser feito para que mais conclusões possam ser desenvolvidas a partir do que já foi demonstrado ao longo deste trabalho e ajudar a propor trabalhos futuros.

2 CONCEITOS BÁSICOS

2.1 MORFOMETRIA

Morfometria é o estudo matemático das formas de objetos pertencentes à mesma população estatística. Uma das suas aplicações é a identificação de populações de organismos vivos, que podem assumir formas ou tamanhos diferentes conforme o ambiente em que se desenvolveram.

A necessidade da utilização de conjuntos multivariados de caracteres no estudo de variação de organismos foi reconhecida há muito tempo por sistematas (a área da biologia dedicada a inventariar e descrever a biodiversidade e compreender as relações filogenéticas entre os organismos) e evolucionistas (é a mudança das características hereditárias de uma população de seres vivos de uma geração para outra) (Gould e Johnston, 1972). A maioria dos esforços iniciais para medir variação ficou, todavia, restrita a poucos caracteres, cujas interrelações nem sempre foram adequadamente avaliadas (Gould e Johnston, 1972). Todavia, o desenvolvimento de métodos de análise multivariada e, em particular, a disponibilidade de "pacotes“ estatísticos, permitiu que evolucionistas e sistematas se dedicassem a problemas relacionados à covariação de caracteres dentro e entre populações (Neff e Marcus, 1980).

A morfometria multivariada tem hoje diversas aplicações em biologia evolutiva. Habitualmente, a análise das funções discriminantes e a análise dos componentes principais são empregadas para detectar variação em caracteres quantitativos e, também, para avaliar padrões de relações fenéticas (Pimentel, 1979, Neff e Marcus, 1980). Estes estudos geralmente avaliam a variação morfométrica dentro das populações e sua relação com a variação entre as populações procurando, frequentemente, relacionar variação ambiental e diferenciação fenotípica. Na maioria dos casos, as relações entre populações são deduzidas a partir dos mesmos dados utilizados nas análises morfométricas, tornando impossível a avaliação independente das relações filogenéticas (Straney e Patton, 1980).

O tópico de morfometria baseado em marcos anatômicos foi introduzido pelos trabalhos de Kendall (1984), Kendall (1977) e Bookstain (1984). A formulação matemática foi apresentada por Kendall (1984) e (1977). Kendall trabalhava com dados de arqueologia e propôs uma metodologia baseada em um certo sistema de coordenadas. Por outro, lado Bookstein (1984) trabalhava com dados de biologia e propôs uma formulação baseada em um outro sistema de

coordenadas. Em seguida, esses dois sistemas foram avaliados como sendo semelhantes. Hoje em dia, os dois sistemas são corriqueiramente utilizados em morfometria.

2.2 FORMA E MARCOS ANATÔMICOS

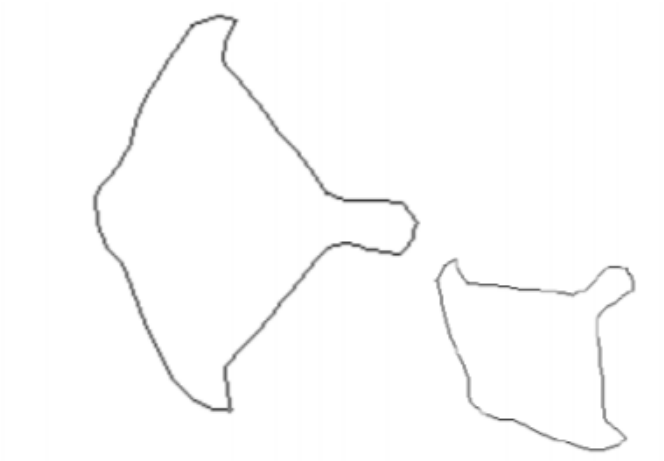

Figura 1 – Segunda vértebra torácica de dois ratos (osso T2), em diferentes locais, escalas e rotações, mas com a mesma forma (Dryden e Mardia (1998)).

A forma (*shape*) de um objeto é toda a informação geométrica que permanece quando retiramos os efeitos de locação, escala e rotação do objeto. Descrevemos a forma de um objeto a partir de pontos de referência (*landmarks* ou marcos anatômicos). A análise estatística de forma é um tópico relativamente recente em estatística e os conceitos fundamentais, como previamente citados, foram propostos por Bookstein (1986) e Kendall (1984). Além disso, existem alguns livros didáticos, tais como os de Dryden e Mardia (2016) e Small (1996).

De acordo com os autores Dryden e Mardia (2016), para uma melhor compreensão do assunto, faz-se necessario apresentarç alguns conceitos básicos:

- Forma (*Shape*) é toda a informação geométrica que permanece quando os efeitos de locação, escala e rotação são removidos de um objeto.
- Tamanho e forma (*Size* e *shape*) é toda a informação geométrica que permanece quando a locação e os efeitos de rotação são removidos de um objeto.
- Marco (*landmark*) é um ponto de correspondência em cada objeto que combina entre e dentro das populações.

Existem três tipos básicos de marcos anatômicos: os marcos anatômicos científicos, os marcos anatômicos matemáticos e os pseudo-marcos anatômicos, que consistem em:

- Um marco anatômico científico é um ponto atribuído por um especialista que corresponde

entre objetos de alguma maneira cientificamente significativa, por exemplo, o canto de um olho ou o encontro de duas suturas em um crânio. Em aplicações biológicas, esses marcos também são conhecidos como marcos anatômicos e designam partes de um organismo que correspondem em termos de derivação biológica, e essas partes são chamadas de homólogas (por exemplo, ver Jardine 1969).

- Marcos anatômicos matemáticos são pontos localizados em um objeto de acordo com alguma propriedade matemática ou geométrica da figura, por exemplo, em um ponto de alta curvatura ou em um ponto extremo. O uso de marcos matemáticos é particularmente útil no reconhecimento e análise automatizados.
- Pseudo-marcos anatômicos são pontos construídos em um objeto, localizados em torno do contorno ou entre marcos científicos ou matemáticos.

Bookstein (1991) também define os marcos anatômicos em três outros tipos, que são de uso particular na biologia. Marcos anatômicos tipo I ocorrem nas junções de tecidos/ossos; marcos anatômicos do tipo II são definidos por propriedades locais, como as curvaturas máximas, e os marcos anatômicos do tipo III ocorrem em pontos extremos ou marcos anatômicos construídos, como diâmetros máximos e centróides.

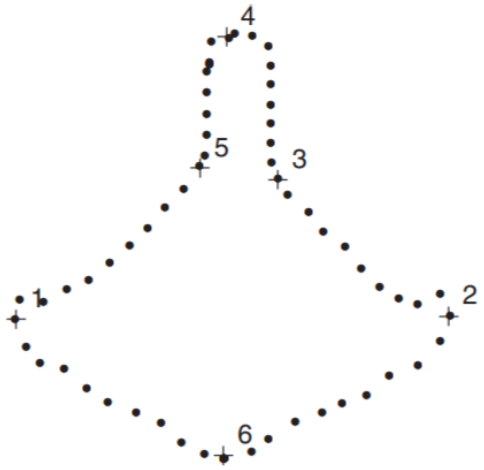

Figura 2 – Seis marcos anatômicos matemáticos (+) em uma segunda vértebra torácica de um rato, juntamente com 54 pseudo marcos anatômicos no contorno. Os pontos de referência são 1 e 2 que são pontos de máximos da função de curvatura (Dryden e Mardia (2016)).

2.3 MATRIZ DE HELMERT

A transformada de Helmert tem sido amplamente utilizada para realizar transformações entre sistemas de referência. Essa transformada permite que um conjunto de pontos em um

sistema seja transformado para outro, utilizando translações, rotações e escalas. A utilização de sistemas de referência é muito importante para qualquer tipo de posicionamento.

A transformação Helmert (em homenagem a Friedrich Robert Helmert , 1843-1917) é um método de transformação dentro de um espaço tridimensional. É frequentemente usado em geodésia para produzir transformações sem distorção de um dado para o outro. A transformação Helmert também é chamada de transformação de sete parâmetros e é uma transformação de similaridade .

Primeiro precisamos definir a submatriz Helmert que é usada para remover a locação. A submatriz de Helmert H tem dimensões $(k-1)\times k$ que é a matriz de Helmert sem a primeira linha. A matriz completa de Helmert H^F é uma matriz quadrada $k\times k$ com sua primeira linha de elementos igual a $1/\sqrt{k}$ e as linhas restantes são ortogonais a primeira linha. Retiramos a primeira linha de H^F para que a transformada HX não dependa da locação original da configuração. Vale notar também que: $H^T H = C$ em que C é matriz de centralização $C = I_k - (1/k)1_k 1_k^T$.

Por definição, a j-ésima linha da submatriz de Helmert é dada por:

$$(h_j \ldots h_j, -jh_j, 0, \ldots, 0), \quad h_j = -\{j(j+1)\}^{-1/2}, \tag{2.1}$$

então a j-ésima linha consiste em h_j repetido j vezes, seguido por $-jh_j$ e então $k-j-1$ zeros, $j = 1, \ldots, k-1$. Exemplo para $k = 3$, a matriz completa de Helmert é:

$$H^F = \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \end{pmatrix}$$

e a submatriz de Helmert é:

$$H = \begin{pmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} \end{pmatrix}.$$

2.4 PRÉ-FORMA

Seja X uma matrix de coordenadas cartesianas de l marcos anatômicos em m dimensões com configuração k x m dimensões, dada por:

$$X = \begin{pmatrix} X_{1,1} & \dots & X_{1,m} \\ X_{21} & \dots & X_{2,m} \\ \vdots & & \vdots \\ X_{k,1} & \dots & X_{k,m} \end{pmatrix}.$$

As coordenadas cartesianas de cada marco anatômico são representadas no espaço real R^m.

Para eliminar os efeitos de locação, escala e rotação do objeto é necessário fazer algumas transformações na matriz X. Inicialmente, escrevemos a matriz como um vetor complexo com dimensão m=2:

$$z^0 = (y_{1,1} + i * y_{1,2}, \dots y_{k,1} + i * y_{k,2})^T = (z^0_{(1)}, \dots, z^0_{(k)})^T \tag{2.2}$$

Em seguida, o efeito da locação pode ser removido multiplicando z^0, equação (2.2), pela submatriz de Helmert H, equação (2.1). Dessa forma:

$$w_{(k-1\text{x}1)} = H_{(k-1\text{x}k)} z^0_{(k\text{x}1)}, \tag{2.3}$$

em que w representa a configuração z_0 sem o efeito de locação, que é denominada de configuração helmertizada.

Pré-multiplicando o vetor w por H^T podemos obter a configuração centrada:

$$H^T w = H^T H z^0 = C * z^0.$$

Para remover o efeito de escala deve-se dividir a configuração helmertizada, equação (2.3), pela sua norma $|w|$, assim teremos:

$$z_{(k-1\times 1)} = \frac{w}{|w|} = \frac{w}{\sqrt{w^* w}}, \tag{2.4}$$

em que $w*$ é o transposto conjugado de w e $|.|$ é a norma complexa de w.

Algoritmo 1: Algoritmo para calcular a pré-forma com m=2

1: Transforme a matriz X em um vetor complexo $(k \times 1)$, dessa forma obtendo z^0, equação (2.2).

2: Para retirar o efeito de locação multiplique z^0 pela submatriz de Helmert H, tendo assim w, equação (2.3).

3: Remova o efeito de escala dividindo w pela sua norma $|w|$, equação (2.4). Assim teremos o pré-shape z.

Assim, a pré-forma da matriz de configuração X com m dimensões é dada por:

$$z_{(k-1\times m)} = \frac{H_{(k-1\times k)}X_{(k\times m)}}{|HX|}, \tag{2.5}$$

A qual é invariante sob locação e escala da configuração inicial.

Note que, pela equação (2.5), podemos obter as pré-formas centralizadas, basta multiplicarmos o numerador e o denominador por H^T, pois dessa forma teremos $C = H^T H$.

$$z_{c(k\times m)} = C_{(k\times k)}X_{k\times m}/|CX|.$$

Nesse caso, z_c é uma matriz $k \times m$, diferente de z, que é $(k-1) \times m$. De acordo com Dryden e Mardia (1998) a vantagem de usar z é que este é de posto completo e sua dimensão é menor que z_c. Entretando, utilizando z_c, a pré-forma centralizada, é que a representação das coordenadas cartesianas serão coerentes com a configuração original.

Os espaços das pré-formas são todos os espaços possíveis das pré-formas z, que são os $(k-1)$ vetores, sem os efeitos de locação e escala. Para formas planas $m = 2$, esse espaço é uma hiperesfera complexa de dimensão $(k-1)$, ou seja:

$$CS^{k-2} = \{z : z^*z = 1, z \equiv C^{k-1}\}, \tag{2.6}$$

em que C^{k-1} é o espaço complexo de dimensão $(k-1)$.

Um importante conceito de morfometria é o conceito de forma média. Considere $z_1^0, \ldots, z_n^0$ uma amostra aleatória de configuração complexa de uma população com n objetos em que z_i^0 foi definido na equação (2). A forma média do grupo $\hat{\mu}$ pode ser encontrada como

autovetor dominante da soma quadrática complexa e matriz produto

$$S = \sum_{i=1}^{n} z_i z_i^*, \tag{2.7}$$

em que $z_i = w_i/||w_i||,\ i = 1,\ldots,n$ são as pré-formas, então $\hat{\mu}$ é o autovetor correspondente ao maior autolavor de S.

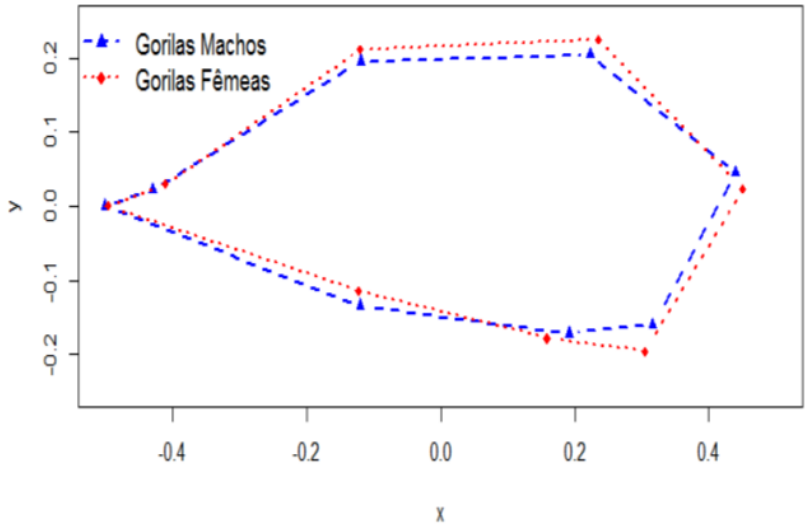

Figura 3 – A forma média procrustes completa de crânios de gorilas machos e fêmeas Dryden and Mardia(2016).

Algoritmo 2: Algoritmo para calcular a forma média

1: Calcule as pré-formas $z_1,\ldots,z_n$ pelo algoritmo (1).

2: Calcule a soma quadrática e matriz produto S pela equação (2.7) .

3: A forma média será o autovetor associado ao maior autovalor de S.

2.5 DISTÂNCIAS

Considere duas matrizes com k pontos e dimensão $m = 2$, em que $z_x = (z_{x1},\ldots,z_{xk})^T$ e $z_y = (z_{y1},\ldots,z_{yk})^T$ são as pré-formas de X e Y com configuração $||z_x|| = 1$ e $||z_y|| = 1$, com $z_x * 1_k = 0$ e $z_y * 1_k = 0$. Sendo assim, a distância entre as duas forma dos dois grupos z_x e z_y é:

$$d_F^2 = 1 - |z_x * z_y|^2.$$

Esta distância é invariante aos efeitos de locação, escala e rotação. Sendo assim, podemos considerar $cos\rho = (1 - d_F^2)^{1/2}$.

Para os dados definidos na esfera complexa (equação (2.6)) o ângulo entre as pré-formas complexas z_x e z_y é

$$\rho = arccos(|z_x * z_y|), \tag{2.8}$$

essa distância também é conhecida como distância geodésica. Ela é o caminho mais curto entre z_x e z_y na hiperesfera da pré-forma e esta não é afetada pela rotação. Dessa forma, fica visível que a distância procruste ρ é o ângulo entre as pré-formas z_x e z_y.

A distância procruste parcial também é invariante quanto à rotação entre z_x e z_y e ela é dada por

$$d_P^2 = 2(1 - |z_x * z_y|) = 2(1 - cos\rho).$$

Outro espaço comumente usado na análise de formas é o espaço tangente, que representa uma versão linearizada do espaço de formas. Através do uso desse conceito por construção, ocorre uma certa perda de informações (contida na curva do espaço original da forma). A vantagem é que a métrica euclidiana do espaço tangente permite o uso de análises multivariadas padrão (Amaral et al., 2010).

2.6 A DISTRIBUIÇÃO BINGHAM COMPLEXA

A distribuição Bingham complexa foi introduzida por Kent (1994) como um modelo tratável para análise de formas baseado em marcos, a qual é definida na esfera complexa unitária.

A distribuição Bingham complexa é relevante para a análise da forma de dados em duas dimensões. O problema de simular dados da distribuição Bingham equivale a simulação de uma distribuição exponencial multivariada truncada (Kent, 2004). A distribuição de Bingham possui simetria antipodal, o que significa que $f(x) = f(-x)$ para todos os vetores unitários x. Isso torna a distribuição Bingham adequada para modelar eixos na esfera em que os vetores x e $-x$ representam o mesmo eixo (Dore 2016).

Densidade: Considere o caso em que temos uma distribuição de probabilidade de pré-formas na esfera $S^{m(k-1)-1}$ correspondente a k pontos em m dimensões. Para o espaço dimensional $m = 2$, utilizando a notação complexa temos que $S^{2(k-1)-1} \equiv CS^{k-2}$ em que $CS^{k-2} =$

$[z : (k-1) - vetores, z^*z = 1]$ é a esfera unitária complexa com $k-1$ dimensões complexas. No nosso caso, considere k marcos anatômicos e $m = 2$ dimensões com coordenadas.

Para $k \leq 2$ seja $CS^{k-1} = \{z = (z_1, z_1, ..., z_k) : \sum |z_i|^2 = 1\} \subset C^k$. A distribuição Bingham complexa tem função densidade de probabilidade:

$$f(z) = c(A)^{-1} \exp(z * Az),$$

em que A é uma matriz hermetiana $k \times k$ e $c(A)$ é uma constante de normalização, dada por:

$$c(A) = 2\pi^{k-1} \sum_{i=1}^{k-1} a_i \exp\lambda_i, \quad a_i^{-1} = \prod_{i \neq j} (\lambda_j - \lambda_i)$$

em que $\lambda_1 < \lambda_2 < ... < \lambda_{k-1} = 0$ denota os autovalores de A. Note que $c(A) = c(\wedge)$ depende somente dos autovalores de A, com $\wedge = \text{diag}(\lambda_1, ..., \lambda_{k-1})$.

Para simular a distribuição Bingham complexa, vários métodos foram propostos por Kent et al. (2004). No trabalho será adotado o primeiro método, truncação pelo simplex. Primeiro, gera-se exponenciais truncadas, $\text{Texp}(\lambda_j)$, pelo método de aceitação e rejeição e, então, essas variáveis aleatórias são expressas em coordenadas polares para obter uma distribuição Bingham complexa.

Algoritmo 3: Algoritmo para gerar a exponencial truncada

1: Gere uma variável aleatória $U_j \sim U[0,1], \quad j = 1, ..., k-1$

2: Seja $S'_j = -(1/\lambda_j)\log(1 - U_j(1 - e^{-\lambda_j})), S' = (S'_1, ..., S'_{k-1})^T$ de modo que S'_j são amostras aleatórias independentes $\text{Texp}(\lambda_j)$

3: Se $\sum_{j=1}^{k-1} S'_j < 1$, então $S = S'$, senão rejeite S' e retorne ao passo 1.

Para simular a Bingham complexa usa-se $k-1$ exponenciais truncadas gerando um vetor de dimensão k de uma distribuição Bingham complexa. É necessário fazer uma mudança nos autovalores de A, de modo que eles são não positivos, com o maior deles igual a 0. A mudança é feita sem perda de generalidade (Kent, 2004). A mudança deve ser feita de forma que $\lambda_1 \geq \lambda_2 \geq ...\lambda_k = 0$ denote os autovalores de $-A$. Então, escreva $\lambda = (\lambda_1, ..., \lambda_{k-1})$ para o vetor dos primeiros $k-1$ autovalores. Os parâmetros λ_j são considerados como parâmetros de

concentração.

Algoritmo 4: Gerando a distribuição Bingham complexa

1: Tendo $S = (S_1, ..., S_{k-1})$, onde $S \sim \text{Texp}(\lambda_j)$ são exponenciais truncadas e independentes.

2: Se $\sum_{j=1}^{k-1} S'_j < 1$, faça $S_k = 1 - \sum_{j=1}^{k-1} S'_j$, senão volte ao passo 1.

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

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

2.6.1 Rotacionando a Distribuição Bingham

No trabalho serão simuladas duas amostras de dados da distribuição Bingham. Para fazer o agrupamento de forma correta, uma das amostras precisa ser rotacionada. Esta seção explica como o método proposto por Amaral, Dryden e Wood (2007) pode ser usado para rotacionar uma das amostras. Deve-se notar que com dados simulados as duas amostras têm mesma rotação, o que implica a mesma forma média, por isso a necessidade de rotacionar uma das amostras.

De acordo com Amaral, Dryden e Wood (2007), seja m o parâmetro de interesse, assumido como um vetor unitário real ou complexo, e denota a estimativa de m com base na amostra i por m_i, $i = 1, ..., k$. Deixe m_0 denotar uma estimativa combinada de m com base nas k amostras. Então, no caso direcional (ou seja, real), para cada i, encontramos uma rotação R_i que gira m_i em direção a m_0 de modo que todos os vetores ortogonais ao plano formado por m_i e m_0 permanecem inalterados, e aplicamos essa transformação a cada observação na amostra, obtendo assim uma amostra transformada ou rotacionada. Na realidade, usamos rotações que alteram as amostras "o mínimo possível“ sujeitas a mover m_i para m_0. Os detalhes são semelhantes no caso de pré-forma (ou seja, no caso complexo), exceto que as transformações apropriadas U_i são unitárias. As opções apropriadas de R_i e U_i são dadas abaixo.

Para o caso real, suponha que a e b são vetores unitários em R^d e que desejamos "mover b para a pelo caminho geodésico na esfera unitária em R^d que liga b a a". Desde que $|a^T b| < 1$, a matriz de rotação é determinada de maneira natural.

Defina $c = \{b - a(a^T b)\}/||b - a(a^T b)||$, onde $||.||$ denota a norma euclidiana em R^d. Porque

$$||b - a(a^T b)||^2 = b^T b + a^T a(a^T b)^2 - (a^T b)^2 - (a^T b)(b^T a) = 1 - (a^T b)^2 > 0$$

quando $|a^T b| < 1$, segue-se que c está bem definido. Observe que, por construção, $c^T c = 1$. Defina $\alpha = \cos^{-1}(a^T b) \in (0, \pi)$ e $A = ac^T - ca^T$.

Suponha que a, $b \in R^d$ sejam vetores unitários tais que $|a^T b| < 1$, e que α, A, e c sejam definidos como anteriormente. Então a matriz

$$Q = \exp(\alpha A) = I_d + \sum_{j=1}^{\infty} \frac{\alpha^j}{j!} A^j,$$

possui as seguintes propriedades: Q é uma matriz de rotação $d \times d$; Q pode ser escrito como $Q = I_d + (\text{sen}\alpha)A + \{(\cos\alpha) - 1\})(aa^T + cc^T)$; $Qb = a$ e para qualquer $z \in R^d$ tal que $a^T z = 0$ e $b^T z = 0$, nós temos $Qz = z$.

Ao aplicar esse resultado à transformação da amostra i, consideramos $b = m_i$, $a = m_0$ e R_i como a matriz de rotação Q.

No caso complexo os detalhes são bastante semelhantes ao caso real, mas com um recurso adicional. No contexto da análise de forma, queremos escolher uma pré-forma $\tilde{b}$ a partir da forma $[b]$ de b, onde $b \in C^d$, $b * b = 1$, de modo que $\tilde{b}$ se move ao longo de uma geodésica "horizontal" no espaço pré-moldado, correspondendo a uma geodésica no espaço da forma para mais detalhes ver Kendall et al. (1999). Na prática, fazemos isso substituindo b por

$$\tilde{b} = b(b * a)/|b * a|, \tag{2.9}$$

, de modo que $\tilde{b}$ tem a mesma norma que b e $\tilde{b} * a$ é real.

Dado que trabalhamos com $\tilde{b}$ em vez de b, os detalhes são análogos ao caso real. Nesse caso, definimos

$$\tilde{c} = \frac{\tilde{b} - a(a * \tilde{b})}{||\tilde{b} - a(a * \tilde{b})||},$$

$$\tilde{\alpha} = \cos^{-1}(a * \tilde{b}) \in (0, \pi)$$

e

$$\tilde{A} = a\tilde{c} * -\tilde{c}a * .$$

Agora suponha que a, $b \in C^d$ satisfaça $||a^2|| = a * a = 1$ e $||b^2|| = b * b = 1$ e suponha que $|a * b| < 1$. Seja $\tilde{\alpha}, \tilde{A}$ e $\tilde{c}$ como foi definida acima. Então a matriz

$$U = \exp(\tilde{\alpha}\tilde{A}) = I_d + \sum_{j=1}^{\infty} \frac{\tilde{\alpha}^j}{j!} \tilde{A}^j,$$

possui as seguintes propriedades: U é uma matriz unitária $d \times d$; U pode ser escrita como $U = I_d + (\text{sen}\tilde{\alpha})\tilde{A} + \{(\cos\tilde{\alpha}) - 1\}(aa* + \tilde{c}\tilde{c}*)$; $U\tilde{b} = a$ e para qualquer $z \in C^d$ tal que $a*z = 0$ e $b*z = 0$, temos $U_z = z$.

Vale notar que o conjunto $\{\tilde{x}(\theta) = \exp(\theta\tilde{A})\tilde{b} : \theta[0, \tilde{\alpha}]\}$ é uma geodésica "horizontal" na esfera de pré-forma e, portanto, corresponde a uma geodésica no espaço da forma (ver Kendall et al. 1999). Ao aplicar esse resultado na prática, tomamos $b = m_i$ e $a = m_0$, definimos $\tilde{b}$ usando (2.9) e assumimos U_i como a matriz unitária.

3 MÉTODOS DE AGRUPAMENTO

3.1 INTRODUÇÃO

Uma das habilidades mais antigas e básicas do seres vivos está entrelaçada com o agrupamento. Com o agrupamento de objetos semelhantes é possivel produzir uma classificação. Essa ideia de categorias vem desde os primórdios onde o homem pré-historico conseguiu perceber que objetos individuais compartilham mesmas propriedades. Por exemplo, alguns objetos são comestíveis, venenosos, etc.

A própria classificação no seu sentido mais abrangente é necessária para a formação das palavras que nos ajudam a reconhecer e discutir os variados tipos de eventos, objetos e pessoas com que nos deparamos. Por exemplo, é preciso um rótulo para descrever uma classe de coisas que tenham características marcantes em comum. Dessa forma os animais são nomeados como gatos, cães, cavalos, etc e esse nome classifica os indivíduos em grupos. Classificar e nomear são em sua essencia sinônimos.

Além de ser um conceito básico para os humanos, a classifição é fundamental para a ciência. Na biologia, por exemplo, a classificação de seres vivos tem sido uma preocupação desde o início dos tempos. Aristóteles, por exemplo, construiu um sistema complexo para classificar as espécies do reino animal, em que começava dividindo os animais em dois grupos primários: os que tinham sangue vermelho (correspondendo aproximadamente aos nossos próprios vertebrados) e os que não os possuíam (os invertebrados). Ele subdividiu ainda mais esses dois grupos de acordo com a maneira como os filhotes são produzidos, tais como: mamiferos, em ovos, como pupas e assim por diante.

No ano de 1737, Carl von Linn publicou uma obra chamada Genera Plantarum, da qual foram tiradas as seguintes citações:

“Todo o conhecimento real que possuímos depende de métodos pelos quais distinguimos o similar do diferente. Quanto maior o número de distinções naturais que esse método compreende, mais clara se torna a nossa ideia das coisas. Quanto mais numerosos os objetos que empregam nossa atenção, mais difícil se torna formar tal método e mais necessário."

Na área biológica existe todo um estudo voltado para classificar organismos, esse estudo é chamado de taxonomia. No início da taxanomia essa era uma prática mais ampla, talvez mais uma arte do que um método científico. Porém, com o passar do tempo, novas técnicas

menos subjetivas foram desenvolvidas.

A classificação de animais e plantas claramente desempenhou um papel importante nos campos da biologia e da zoologia, como por exemplo, na base da teoria da evolução de Darwin, mas esta também desempenhou um papel muito importante no desenvolvimento de teorias em outros campos da ciência. Como, por exemplo, a classificação dos elementos na tabela periódica, produzida por Mendeleyev no ano de 1860, que teve um impacto profundo no entendimento e compreensão da estrutura do átomo. Novamente, em astronomia, a classificação de estrelas em estrelas anãs e estrelas gigantes, usando o gráfico de temperatura Hertzsprung-Russell contra a luminosidade (Figura 4) afetou fortemente as teorias da evolução estelar.

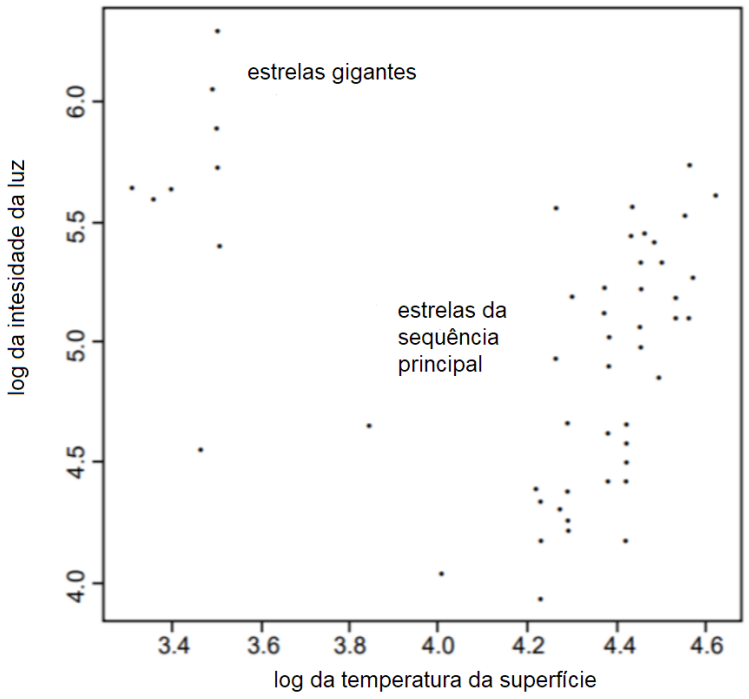

Figura 4 – Gráfico de temperatura de Hertzsprung – Russell contra a luminosidade (veja Everitt Et. Al. (2011))

A classificação pode envolver pessoas, animais, elementos químicos, estrelas, etc., como entidades a serem agrupadas. Neste texto, geralmente usaremos o termo objeto para cobrir todas essas possibilidades.

3.2 RAZÕES PARA A CLASSIFICAÇÃO

De forma mais simples, um esquema de classificação pode ilustrar um método conveniente para organizar grandes conjuntos de dados para que estes possam ser entendidos

mais facilmente e suas informações possam ser extraídas ou recuperadas de forma mais eficiente. Se um pequeno número de objetos puderem resumir de forma coesa os dados, então os rótulos dos grupos poderão fornecer uma descrição muito concisa dos padrões de semelhanças e diferenças nos dados. Em pesquisa de mercado, por exemplo, é possivel facilitar a pesquisa agrupando um grande número de pessoas de acordo com suas preferências para produtos específicos. Isso pode ajudar a identificar um “produto de nicho"para um tipo específico de consumidor. Cada vez mais é importante a necessidade de resumir conjuntos de dados dessa maneira devido ao crescente número de grandes bancos de dados atualmente disponíveis em praticamente todas as áreas da ciência. Com a exploração correta desses bancos de dados usando análise de agrupamento e outras técnicas de análise multivariada é chamada agora de mineração de dados. No século 21 a mineração de dados ganhou particular interesse para investigação de material na *World Wide Web* onde o objetivo é extrair informações ou conhecimentos úteis do conteúdo da página da internet, (para mais detalhes ver Liu (2007)).

Entretanto, nas suas muitas aplicações, um pesquisador pode estar procurando uma classificação que lhe forneça um resumo útil dos dados e também sirva para outros propósitos mais fundamentais. Na medicina, por exemplo, para tratar e entender uma doença, esta deve ser classificada e de forma geral essa classificação deve ter dois objetivos principais: o primeiro será a previsão - separando doenças que requerem tratamentos diferentes; o segundo será fornecer uma base para a pesquisa em etiologia - as causas de diferentes tipos de doenças. São esses dois principais objetivos que um clínico tem em mente quando faz um diagnóstico.

Quase sempre existe uma grande variedade de classificações alternativas para um mesmo conjunto de objetos. Por exemplo, os seres humanos podem ser classificados em relação à sua renda como classe baixa, média e alta; também podem ser classificados pelo seu consumo anual de álcool em baixo, médio e alto; etc. Obviamente essas diferentes classificações irão reunir diferentes indivíduos em grupos. No entanto algumas classificações são mais prováveis de serem de uso geral do que outras, um argumento bem formulado por Needham (1965) ao discutir a classificação de humanos em homens e mulheres:

“A utilidade dessa classificação não começa e termina com tudo o que, em certo sentido, pode ser estritamente deduzido dela - a saber, uma afirmação sobre o gênero. É uma classificação muito útil porque classificar uma pessoa como homem ou mulher transmite muito mais informações sobre provável tamanho relativo, força, certos tipos de destreza e assim por diante. Quando dizemos que pessoas da classe homem são mais adequadas do que pessoas da classe mulher para determinadas tarefas e, inversamente, estamos apenas fazendo um comentário

incidental sobre sexo, nossa principal preocupação é com força, resistência, etc. O ponto é que fomos capazes de usar uma classificação de pessoas que transmite informações sobre muitas propriedades. Pelo contrário, uma classificação de pessoas com tamanho dos antebraços entre 10cm e 20cm de comprimento e aquelas que não estão nesse intervalo, embora possa servir para algum uso específico, certamente não é de uso geral. Em outras palavras, não existem propriedades conhecidas que dividam um conjunto de pessoas de maneira semelhante."

Argumentos semelhantes podem ser feito com relação a classificação de livros baseando-se, por exemplo, no assunto ou na cor do livro. Claramente à classificação do livro com base no assunto, com classes como dicionário, romances, biografias, etc, será de uso muito mais positivo que o segundo, com base na cor do livro como verde, azul, vermelho, etc. É clara a razão pela qual a primeira classificação é mais util que a segunda. A classificação pelo assunto indica muito mais as características do livro que a classificação pela cor.

Portanto, deve-se ressaltar e lembrar que, de forma geral, uma classificação de um conjunto de objetos não é como uma teoria científica e talvez deva ser julgada amplamente pela sua utilidade, e não em termos binários “verdadeiro"ou “falso".

3.3 MÉTODOS DE PARTIÇÃO PARA FORMAS PLANAS

As técnicas numéricas para derivar classificações originaram-se de forma geral nas ciências naturais, como biologia e zoologia, em um grande esforço para livrar a taxonomia de sua natureza tradicionalmente subjetiva e torná-la mais objetiva. O foco era fornecer classificações objetivas e estáveis. Objetivo no sentido de que a análise do mesmo conjunto de organismos pela mesma sequência de métodos numéricos produza uma mesma classificação; de forma estável na medida em que a classificação permanece a mesma em uma ampla variedade de adições de organismos ou de novas características que os descrevem.

Um grande número de nomes foi utilizado para esses métodos numéricos de acordo com a área de aplicação considerada. A *taxonomia numérica* é geralmente usada em biologia. Em psicologia, o termo *análise Q* é empregado algumas vezes. Na literatura sobre inteligência artificial, o *reconhecimento não supervisionado de padrões* é o rótulo preferido, e os pesquisadores de mercado costumam falar sobre *segmentação*. Atualmente, porém, a análise de agrupamento é provavelmente o termo genérico preferido para procedimentos que buscam descobrir grupos de dados.

Na maioria das aplicações de análise de agrupamento é procurado dividir os objetos

em partições dos dados na qual cada indivíduo ou objeto pertence a um único grupo e o conjunto completo dos grupos contém todos os indivíduos. Em algumas circunstâncias, no entanto, os grupos sobrepostos podem fornecer uma solução mais aceitável.

A entrada de dados básica para a maioria das aplicações em análise de grupo é a matriz X de dados multivariada $n \times p$, contendo os valores das variáveis que descrevem cada objeto a ser agrupado; isso é:

$$X = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ x_{21} & \dots & x_{2,p} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix}$$

Os elementos de entrada x_{ij} em X fornecem o valor da j-ésima variável no objeto i.

As variáveis na matrix X podem ser frequentemente uma mistura de dados contínua e/ou discreta e algumas vezes pode ocorrer nenhuma entrada. As variáveis mistas e os valores ausentes complicam o agrupamento dos dados e em algumas aplicações as linhas da matriz X podem conter dados repetidos da mesma variável como, por exemplo, em condições diferentes, momentos diferentes, ou em posições espaciais, etc. Um exemplo simples de dado repetido com relação ao tempo pode ser a altura de crianças todos os meses com o passar de alguns anos. Esses dados estruturados são de natureza especial, pois todas as variáveis são medidas na mesma escala, e a análise de grupos de dados estruturados pode exigir abordagens diferentes do agrupamento de dados não estruturados.

A análise de agrupamento é essencial para a descoberta e agrupamento de dados. Os métodos de agrupamento não devem ser confundidos com os métodos discriminantes e atribuição (que no contexto da inteligência artificial é usado o termo aprendizado supervisionado), onde os grupos são conhecidos a priori e o objetivo da análise é construir regras para classificar novos indivíduos em um ou outro dos grupos conhecidos. Um texto didático sobre tais métodos é apresentado por Hand (1981).

3.4 O QUE É UM GRUPO?

Até esse momento, os termos de grupos, grupo e classe foram usados de maneira totalmente intuitiva, sem sua devida definição específica. A definição formal não é apenas difícil,

trabalhosa, mas pode até ser equivocada. Bonner (1964), sugeriu que o critério final para avaliar o significado de tais termos é o julgamento do valor do usuário. Se o uso de um termo como “grupo"produz uma resposta de valor para o investigador, isso é tudo o que é necessário.

Em parte Bonner está certo, mas seu argumento não é inteiramente convincente, e muitos autores, por exemplo, Cormack (1971) e Gordon (1999), tentam definir exatamente o que é um grupo em termos de coesão interna - homogeneidade - e isolamento externo - separação. Tais propriedades pelo menos podem ser ilustradas informalmente, como na Figura 5. Os “grupos"presentes nesta figura serão claros aos observadores sem a necessidade de uma definição formal explícita do termo. A valer o exemplo demonstra que não há nenhuma definição única e que provavelmente não será suficiente para todas as situações. Isso pode explicar por que as tentativas de tornar os conceitos de homogeneidade e separação matematicamente precisos em termos de índices numéricos explícitos levaram a numerosos e diversos critérios.

Figura 5 – Gráfico mostrando a seleção de grupos intuitivamente (veja Gordon (1980)).

Não está totalmente claro como um “grupo"é reconhecido quando exibido no plano, mas um recurso do processo de reconhecimento parece envolver a avaliação das distâncias relativas entre os pontos. Como os observadores humanos extraem grupos perceptivamente coerentes dos campos de “pontos"será considerado brevemente no trabalho.

Um outro conjunto de dados bidimensionais é plotado na Figura 6. Aqui a maioria dos observadores conclui que não existe uma estrutura de grupos “naturais", simplesmente uma única coleção homogênea de pontos. Idealmente, então, pode-se esperar que um método de análise de agrupamento aplicado a esses dados chegue a uma conclusão semelhante. Em um estudo mais aprofundado, este pode não ser o caso, e muitos (a maioria) dos métodos de análise de agrupamento dividirão o tipo de dados visto na Figura 6 em ‘grupos’. Geralmente, o processo de dividir um conjunto de dados homogêneo em diferentes partes é chamado de dissecção, e esse procedimento pode ser útil em circunstâncias específicas. Se, por exemplo, os pontos na Figura

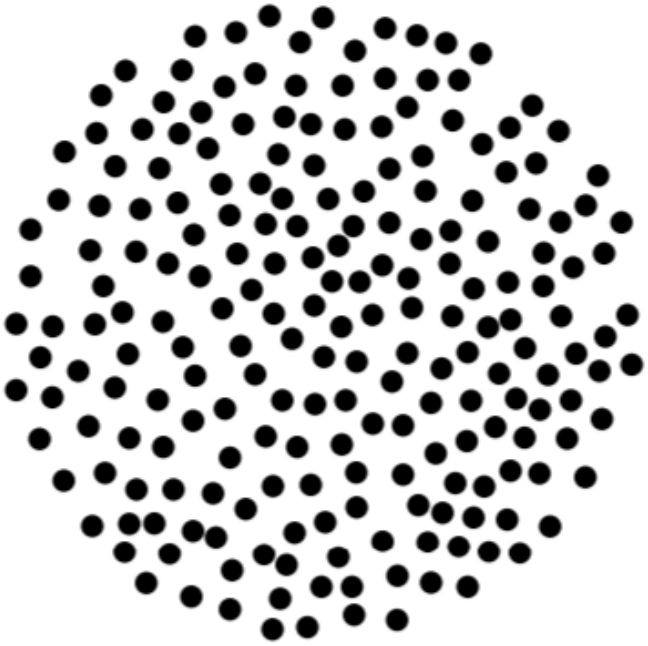

Figura 6 – Conjunto de dados sem grupo (veja Gordon (1980)).

6 representassem a localização geográfica das casas em uma cidade, a dissecção poderia ser uma maneira útil de dividir a cidade em distritos postais compactos que contêm números comparáveis de casas - veja a Figura 7. (Este exemplo foi sugerido por Gordon, 1980.) O problema é que, é claro, uma vez que na maioria dos casos o investigador não conhece a priori a estrutura dos dados (a análise de grupos é, afinal, destinada a ajudar a descobrir qualquer estrutura). Existe o risco de interpretar todas as soluções do agrupamento em termos da existência de grupos (naturais) distintos. O investigador pode, então, convenientemente “ignorar"a possibilidade de que a classificação produzida por uma análise de grupos seja um artefato do método e que na verdade ela esteja impondo uma estrutura em seus dados, em vez de descobrir algo sobre a estrutura real. Esse é um problema muito real na aplicação de técnicas de agrupamento.

Segundo Hair (1995), a análise de agrupamento, também conhecida como análise de conglomerados, é um conjunto de técnicas estatísticas cujo objetivo é agrupar objetos segundo suas características, formando grupos ou conglomerados homogêneos. Os objetos em cada conglomerado tendem a ser semelhantes entre si, e diferentes dos demais objetos dos outros conglomerados. Os conglomerados obtidos devem apresentar tanto uma homogeneidade interna (dentro de cada conglomerado), como uma grande heterogeneidade externa (entre conglomerados).

De acordo com Malhotra (2001), a análise de grupos tem uma ampla aplicação nas áreas de marketing para vários objetivos, como segmentação de mercado, compreensão do

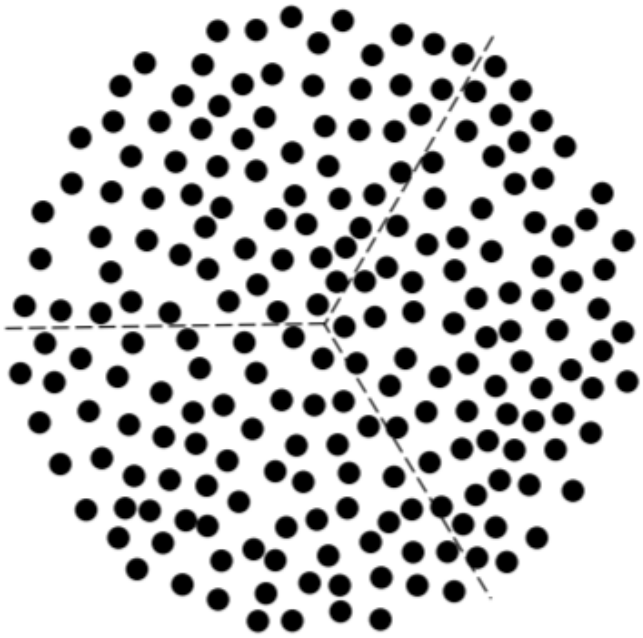

Figura 7 – Conjunto de dados sem grupo, dividido (veja Gordon (1980)).

comportamento do comprador, identificação das oportunidades de um novo produto, seleção de mercados de testes e redução de dados. A análise de agrupamento também vem sendo aplicada em áreas como a de investimentos, economia, saúde, etc.

Segundo Malhotra (2001), as etapas para a aplicação da análise de grupos são: formular o problema; escolher uma medida de distância; escolher um processo de aglomeração; decidir quanto ao número de conglomerados; interpretar e perfilar os conglomerados; avaliar a validade do processo de aglomeração. No contexto de mofometria deve-se notar que é preciso trabalhar com vetores de pré-formas e usar uma distância apropriada para as pré-formas.

Como um processo de aprendizado não supervisionado, o agrupamento de dados é frequentemente usado como uma etapa preliminar da análise de dados. Por exemplo, o agrupamento de dados é usado para identificar os padrões ocultos nos dados de expressão gênica (MacCuish, 2010), para produzir uma boa qualidade de grupos ou resumos de big data para tratar dos problemas analíticos e de armazenamento associados (Fahad et al., 2014), para selecionar apólices de seguro representativas de um grande portfólio para construir modelos de metamodelos (Gan, 2013; Gan e Lin, 2015).

Tambem é necessário ter cuidado com o uso de métodos de agrupamento, porque o particionamento de dados específico (e as pontuações *outliers* correspondentes) podem variar significativamente com a escolha da metodologia do agrupamento. Portanto, geralmente é aconselhável agrupar os dados várias vezes e calcular a média das pontuações obtidas nas

diferentes execuções. Os resultados dessa abordagem geralmente são surpreendentemente robustos (Aggarwall 2017).

3.5 TIPOS DE AGRUPAMENTOS

A análise de agrupamento é uma das técnicas mais conhecidas e populares para o reconhecimento de padrões. Dessa forma, há vários modelos de agrupamentos e algoritmos que analisam distribuições, densidades, possíveis pontos centrais e um conjunto de dados. Os métodos hierárquicos e não-hierarquicos são os mais conhecidos.

Os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os elementos. São populares na área de bioinformática e funcionam muito bem. Os métodos hierárquicos se dividem em dois grupos - os aglomerativos e divisivos. Para mais detalhes veja (JOHNSON et al., 1992).

Nos métodos hierárquicos aglomerativos todos elementos formam seu próprio grupo inicial. Então, os grupos mais próximos são mesclados juntos de uma maneira iterativa, até que finalmente todos indivíduos se unem em apenas um grupo, o qual inclui todos elementos do conjunto original dos dados. O grande problema dessa abordagem é que as distâncias entre os agrupamentos devem ser recalculadas a cada iteração, o que torna o processo extremamente lento para grandes massas de dados, elevando o custo computacional.

Os métodos hierárquicos divisivos, por outro lado, possuem abordagem inversa. O agrupamento começa a partir de um único grupo que vai se dividindo iterativamente em grupos menores até cada elemento tornar-se seu próprio grupo. Cada um desses métodos podem ser vistos em (KAUFMAN e ROUSSEEUW, 2009), (JAIN e DUBES, 1988).

Os métodos não-hierárquicos são organizados em dois paradigmas: rígido (*hard*) e difuso (fuzzy). O método rígido tem como princípio particionar os dados em grupos mutuamente exclusivos. Um dos primeiros algoritmos com essa técnica é o algoritmo K-means (MACQUEEN et al., 1967). O objetivo do agrupamento *hard* é alocar cada elemento do conjunto de dados a um e somente um grupo. Em contraste, o agrupamento *fuzzy* permite que cada elemento pertença a mais de um grupo por meio de graus de pertinência. Portanto, propicia informação muito mais detalhada da estrutura dos dados. Isso cria um conceito de limites *fuzzy* o qual difere do conceito *hard* de limites bem definidos.

3.6 K-MEANS PARA FORMAS PLANAS

Antes de definir o k-means é necessário definir um algoritmo para gerar a partição inicial. O algoritmo de partição inicial será apresentado no algoritmo 5. Deve-se notar que essa partição inicial também é utilizada nos outros algoritmos de agrupamento que serão estudados.

Algoritmo 5: Algoritmo de partição inicial

1: Calcule as pré-formas de cada objeto pela equação 2.4.

2: Escolha de forma aleatória k objetos para serem os centros dos grupos.

3: Atribua cada objeto ao centro k mais próximo.

4: Calcule a forma média dos grupos que é definida logo acima da equação (2.7).

5: Calcule $J = \sum_{r=1}^{k} \sum_{i \in C_r} \text{Dist}^2(z_i, \mu_r)$.

6: Repita o passo 1 ao 5 um número M de vezes. Em seguida, escolha a partição associada ao menor valor de J.

Muitos algoritmos de agrupamento foram desenvolvidos nos últimos sessenta anos. Entre esses algoritmos, o algoritmo k-means é um dos algoritmos de agrupamento não hierárquicos mais antigos e mais usados (Macqueen, 1967; Jain, 2010)

O algoritmo de agrupamento k-means foi desenvolvido por MacQueen (1967) posteriormente sendo adaptado a uma nova versão feita para trabalhar com dados de análise estatística de formas. O algoritmo k-means possui como objetivo particionar n observações dentre K grupos de modo que cada observação pertença ao grupo cuja distância entre essa observação e o protótipo (forma média) do grupo é mínima. O algoritmo de agrupamento k-means para formas planares está definido no Algoritmo 6.

Algoritmo 6: Algoritmo k-means

1: Utilize o algoritmo de partição incial 5 aplicado as pré-formas dos objetos z_i.

2: Calcule a forma média dos grupos.

3: Atribua cada objeto a forma média do grupo mais próximo.

4: Calcule a forma média de cada grupo.

5: Repita o passo 3 e 4 até que não haja mais mudança de elementos para outros grupos ou um número máximo de iterações seja atingido.

Para adaptar o método k-means ao agrupar objetos com base em marcos anatômicos, é desnecessário considerar alguns conceitos básicos da análise de formas (Amaral 2010).

De acordo com Hastie, Tibshirani e Friedman (2008) o agrupamento k-means é um método para localizar grupos e centros de grupos em um conjunto de dados não rotulados. Primeiramente escolhe-se o número desejado de centros dos grupos, digamos K, e o procedimento k-means move iterativamente os centros para minimizar o total dentro da variação do grupo.

O método k-means é muito semelhante ao método de Forgy, a única diferença é que, cada vez que um objeto muda de grupo, os centróides do seu grupo antigo e novo são recalculados. Isso pode ser feito facilmente usando uma equação de atualização para as coordenadas do centróide (Kaufman Rousseeuw 1990).

O k-means é pode ser implementado para resolver muitos problemas corriqueiros. Ele é adequado para grupos compactos e hiperesféricos. Sua complexidade computacional com relação ao tempo do k-means é $O(nKp)$. Uma vez que K (número de grupos) e p (dimensão) são geralmente menores que n (número de objetos), k-means pode ser utilizado para agrupar grandes conjuntos de dados. As desvantagens do k-means são também bem estudadas e como resultado disto muitas variações do k-means têm sido propostas para superar estes obstáculos(Assis 2019).

Não há nenhum método eficiente e universal perfeito para identificar o número de grupos e a partição inicial (Fayyad et al., 1996). Os protótipos encontrados no estado de convergência variam com a partição inicial. Uma estratégia geral para este problema é executar o algoritmo várias vezes com diferentes partições aleatórias iniciais (Bradley et al., 1998).

O k-means também é sensitivo a ruído, anomalias e outliers. Mesmo se um indivíduo está bem afastado dos protótipos dos grupos ele é obrigado a ser incluído em algum grupo desta forma distorcendo a forma daquele grupo. Em geral, o k-means converge para um ponto estacionário em tempo finito, não há garantias sobre se o ponto de convergência é de fato o mínimo global ou, se não, a que distância estamos dele, uma inicialização ruim dos centróides pode nos levar a uma solução não ótima. Na prática, é aconselhável executar o algoritmo várias vezes e selecionar a solução com a menor dispersão dentro do grupo (Peter Flach 2015).

Uma grande desvantagem dos algoritmos k-means é que eles são baseados em valores médios e, consequentemente, são muito sensíveis a valores discrepantes. Tais valores atípicos, que podem não ser incomuns em amostras grandes, podem deteriorar significativamente o desempenho desses algoritmos, mesmo que representem apenas uma pequena fração dos dados, conforme explicado em García et al. (2012) ou Croux et al. (2007). A abordagem k-medians é uma primeira tentativa de obter algoritmos de clustering mais robustos; foi sugerido

por MacQueen (1967) e desenvolvido por Kaufman e Rousseeuw (1990). Consiste em considerar critérios baseados em normas mínimas em vez de normas quadráticas, de modo que os centros de aglomerados sejam as medianas espaciais, também chamadas de medianas geométricas ou L1 (veja Small, 1990), dos elementos pertencentes a cada aglomerado. Muitos algoritmos foram propostos na literatura para encontrar esse mínimo. O mais popular é certamente o algoritmo PAM (particionamento em torno dos medóides), desenvolvido por Kaufman e Rousseeuw (1990), a fim de buscar mínimos locais entre os elementos da amostra. Seu tempo de computação é $O(kn^2)$ e, como conseqüência, não está muito bem adaptado para amostras grandes.

3.7 ALGORITMO K-MEDÓIDE PARA FORMAS PLANAS

O algoritmo k-medóide, diferente do k-means, não utiliza a média como valor de referência para o centro dos grupos, ele utiliza o objeto mais central de cada grupo para ser o centro. O algoritmo k-medóide, requer que os centros sejam pontos dos dados. Observe que calcular o medóide de um grupo requer o exame de todos os pares de pontos - enquanto o cálculo da média requer apenas uma única passagem pelos pontos - o que pode ser excessivo para grandes conjuntos de dados. A qualidade de um agrupamento Q é calculada como a distância total entre todos os pontos até o medóide mais próximo. Observe que existem $k(n-k)$ pares de um medóide e um não-medóide, e avaliar Q requer iteração sobre $n-k$ pontos de dados, portanto, o custo computacional de uma iteração é quadrático no número de pontos de dados. No k-medóide os objetos são $D \subseteq X$ o número de grupos $K \in \mathbb{N}$ e $\mu_1, ..., \mu_k \in D$, representa os grupos preditivos de X.

O algoritmo de agrupamento k-medóides esta definido no algoritimo abaixo:

Algoritmo 7: Algoritmo k-medóide

1: Utilize o algoritmo de partição inicial 5 para ter a primeira divisão dos grupos.

2: Escolha aleatoriamente k centros para representar os $\mu_1, ...\mu_k$ protótipos dos grupos iniciais.

3: Atribua cada objeto x_i ao grupo cujo centro estiver mais próximo $argmin_j Dis(z_i, \mu_j)$.

4: Calcule a distância de cada objeto para o seu centro $D_l = \sum_{i=1}^{k} Dist(z_i, \mu_j)$

5: Repita o passo 1 ao 3 se $D_{l+1} < D_l$ então fique com o grupo D_{l+1}

6: Repita do passo 1 ao 5 um número grande de vezes.

De acordo com Kaufman Rousseeuw (1990) para obter k grupos, o método seleciona

k objetos (que são chamados de objetos representativos) no conjunto de dados. Os grupos correspondentes são encontrados, atribuindo cada objeto restante ao objeto representativo mais próximo.

Para ser exato, a distância média (ou dissimilaridade média) do objeto representativo para todos os outros objetos do mesmo grupo está sendo minimizada. Por esse motivo, um objeto representativo ótimo que chamamos de medóide de seu aglomerado e o método de particionamento em torno de medóides, chamamos de técnica de k-medóide (Kaufman Rousseeuw 1990).

3.8 ALGORITMO FUZZY C-MEANS PARA FORMAS PLANAS

Nesta seção é descrito como o algoritmo de fuzzy c-means proposto por Bezdec (1984) pode ser adaptado para o contexto de morfometria em formas planas onde é utilizado uma distância para espaços não-euclideanos. Nesse caso, cada objeto é representado por um vetor de pré-formas e as distâncias entre os objetos são medidas a partir de distâncias de pré-formas tais como a geodésica.

Normalmente, as abordagens de agrupamento geram partições de forma que um indivíduo pertença somente a um grupo. Este tipo de agrupamento é chamado de agrupamento *hard*, em que os grupos nesse tipo de abordagem são disjuntos. O método de agrupamento fuzzy muda essa noção permitindo associar um indivíduo com todos os grupos, nesse tipo de abordagem os grupos são disjuntos. O agrupamento fuzzy estende essa noção para permitir associar um indivíduo com todos os grupos usando uma função de pertinência (Zadeh, 1965). O conceito do método fuzzy oferece a vantagem de expressar o tipo de situação em que um indivíduo compartilha similaridade com vários grupos através da possibilidade do algoritmo que associa cada indivíduo parcialmente a todos os grupos.

Seja $\{Y_1, \ldots Y_c\}$ uma partição dos objetos definidos em Y. Uma partição *hard* de Y é formada a partir dos subconjuntos que correspondem a:

$$Y_i \neq \phi, \quad 1 \leq i \leq c;$$

$$Y_i \cap Y_j = \phi; \quad i \neq j$$

e

$$\cup_{i=1}^{c} Y_i = Y,$$

em que ϕ denota o conjunto vazio, $\cup$ e $\cap$ denotam as operações de união e intersecção.

Seja c um inteiro com $1 < c < n$ e seja $Z = \{z_1, ..., z_n\}$ um conjunto das pré-formas dos n indivíduos na esfera complexa de dimensão $k-1$. Dado Z, pode-se dizer que c subconjuntos fuzzy $\{u_i : Z \rightarrow [0;1]\}$ são uma c partição fuzzy de Z se os $c \times n$ valores $\{u_{ik} = u_i(z_k), 1 \leq k \leq n, 1 \leq i \leq c\}$ satisfazem as seguintes três condições.

$$0 \leq u_{ik} \leq 1, \quad \forall i,k;$$

$$\sum_i u_{ik} = 1, \quad \forall k$$

e

$$0 < \sum_k u_{ik} < n \quad \forall i.$$

Deve-se notar a diferença entre as partições *hard* e fuzzy. No caso da partição fuzzy cada objeto pertence parcialmente aos c grupos. Por outro lado, no caso da partição *hard*, cada objeto pertence apenas a um dos grupos.

Qualquer conjunto de valores $c \times n$ que satisfaça às condições acima pode formar uma matriz $(c \times n)$, $U = [u_{ik}]$.

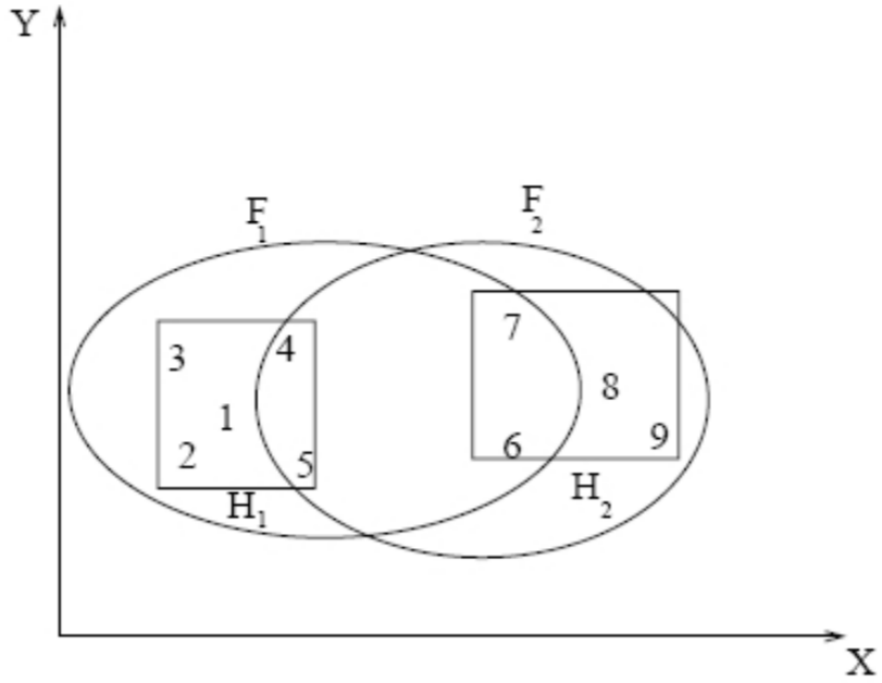

Figura 8 – Agrupamento fuzzy x *hard*

Na Figura 8, os retângulos dividem o conjunto de dados em dois agrupamentos *hard*: $H1 = \{1,2,3,4,5\}$ e $H2 = \{6,7,8,9\}$. Um algoritmo de agrupamento fuzzy poderia produzir os dois grupos fuzzy F1 e F2 representados por elipses. Os indivíduos terão pertinência definida no intervalo [0; 1] para cada grupo. Por exemplo, o grupo fuzzy F1 e F2 pode ser descrito $F1 = \{(1,0.9)(2,0.8)(3,0.7)(4,0.6)(5,0.55)(6,0.2)(7,0.2)(8,0.0)(9,0.0)\}$ e $F2 = \{(1,0.0)(2,0.0)(3,0.0)(4,0.1)(5,0.15)(6,0.4)(7,0.35)(8,1.0)(9,0.9)\}$

Os pares ordenados (n,u_i) em cada grupo representam, respectivamente, o indivíduo n e sua pertinência ao grupo i, u_i. Quanto maior o valor de pertinência indica alta confiança na associação de um indivíduo ao grupo. Uma partição *hard* pode ser feita a partir da partição fuzzy, para isso basta aplicar um limiar aos valores de pertinência.

O algoritmo de agrupamento fuzzy c-means esta definido a seguir:

Algoritmo 8: Algoritmo fuzzy c-means

1: Escolha o número de grupos c , $2 \leq c \leq n$, determine m, $1 \leq m \leq \infty$ e fixe o número máximo de iterações N.

2: Utilize o algoritmo de partição inicial 5 e a partir daí defina a primeira matriz de pertinência cujos elementos são dados por u_{ik}.

3: Calcule o protótipo dos grupos p_i pela equação: $p_i = \frac{\sum_{k-1}^{n} (u_{ik})^m z_k}{\sum_{k-1}^{n} u_{ik}}$, em que z_k denota a pré-forma do k-ésimo objeto.

4: Atualize a matriz de pertinência u_{ik} pela equação: $u_{ik} = (\sum_{j=1}^{c} a_{ijk})^{-1}$.

5: Em que: $a_{ijk} = (d_{ik}/d_{ik})^{\frac{1}{(m-1)}}$ e $d_{ik} = ||z_k - p_i||^2$.

6: Repita o passo 3, 4 e 5 N vezes.

A abordagem fuzzy c-means pode ser muito eficaz para resolver muitos problemas de análise de grupos e o comportamento da abordagem fuzzy c-means na prática está bem documentado. Entretando, existem muitas questões substanciais não respondidas sobre o fuzzy c-means (Hathaway Bezdek 1988).

3.9 ÍNDICE DE RAND

Nesse trabalho será calculado o índice Rand ajustado (CR) ou a medida Rand. Na estatística e, em particular, na análise de grupos de dados, é uma medida da semelhança entre dois

agrupamentos de dados. O índice CR mede o quão bem ajustados ficaram os grupos, fazendo uma comparação entre uma partição física a priori e a posterior (posteriori) obtida da partição fornecida pelo algoritmo de agrupamento. O CR tem valores no intervalo [-1, 1], onde o valor 1 indica perfeito ajuste dos dados aos grupos específicos, enquanto valores próximos a 0 (ou negativos) correspondem a um grupo mais aleatório, encontrada ao acaso.

Sendo $X = x_1, \ldots x_i, \ldots, x_R$ e $Y = y_1, \ldots, y_j, \ldots, y_C$ as duas partições do mesmo conjunto de dados que possuem, respectivamente, grupos R e S. Então, o índice Rand ajustado é:

XY	Y_1	Y_2	$\ldots$	Y_S	Soma
X_1	n_{11}	n_{12}	$\ldots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\ldots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
X_R	n_{1r}	n_{2r}	$\ldots$	n_{rs}	a_t
Soma	b_1	b_2	$\ldots$	b_s	

e

$$CR = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{a_i}{2}\sum_{j=1}^{C}\binom{b_j}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{a_i}{2} + \sum_{j=1}^{C}\binom{b_j}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{a_i}{2}\sum j = 1^{C}\binom{b_j}{2}}, \quad (3.1)$$

em que $\binom{a_i}{2} = \frac{n(n-1)}{2}$ e n_{ij} representa o número de objetos que estão nos grupos x_i e y_j, assim como b_i indica o número de objetos no grupo y_i e a_i indica o número de objetos no grupo x_i.

Além do índice de Rand, há vários outros critérios para se definir se o agrupamento foi feito de forma adequada ou não, tais como o índice de Jaccard, índice Folkes Mallows, entre outros.

4 ANÁLISE NUMÉRICA

Nessa seção serão ilustrados os métodos apresentados acima com dados reais e dados simulados. Isso servirá para avaliar a eficiência de cada método para diferentes parâmetros da distribuição Bingham complexa apresentada no algoritmo 4 e para os dados reais. A distância utilizada para a alocação dos dados nos grupos foi a distância geodésica apresentada, na equação (2.8). Cada simulação terá 100 réplicas diferentes de Monte Carlo e a média do índice de Rand apresentada na equação (3.1) será exibida nas tabelas. Para os dados simulados, utilizando a distribuição Bingham complexa, foram gerado dados utilizados três parâmetros diferentes, três diferentes tamanhos de amostra e três rotações diferentes. Em todas as simulações e análise dos dados foram utilizados os métodos de agrupamentos k-means, k-medóide e c-means aprensentados nos algoritmos 6, 7 e 8, respectivamente.

4.1 DADOS MACACOS

Em uma investigação para avaliar as diferenças cranianas entre os macacos pelo sexo, foram estudados 29 gorilas machos e 30 fêmeas adultas, 28 chimpanzés machos e 26 fêmeas adultas e 30 orangotangos machos e 24 fêmeas adultas. Os dados são descritos em detalhes por O'Higgins (1989) e O'Higgins e Dryden (1993). Oito marcos anatômicos são escolhidos no plano para cada crânio. Os pontos de referência são marcos anatômicos e foram selecionadas por um biólogo especialista.

Uma ilustração dos crânios dos gorilas, chinpanzés e orangotangos, machos e fêmeas é apresentada a seguir na figura 9.

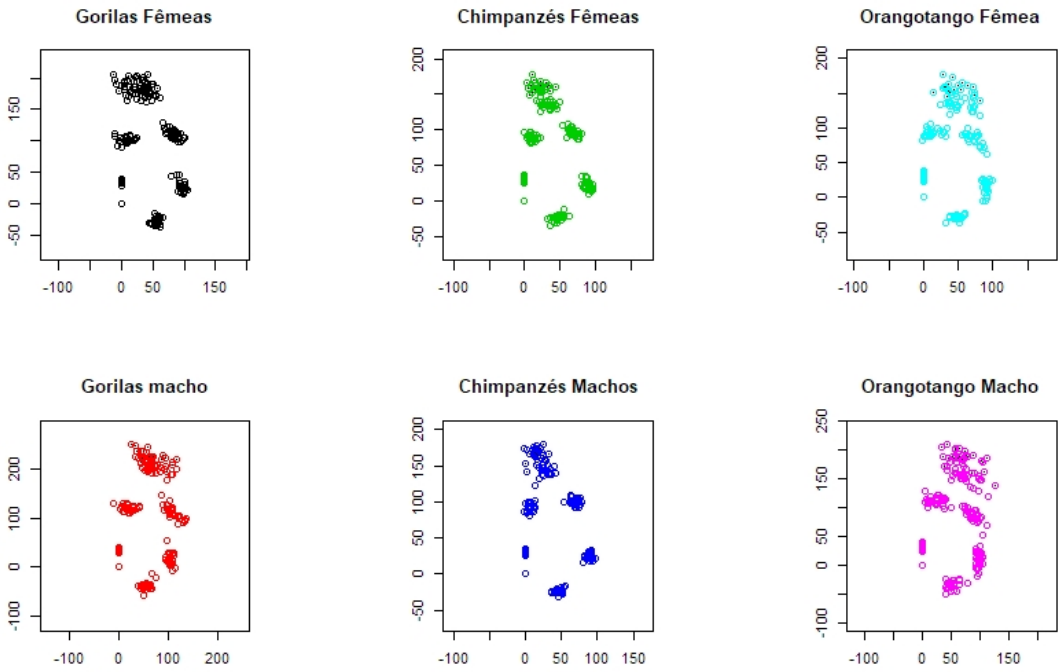

Figura 9 – Os seis grupos com os marcos anatômicos dos crânios dos macacos: (coluna da esquerda) gorilas fêmeas e machos; (coluna do meio) chimpanzés fêmeas e machos; e (coluna da direita) orangotangos fêmeas e machos (Dryden e Mardia (2016)).

4.1.1 Gorilas

K-means	Macho	Fêmea	Índice de Rand
Macho	27	2	0,7428
Fêmea	2	28	

Tabela 1 – Tabela do k-means para os dados dos crânios dos gorilas machos e gorilas fêmeas.

K-medóide	Macho	Fêmea	Índice de Rand
Macho	27	2	0,7428
Fêmea	2	28	

Tabela 2 – Tabela do k-medóide para os dados dos crânios dos gorilas machos e gorilas fêmeas.

C-means	Macho	Fêmea	Índice de Rand
Macho	27	2	0,7428
Fêmea	2	28	

Tabela 3 – Tabela do c-means para os dados dos crânios dos gorilas machos e gorilas fêmeas.

Pelas Tabelas 1, 2 e 3, contendo dados de 29 gorilas machos e 30 gorilas fêmeas, todos os métodos foram igualmente efetivos. Dados de 27 dos 29 machos e 28 das 30 fêmeas foram agrupados de forma correta e cada método teve o mesmo valor no CR = 0,7428

4.1.2 Chimpanzés

K-means	Macho	Fêmea	Índice de Rand
Macho	13	8	-0,0026
Fêmea	15	18	

Tabela 4 – Tabela do k-means para os dados dos crânios dos chimpanzés machos e chimpanzés fêmeas.

K-medóide	Macho	Fêmea	Índice de Rand
Macho	12	8	-0,0054
Fêmea	16	18	

Tabela 5 – Tabela do k-medóide para os dados dos crânios dos chimpanzés machos e chimpanzés fêmeas.

C-means	Macho	Fêmea	Índice de Rand
Macho	20	7	0,13
Fêmea	9	18	

Tabela 6 – Tabela do c-means para os dados dos crânios dos chimpanzés machos e chimpanzés fêmeas.

Nas Tabelas 4, 5 e 6, que contêm os dados referentes aos chimpanzés, o k-means, que se utiliza da forma média, agrupou quase metade dos dados de maneira inadequada, com $CR = -0,0026$. O resultado do agrupamento pelo método k-medóide foi semelhante, com $CR = -0,0054$. Já o método c-means apresentou o melhor resultado, com $CR = 0,13$, identificando 16 chimpanzés como pertencendo ao grupo errado. Estes resultados indicam que o algoritmo proposto, c-means, funciona bem para esse conjunto de dados reais.

4.1.3 Orangotangos

K-means	Macho	Fêmea	Índice de Rand
Macho	24	2	0,4856
Fêmea	6	22	

Tabela 7 – Tabela do k-means para os dados dos crânios dos orangotangos machos e orangotangos fêmeas.

K-medóide	Macho	Fêmea	Índice de Rand
Macho	24	5	0,4339
Fêmea	6	19	

Tabela 8 – Tabela do k-medóide para os dados dos crânios dos orangotangos machos e orangotangos fêmeas.

C-means	Macho	Fêmea	Índice de Rand
Macho	23	5	0,4191
Fêmea	7	19	

Tabela 9 – Tabela do c-means para os dados dos crânios dos orangotangos machos e orangotangos fêmeas.

Para o conjunto de dados de crânios de orangotangos, o índice de Rand dos três métodos foram semelhantes, de acordo com as Tabelas 7, 8 e 9. O algoritmo k-means classificou 8 crânios em um grupo errado e ficou com $CR = 0,485$. O algoritmo k-medóide classificou 11 objetos em um grupo errado e ficou com $CR = 0,4339$. O algoritmo c-fuzzy agrupou 12 crânios no grupo errado, ficando com $CR = 0,419$.

Nenhum algoritmo teve superioridade definitiva sobre os outros em todos os dados reais analisados. Diferentes algoritmos tiveram resultados superiores a depender dos dados analisados.

4.2 DADOS SIMULADOS

Todos os dados simulados foram feitos a partir da distribuição Bingham com 3 parâmetros. Para comparar os agrupamentos, foram selecionados 4 valores diferentes para os parâmetros $P_1 = (\ dados_1 = (\ \lambda_1 = 890,\ \lambda_2 = 895,\ \lambda_3 = 900)$; $dados_2 = (\lambda_1 = 800,\ \lambda_2 = 850,\ \lambda_3 = 900))$, $P_2 = (\ dados_1 = (\ \lambda_1 = 100,\ \lambda_2 = 200,\ \lambda_3 = 300)$, $dados_2 = (\ \lambda_1 = 898,\ \lambda_2 = 899,\ \lambda_3 = 900)), P_3 = (\ dados_1 = (\ \lambda_1 = 100,\ \lambda_2 = 250,\ \lambda_3 = 500)$, $dados_2 = (\ \lambda_1 = 150,\ \lambda_2 = 300,\ \lambda_3 = 600), P_4 = (\ dados_1 = (\ \lambda_1 = 100,\ \lambda_2 = 500,\ \lambda_3 = 1000)$, $dados_2 = (\ \lambda_1 = 980,\ \lambda_2 = 990,\ \lambda_3 = 1000))$, 3 tipos de rotação para os dados da segunda amostra $r_1 = (0.6 * \pi),\ r_2 = (20 * \pi)$ e $r_3 = (50 * \pi)$ e 3 tamanhos de amostra $n_1 = 25,\ n_2 = 50$ e $n_3 = 100$. Em todas as simulações, antes de aplicar o método de agrupamento, foi feita a partição inicial dos dados, que foi definida no algoritmo (5).

n=25	Índice de Rand	Tempo
K-means	0,128	1h20min
K-medóide	0,083	1h36min
C-means	0,103	1h25min
Rotação=0.6*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 10 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_1 e rotação r_1

n=25	Índice de Rand	Tempo
K-means	0,140	1h22min
K-medóide	0,105	1h36min
C-means	0,137	1h27min
Rotação=20*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 11 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_1 e rotação r_2

n=25	Índice de Rand	Tempo
K-means	0,130	1h21min
K-medóide	0,096	1h36min
C-means	0,102	1h28min
Rotação=50*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 12 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_1 e rotação r_3

De acordo com as Tabelas 10, 11 e 12, para os dados simulados com os parâmetros $\lambda_1 = (890, 895, 900)$ e $\lambda_2 = (800, 850, 900)$, rotação $r_1 = 0.6 * \pi$ e $n_1 = 25$, os resultados do

índice de Rand indicaram que os métodos têm desempenho semelhante. Houve um pequeno destaque para o k-means. Na Tabela 12, com $\lambda_1 = (890, 895, 900)$ e $\lambda_2 = (800, 850, 900)$ em que temos ambos parâmetros com baixa concentração, o k-means teve índice de Rand igual a 0,130 e o seu tempo de execução foi de 1 hora e 21 minutos, enquanto que o c-means teve índice de Rand igual a 0.102 e tempo de execução de 1 hora e 28 minutos. Apesar de, neste caso, o c-means não ter obtido o melhor desempenho, sob outras configurações que serão apresentadas a seguir este método mostra-se superior aos demais.

n=25	Índice de Rand	Tempo
K-means	0,404	1h20min
K-medóide	0,258	1h35min
C-means	0,398	1h28min
Rotação=0.6*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 13 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_1 e rotação r_1

n=25	Índice de Rand	Tempo
K-means	0,729	1h15min
K-medóide	0,736	1h37min
C-means	0,827	1h26min
Rotação=20*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 14 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_1 e rotação r_2

n=25	Índice de Rand	Tempo
K-means	0,790	1h52min
K-medóide	0,607	1h37min
C-means	0,878	1h32min
Rotação=50*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 15 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_1 e rotação r_3

Nas Tabelas 13, 14 e 15 os parâmetros utilizados para simular os dados foram: $\lambda_1 = (100, 200, 300)$ $\lambda_2 = (898, 899, 900)$, e $n_1 = 25$. Nesta configuração, com o aumento da rotação e a maior separação dos dados, o método c-means passou a se destacar em sua capacidade de agrupar corretamente os dados, mostrando-se melhor que os demais. Na Tabela 15, por exemplo, com a rotação igual a $r_3 = 50 * \pi$, o k-means apresentou índice de Rand igual a 0,790, enquanto o do c-means foi igual a 0,878, mostrando-se superior àquele.

n=25	Índice de Rand	Tempo
K-means	0,034	1h21min
K-medóide	0,012	1h35min
C-means	0,025	1h24min
Rotação=0.6*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 16 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_1 e rotação r_1

n=25	Índice de Rand	Tempo
K-means	0,151	1h21min
K-medóide	0,086	1h36min
C-means	0,091	1h25min
Rotação=20*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 17 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_1 e rotação r_2

n=25	Índice de Rand	Tempo
K-means	0,156	1h21min
K-medóide	0,061	1h36min
C-means	0,075	1h27min
Rotação=50*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 18 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_1 e rotação r_3

Nas Tabelas 16, 17 e 18 os dados foram simulados com os parâmetros $\lambda_1 = (100,250,500)$, $\lambda_2 = (150,300,600)$ e $n = 25$. Na Tabela 16, com rotação igual a $r_1 = 0,6 * \pi$, quase não houve diferença entre os métodos k-means e c-means, sendo o índice de Rand igual a 0,034 e 0,025, respectivamente. Já nas Tabelas 17 e 18 com rotação igual a $r_2 = 20 * \pi$ e $r_3 = 50 * \pi$, respectivamente, o método que teve melhor desempenho foi o k-means, com índice de Rand acima dos demais métodos.

n=25	Índice de Rand	Tempo
K-means	0,096	1h35min
K-medóide	0,080	1h36min
C-means	0,083	1h28min
Rotação=0.6*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 19 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_1 e rotação r_1

n=25	Índice de Rand	Tempo
K-means	0,182	1h35min
K-medóide	0,160	1h36min
C-means	0,403	1h28min
Rotação=20*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 20 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_1 e rotação r_2

n=25	Índice de Rand	Tempo
K-means	0,215	1h36min
K-medóide	0,188	1h36min
C-means	0,415	1h29min
Rotação=50*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 21 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_1 e rotação r_3

Nas Tabelas 19, 20 e 21 os dados foram simulados com os parâmetros $\lambda_1 = (100, 500,$
$1000)$, $\lambda_2 = (980, 990, 1000)$ e $n_1 = 25$. Nas tabelas 20 e 21, com rotação igual a $r_2 = 20 * \pi$ e $r_3 = 50 * \pi$, respectivamente, o método de agrupamento c-means demonstrou resultados consideravelmente melhores que os demais métodos, com $CR = 0,403$ e $CR = 0,415$. Na Tabela 19, a diferença entre os métodos k-means, k-medóide e c-means se mostrou pequena, sendo os índices de Rand muito semelhantes entre os três.

n=50	Índice de Rand	Tempo
K-means	0,120	4h53min
K-medóide	0,078	5h23min
C-means	0,085	5h12min
Rotação=0.6*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 22 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_2 e rotação r_1

n=50	Índice de Rand	Tempo
K-means	0,116	4h57min
K-medóide	0,097	5h22min
C-means	0,088	5h13min
Rotação=20*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 23 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_2 e rotação r_2

n=50	Índice de Rand	Tempo
K-means	0,122	5h8min
K-medóide	0,105	5h25min
C-means	0,097	5h18min
Rotação=50*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 24 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_2 e rotação r_3

Nas Tabelas 22, 23 e 24 os dados da distribuição Bingham foram simulados com os parâmetros $\lambda_1 = (100, 250, 500)$, $\lambda_2 = (150, 300, 600)$ e $n_2 = 50$. Na Tabela 22, com rotação igual a $r_1 = 0,6 * \pi$, o método k-means se mostrou melhor que os demais, com $CR = 0,120$ e tempo de execução de 5 horas e 8 minutos, enquanto nas Tabelas, 23 e 24, com rotações de $r_2 = 20 * \pi$ e $r_3 = 50 * \pi$, os métodos k-means e k-medóide tiveram índices de Rand muito próximos, sendo a diferença de $0,019$ e $0,017$ respectivamente.

n=50	Índice de Rand	Tempo
K-means	0,290	4h53min
K-medóide	0,223	5h21min
C-fuzzy	0,264	5h15min
Rotação=0.6*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 25 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_2 e rotação r_1

n=50	Índice de Rand	Tempo
K-means	0,637	5h52min
K-medóide	0,580	5h20min
C-fuzzy	0,746	5h18min
Rotação=20*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 26 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_2 e rotação r_2

n=50	Índice de Rand	Tempo
K-means	0,659	5h2min
K-medóide	0,588	5h21min
C-fuzzy	0,851	5h20min
Rotação=50*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 27 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_2 e rotação r_3

Nas Tabelas 25, 26 e 27 os dados foram simulados com os parâmetros $\lambda_1 = (100, 200, 300)$, $\lambda_2 = (898, 899, 900)$ e $n_2 = 50$. Na Tabela 25, com rotação igual a $r_1 = 0,6 * \pi$, o

k-means se mostrou o melhor método, com $CR = 0,290$, enquanto c-means teve um $CR = 0,264$. Entretanto, com o aumento da rotação nas tabelas 26 e 27 para $r_2 = 20 * \pi$ e $r_3 = 50 * \pi$, respectivamente, o melhor método de agrupamento foi o c-means, com $CR = 0,746$ e $CR = 0,851$, respectivamente, sendo superior ao dos demais métodos.

n=50	Índice de Rand	Tempo
K-means	0,032	4h53min
K-medóide	0,023	5h22min
C-means	0,035	5h16min
Rotação=0.6*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 28 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_2 e rotação r_1

n=50	Índice de Rand	Tempo
K-means	0,124	4h53min
K-medóide	0,086	5h22min
C-means	0,081	5h17min
Rotação=20*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 29 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_2 e rotação r_2

n=50	Índice de Rand	Tempo
K-means	0,138	4h55min
K-medóide	0,085	5h23min
C-means	0,089	5h18min
Rotação=50*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 30 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_2 e rotação r_3

Nas Tabelas 28, 29 e 30 os dados da distribuição Bingham foram simulados com os parâmetros de baixa concentração $\lambda_1 = (100,250,500)$, $\lambda_2 = (150,300,600)$ e $n_2 = 100$. Na Tabela há diferença entre os métodos k-means e c-means, com relação ao índice de Rand de ambos igual a $CR = 0,003$. Para a mesma rotação o k-medóide apresentou $CR = 0,023$. Com o aumento da rotação para $r_2 = 20 * \pi$ e $r_3 = 50 * \pi$, o melhor método foi o k-means, com $CR = 0,124$ e $CR = 0,138$, respectivamente.

n=50	Índice de Rand	Tempo
K-means	0,098	4h58min
K-medóide	0,078	5h22min
C-means	0,061	5h16min
Rotação=0.6*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 31 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_2 e rotação r_1

n=50	Índice de Rand	Tempo
K-means	0,217	5h1min
K-medóide	0,196	5h41min
C-means	0,289	5h21min
Rotação=20*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 32 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_2 e rotação r_2

n=50	Índice de Rand	Tempo
K-means	0,199	4h56min
K-medóide	0,197	5h38min
C-means	0,304	5h22min
Rotação=50*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 33 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_2 e rotação r_3

Nas Tabelas 31, 32 e 33 os dados foram simulados com os parâmetros $\lambda_1 = (100,500,$
$1000)$, $\lambda_2 = (980,990,1000)$ e $n_2 = 50$. Na Tabela 31, com rotação de $r_1 = 0,6 * \pi$, o CR do método k-means foi de $0,098$, maior que dos demais métodos. Já nas Tabelas 32 e 33, o método c-means demonstrou superioridade no agrupamento, apresentando índices de Rand de $0,289$ e $0,304$, com rotações de $r_2 = 20 * \pi$ e $r_3 = 50 * \pi$, respectivamente.

n=100	Índice de Rand	Tempo
K-means	0,142	18h38min
K-medóide	0,089	20h22min
C-means	0,106	20h35min
Rotação=0.6*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 34 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_3 e rotação r_1

n=100	Índice de Rand	Tempo
K-means	0,164	18h22min
K-medóide	0,145	20h22min
C-means	0,151	20h32min
Rotação=20*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 35 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_3 e rotação r_2

n=100	Índice de Rand	Tempo
K-means	0,282	18h26min
K-medóide	0,279	20h23min
C-means	0,281	20h33min
Rotação=50*π	λ_1=890,895,900	λ_2=800,850,900

Tabela 36 – Tabela com o índice de Rand para os dados simulados com parâmetros P_1, tamanho de amostra n_3 e rotação r_3

Nas Tabelas 34, 35 e 36 os dados foram simulados com $n_3 = 100$ e parâmetros de alta concentração $\lambda_1 = (890, 895, 900)$ e $\lambda_2 = (800, 850, 900)$. Nas Tabelas 34 e 35 o k-means foi o melhor método de agrupamento, com índices de Rand de $0,142$ e $0,164$ e tempo de execução em torno de 18 horas e 30 minutos. Já com rotação igual a $r_3 = 50 * \pi$, os métodos k-means e c-means tiveram desempenho semelhante, com índices de Rand igual a $0,282$ e $0,281$, e tempo de execução de 18 horas e 26 minutos e 20 horas e 33 minutos respectivamente.

n=100	Índice de Rand	Tempo
K-means	0,346	18h33min
K-medóide	0,173	20h22min
C-fuzzy	0,332	20h35min
Rotação=0.6*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 37 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_3 e rotação r_1

n=100	Índice de Rand	Tempo
K-means	0,516	18h35min
K-medóide	0,424	20h21min
C-fuzzy	0,729	20h34min
Rotação=20*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 38 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_3 e rotação r_2

n=100	Índice de Rand	Tempo
K-means	0,504	18h43min
K-medóide	0,487	20h20min
C-fuzzy	0,761	20h34min
Rotação=50*π	λ_1=(100,200,300)	λ_2=(898,899,900)

Tabela 39 – Tabela com o índice de Rand para os dados simulados com parâmetros P_2, tamanho de amostra n_3 e rotação r_3

Nas Tabelas 37, 38 e 39 os dados foram simulados com $n_3 = 100$ e parâmetros de baixa concentração $\lambda_1 = (100, 200, 300)$ e alta concentração $\lambda_2 = (898, 899, 900)$. Com baixa rotação igual a $r_1 = 0,6 * \pi$, como mostrado na Tabela 37, os resultados utilizando-se os métodos k-means e c-means foram muito semelhantes, com índices de Rand de $0,346$ e $0,332$, respectivamente. Com o aumento da rotação para $r_2 = 20 * \pi$ e $r_3 = 50 * \pi$, como apresentado nas Tabelas 38 e 39, respectivamente, o método c-means demonstrou melhores resultados, com CR de $0,729$ e $0,761$.

n=100	Índice de Rand	Tempo
K-means	0,053	18h19min
K-medóide	0,043	20h21min
C-means	0,043	20h35min
Rotação=0.6*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 40 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_3 e rotação r_1

n=100	Índice de Rand	Tempo
K-means	0,286	18h26min
K-medóide	0,155	20h23min
C-means	0,174	20h35min
Rotação=20*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 41 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_3 e rotação r_2

n=100	Índice de Rand	Tempo
K-means	0,379	18h25min
K-medóide	0,180	20h23min
C-means	0,187	20h34min
Rotação=50*π	λ_1=(100,250,500)	λ_2=(150,300,600)

Tabela 42 – Tabela com o índice de Rand para os dados simulados com parâmetros P_3, tamanho de amostra n_3 e rotação r_3

Nas Tabelas 40, 41 e 42 os dados foram simulados com $n_3 = 100$ e parâmetros de

baixa concentração $\lambda_1 = (100, 250, 500)$ e $\lambda_2 = (150, 300, 600)$. Como observado na Tabela 40, os métodos k-means e c-means tiveram desempenho semelhante com uma rotação de $r_1 = 0,6 * \pi$, apresentando CR de $0,053$ e $0,043$, respectivamente. Com o aumento da rotação para $r_2 = 20 * \pi$ na Tabela 31 e $r_3 = 50 * \pi$ na tabela 42, o método k-means mostrou-se melhor que os demais métodos, com índices de Rand de $0,286$ e $0,189$.

n=100	Índice de Rand	Tempo
K-means	0,092	18h26min
K-medóide	0,002	20h25min
C-means	0,056	20h35min
Rotação=0.6*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 43 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_3 e rotação r_1

n=100	Índice de Rand	Tempo
K-means	0,247	18h21min
K-medóide	0,191	20h24min
C-means	0,403	20h35min
Rotação=20*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 44 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_3 e rotação r_2

n=100	Índice de Rand	Tempo
K-means	0,370	18h21min
K-medóide	0,368	20h25min
C-means	0,423	20h34min
Rotação=50*π	λ_1=(100,500,1000)	λ_2=(980,990,1000)

Tabela 45 – Tabela com o índice de Rand para os dados simulados com parâmetros P_4, tamanho de amostra n_3 e rotação r_3

Nas Tabelas 43, 44 e 45 os dados foram simulados com $n_3 = 100$, parâmetros de baixa concentração $\lambda_1 = (100, 500, 1000)$ e parâmetros de alta concentração $\lambda_2 = (980, 990, 1000)$. Na Tabela 43 o método k-means foi melhor que os demais, com $CR = 0,092$, enquanto que nas Tabelas 44 e 45 o método fuzzy c-means foi melhor, com $CR = 0,403$ e $CR = 0,423$.

Em suma, o método proposto para formas planas, o c-means, é útil e tem melhor desempenho que os demais quando trabalha com dados de alta concentração e baixa concentração. O algoritmo c-means consegue agrupar melhor os dados nesse contexto.

5 CONSIDERAÇÕES FINAIS

O último capítulo desta dissertação visa estabelecer conclusões a respeito do trabalho e sugerir algumas recomendações para futuros trabalhos relacionados ao tema principal.

O presente trabalho objetivou apresentar métodos já existentes para análise de formas em morfometria e desenvolver novos métodos de agrupamento. Os métodos utilizados foram o já existente na literatura, k-means e os novos métodos desenvolvidos, k-medóide e fuzzy c-means. Comparando-se os métodos citados, foi possível observar que quando a distribuição Bingham complexa possui um conjunto de dados com parâmetros de alta concentração e outro conjunto de dados com parâmetros de baixa concentração, o método proposto c-means obtém melhores resultados, ou seja, é capaz de agrupar os dados de maneira mais compatível com o esperado, tendo assim um melhor índice de Rand quando se tem um alto valor de rotação nos dados. Sendo assim, seu uso é recomendado nessas situações.

Para isso, foram realizado os três agrupamentos: k-means, k-medóide e c-means, tanto para dados reais como para dados simulados, utilizando a distância geodésica onde em situações pontuais o c-means se mostrou-se melhor que o k-means. Com isso, considera-se que o objetivo do trabalho tenha sido atingido - desenvolver um método de agrupamento para pré-formas planas em análise de grupos.

Com relação aos objetivos secundários desta dissertação, foi elaborada uma revisão bibliográfica consistente com o tema, incluindo conceitos ligados à morfometria, à formas e marcos anatômicos, à matriz de Helmert, às pré-formas, à distância, à distribuição Bingham complexa e como rotaciona-lá, aos métodos de agrupamento k-means, k-medóide, c-fuzzy, às razões para a classificação, às definições de grupo e ao índice de Rand, além de ter sido apresentado um panorama dessas definições e aplicações dos métodos. Estes objetivos foram alcançados no decorrer dos Capítulos 2 e 3. É importante ressaltar que os objetivos secundários foram meios para se obter o objetivo central desta dissertação. No capítulo 4, foi atingido o objetivo central dessa dissertação. Através da análise numérica, foi possível obter os resultados positivos que demonstraram o bom índice do agrupamento fuzzy c-means tanto para os dados reais dos crânios dos chimpanzés quanto para os alguns dos dados simulados da distribuição bingham complexa.

Uma ideia para trabalhos futuros é utilizar *bagging (Bootstrap Aggregating)*, um método proposto por Breiman em 1996, em que um conjunto de dados é gerado por amostragem

bootstrap dos dados originais. O conjunto de dados gera um conjunto de modelos utilizando um algoritmo de aprendizagem simples por meio da combinação por votos para classificação.

REFERÊNCIAS

AGGARWAL, Charu C.. **Outlier Analysis**. 2. ed. New York: Springer, 2017. 481 p.

AMARAL, G. J. A.; Dore, L. H. ; Lessa, R. P. ; STOSIC, B. . k-Means Algorithm in Statistical Shape Analysis. Communications in Statistics. Simulation and Computation, v. 39, p. 1016-1026, 2010.

AMARAL, G. J. A.; DRYDEN, I. L; A WOOD, Andrew T.. Pivotal Bootstrap Methods fork-Sample Problems in Directional Statistics and Shape Analysis. Journal Of The American Statistical Association, [s.l.], v. 102, n. 478, p.695-707, jun. 2007. Informa UK.

ASSIS, E. C. ; SOUZA, R. ; AMARAL, GETULIO J.A. . **Using bagging to enhance clustering procedures for planar shapes**. International Journal of Business Intelligence and Data Mining, 2019.

BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: THE FUZZY C-MEANS CLUSTERING ALGORITHM. **Computers & Geosciences**, Printed, v. 10, n. 2, p.191-203, 1984.

BISHOP, Christopher M.. Pattern Recognition and Machine Learning. Cambridge: Springer, 2006. 758 p.

BONNER, R. E. On some clustering techniques. **International Business Machines Journal of Research and Development**, v. 8, 1964. p. 22–32.

BOOKSTEIN **A statistical method for biological shape comparasions**. Journal of Theoretical Biology. v. 107, 1984. p. 475 520.

BRADLEY, R. A.; ASAKAWA, N.; LATORRACA, S.; MAGALHÃES, F. M. M.; OLIVEIRA, L. A.; PEREIRA, R. M. **Levantamento quantitativo de microrganismos solubilizadores de fosfato na rizosfera de gramíneas e leguminosas forrageiras na amazônia. Acta Amazônica, Manaus**, v.1, p.12-22, 1988.

BREIMAN, Leo. Bagging predictors. Machine Learning, [s.l.], Springer Science and Business Media LLC. v. 24, n. 2, p.123-140, ago. 1996.

CARDOT, Hervé; CÉNAC, Peggy; MONNEZ, Jean-marie. A fast and recursive algorithm for clustering large datasets with -medians. **Computational Statistics & Data Analysis**, [s.l.], v. 56, n. 6, p.1434-1449, jun. 2012. Elsevier BV.

CORMACK, R. M. A review of classification. **Journal of the Royal Statistical Society A**, v. 134. 1971. 321–367.

CROUX, Christophe et al. Machine Learning and Robust Data Mining. **Computational Statistics & Data Analysis**, [s.l.], v. 52, n. 1, p.151-154, set. 2007. Elsevier BV.

DORE, Luiz H. G. et al. Bias-corrected maximum likelihood estimation of the parameters of the complex Bingham distribution. **Brazilian Journal Of Probability And Statistics**, [s.l.], v. 30, n. 3, p.385-400, ago. 2016. Institute of Mathematical Statistics.

DRYDEN, Ian L.; MARDIA, Kanti V.. Statistical Shape Analysis. 2. ed. West Sussex: John Wiley & Sons, 2016. 520 p.

EVERITT, Brian S. et al. **Cluster Analysis**. 5. ed. West Sussex: 1 John Wiley & Sons, 2011. 348 p.

FAHAD, Adil et al. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. **Ieeee Transactions On Emerging Topics In Computing**, [s.l.], Institute of Electrical and Electronics Engineers (IEEE). v. 2, n. 3, p.267-279, set. 2014.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery & Data Mining**. 1 ed. American Association for Artificial Intelligence, Menlo Park, Califórnia, 1996. 611 p.

FLACH, Peter. **MACHINE LEARNING**: The Art and Science of Algorithms that Make Sense of Data. New York: Mpg Books Group, 2012. 416 p.

GAN, Guojun. Application of data clustering and machine learning in variable annuity valuation. **Insurance: Mathematics and Economics**, [s.l.], v. 53, n. 3, p.795-801, nov. 2013. Elsevier BV.

GAN, Guojun; LIN, X. Sheldon. Valuation of large variable annuity portfolios under nested simulation: A functional data approach. **Insurance: Mathematics and Economics**, [s.l.], v. 62, p.138-150, maio 2015. Elsevier BV.

GARCÍA-TREVIÑO, E.s.; BARRIA, J.a.. Online wavelet-based density estimation for non-stationary streaming data. **Computational Statistics & Data Analysis**, [s.l.], v. 56, n. 2, p.327-344, fev. 2012. Elsevier BV.

GOULD, SJ. & R. F. lohnston. **Geographic variation**. Ann. Rev. Eco!. Syst., 3: 457 - 498. 1972.

GORDON, A. D. **Classification**, 1st ed. Chapman and Hall CRC, London. 1980.

GORDON, A. D. **Classification**, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL. 1999.

HAIR, Joseph F. et al. **Multivariate Data Analysis: With Readings**. 4. ed. Pearson College Div, 1995.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning:** Data Mining, Inference, and Prediction. 2. ed. Stanford: Springer, 2008. 764 p.

HATHAWAY, Richard J.; BEZDEK, James C.. Recent Convergence Results for the Fuzzy c-Means Clustering Algorithms. **Journal Of Classification.** p. 237-247. 1988

JAIN, Anil K.. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, [s.l.], v. 31, n. 8, p.651-666, jun. 2010. Elsevier BV.

JARDINE, N.. A Logical Basis for Biological Classification. **Systematic Biology**, [s.l.], Oxford University Press (OUP). v. 18, n. 1, p.37-52, 1 mar. 1969.

KAUFMAN, L., Rousseeuw, P., **Finding Groups in Data: An Introduction to Cluster Analysis**. John Wiley & Sons, New York. p. 255. 1990.

KENDALL, D. G. The diffusion of shape. **Advances in applied Probability**; v 9, p. 428 430. 1977.

KENDALL, D. G. Shape manifolds, **Procrustean metrics and complex projective space**. Bulletin of the London Mathematical Society, v 16. p. 81 121. 1984.

KENDALL, D. G. et al. **Shape and Shape Theory**. Cambridge: Wiley-blackwell, 1999. 567 p.

KENT, John T.; CONSTABLE, Patrick D.l.; ER, Fikret. Simulation for the complex Bingham distribution. **Statistics And Computing**, [s.l.], Springer Science and Business Media LLC. v. 14, n. 1, p.53-57, jan. 2004.

LIU, Bing. **Web Data Mining**: Exploring Hyperlinks, Contents, and Usage Data. 2. ed. New York: Springer, 2007.

MACCUISH, J.D., MACCUISH, N.E., **Clustering in Bioinformatics and Drug Discovery**. CRC Press, Boca Raton, FL. 2010.

MACQUENN, J. **Some methods for classification and analysis of multivariate observations**. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281–297, University of California Press, Berkeley, Calif., 1967.

MALHOTRA, N. K. **Pesquisa de marketing**: uma orientação aplicada. 3. Ed. Porto Alegre: Bookman, 2001.

NEEDHAM, R. M. Computer methods for classification and grouping. **The Use of Computers in Anthropology** (I. Hymes, ed.) 345–356. Mouton, The Hague. 1965.

NEFF, N. A. & Marcus, L. F. **A survey of multivariate methods for systematics**. Privately published, N.Y. 1980.

PIMENTEL, R. A. **Morphometrics**. Kendall/Hunt, Dubuque. 1979.

SMALL, C.G., **A survey of multidimensional medians**. International Statistical Review/Revue Internationale de Statistique 58 (3), 263–277. 1990.

STRANEY, D.O. & J.L. Patton. **Phyllogenetic and environmental determinants of geographic variation of the pocket mouse Perognathus goldmani Osgood**. Evolution, 34: 888-933. 1980.

TUKEY, John. **Exploratory Data Analysis**. Pearson, 1977. 688 p.

ZADEH, L. A.. Fuzzy sets. **Information And Control**, California, v. 8, n. 0, p.338-353, 1965.