UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
CURSO DE GRADUAÇÃO EM ESTATÍSTICA

GABRIEL HARRISON FIDELIS TEOTONIO

ANÁLISE ESTATÍSTICA DE FORMAS E ALGORITMO $K-MEANS$ PARA DADOS EM 3D: UMA APLICAÇÃO EM AGRUPAMENTO DE CÓRTEX CEREBRAIS

RECIFE
2020

GABRIEL HARRISON FIDELIS TEOTONIO

ANÁLISE ESTATÍSTICA DE FORMAS E ALGORITMO $K-MEANS$ PARA DADOS EM 3D: UMA APLICAÇÃO EM AGRUPAMENTO DE CÓRTEX CEREBRAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientador: Prof. Dr. Getúlio J. A. Amaral

Recife

2020

Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística
Coordenação de Graduação
Av. Prof. Luiz Freire, s/n, Cidade Universitária 50740-540 Recife - PE
Fone: 81 2126-8423/8420 Fax:81 2126-8421
www.de.ufpe.br/graduacao.html e-mail: bacharelado@de.ufpe.br

DECLARAÇÃO

Declaro para os devidos fins, que **GABRIEL HARRISON FIDELIS TEOTÔNIO** portador do CPF: 118.826.924-03, apresentou o Trabalho de Conclusão de Curso intitulado **"ANÁLISE ESTATÍSTICA DE FORMAS E ALGORITMO K-MEANS PARA DADOS EM 3D: UMA APLICAÇÃO EM AGRUPAMENTO DE CÓRTEX CEREBRAIS."** em 27 de outubro de 2020.

A banca examinadora foi composta pelos seguintes membros: Professor Dr. Getúlio José Amorim do Amaral (Presidente), Professora Dra. Calitéia Santana de Souza e o Professor Dr. Abraão David C. Nascimento.

Recife, 27 de outubro de 2020.

Coordenação da Graduação em Estatística

À minha família.

AGRADECIMENTOS

Aos meus pais, Antônia e Joselito, que diante de muitas adversidades durante minha vida, conseguiram me guiar para os melhores caminhos e me ajudar nos momentos difíceis. Além de me proverem uma excelente educação e valores íntegros.

À minha tia Tilde e suas filhas, Daisy, Daiany e Débora, que foram essenciais na minha criação e desenvolvimento. Sem vocês, eu não estaria aqui.

Aos amigos que o DE me proporcionou conhecer, pela amizade e companheirismo ao longo dessa jornada. Em especial, Alan Douglas, Antônio Soares, Arthur Machado, José Henrique, Pedro Estevão e Willams Batista.

À minha companheira, Letícia, que me permitiu compartilhar minha jornada com a dela ao longo dos anos na universidade e me apoiou em momentos difíceis. Agradeço, até hoje, por ter pego o mesmo ônibus que você 4 anos atrás.

Ao professor Getúlio Amaral, pela orientação deste trabalho.

Ao professor Ian Dryden que, mesmo distante, proporcionou-me várias discussões e soluções ao longo da pesquisa.

Aos meus professores do DE, pela dedicação em me fazer questionar. Em especial, Audrey e Francisco Cysneiros, que foram essenciais na minha inserção no mundo da pesquisa acadêmica. Como, também, Klaus Vasconcellos, Abraão Nascimento, Renato Cintra, André Leite e Raydonal Ospina, que contribuíram de maneira ímpar na minha formação.

Aos colegas da In Loco, pela parceria e ajuda em muitos momentos complicados. Em especial, Raíza Oliveira, Abel Borges e Tiago Lima.

Aos membros da banca, por investirem seu tempo na crítica deste texto.

À Raquel e Jardiclebson, pela solicitude de sempre.

À PROAES, pelo apoio financeiro ao longo do curso.

"The art of measuring, as precisely as possible, probabilities of things, with the goal that we would be able always to choose or follow in our judgments and actions that course, which will have been determined to be better, more satisfactory, safer or more advantageous."

-Jacob Bernoulli (1655-1705)

RESUMO

Algoritmos de agrupamento permitem a criação de duas ou mais partições distintas, a partir de um único conjunto de dados. Neste trabalho, abordamos o problema de agrupamento de formas pertencentes ao espaço tridimensional, a partir de um conjunto de dados do córtex cerebral de indivíduos que possuem o diagnóstico de esquizofrenia. Primeiramente, revisamos a literatura da Análise Estatística de Formas com enfoque nos principais conceitos e exemplificando como seus métodos podem ser aplicados em problemas nos quais a estrutura dos objetos de estudo é relevante. Em seguida, revisamos o método de agrupamento *K-means* numa visão geral e, também, no contexto da Análise Estatística de Formas, recapitulando as principais contribuições desse tipo de agrupamento, nos espaços bidimensional e tridimensional. No caso tridimensional, quando temos um grande número de *landmarks*, o método de *Procrustes* para o cálculo da forma média se torna inviável, devido ao grande tempo de processamento necessário para esse método, neste caso, iterativo. A literatura apresenta a forma média ϕ como uma alternativa ao método de *Procrustes*, quando consideramos o $\mathbb{R}^3$, já que essa possui um estimador de forma fechada para a forma média. Mostramos que a forma média ϕ possui, de fato, uma performance computacional para o cálculo da forma média acima do método de *Procrustes*, avaliando-as em cenários com diferentes números k de *landmarks*. Dessa forma, propusemos o uso da forma média ϕ no algoritmo *K-means*, com o objetivo de reduzir o tempo computacional do agrupamento, tornando-o competitivo no momento de atualização dos centroides dos grupos definidos. Mostramos uma redução relevante do tempo de processamento do algoritmo, quando utilizada a forma média ϕ, além do resultado do agrupamento ser similar a quando utilizamos o método de *Procrustes*. Finalmente, analisamos o resultado do agrupamento e não foram encontrados indícios de associação entre os grupos obtidos e a presença do diagnóstico de esquizofrenia. O índice ajustado de Rand foi utilizado para avaliar o método de agrupamento abordado.

Palavras-chave: Agrupamento. Espaço não-Euclidiano. Análise estatística de formas. Computação paralela.

ABSTRACT

Clustering algorithms allow you to partition a single data set into two or more different sets. In this work, we approach the problem of grouping shapes in three dimensions, using a set of data from the cerebral cortex of individuals who have schizophrenia. First, we review the literature on Statistical Shape Analysis with a focus on the main concepts and exemplifying how their methods can be applied to problems in which the structure of the objects of study is relevant. Then, we review the K-means grouping method in an overview and, also, in the context of Statistical Shape Analysis, reviewing the main contributions of this type of grouping, in bidimensional and tridimensional spaces. In the case of three dimensions, when we have a large number of landmarks, the Procrustes method for calculating the mean shape becomes impracticable, due to the large processing time required for this iterative method. The literature presents the mean ϕ-shape as an alternative to the Procrustes method, when in $\mathbb{R}^3$, since it has a closed form estimator for the mean shape. We show that the mean ϕ-shape has, in fact, a computational performance, for the calculation of the mean shape, above the method of Procrustes, evaluating them in scenarios with different numbers k of landmarks. Thus, we proposed the use of the mean ϕ-shape in the K-means algorithm, with the objective of reducing the computational time of the grouping, making it more competitive when updating the centroides of the defined groups. We show a significant reduction in the processing time of the algorithm, when using the mean ϕ-shape, in addition to the result of the grouping being similar to when using the method of Procrustes. Finally, we analyzed the result of the grouping and there was no evidence of an association between the groups obtained and the presence of the diagnosis of schizophrenia. The Rand's adjusted index was used to evaluate the grouping method addressed.

Keywords: Clustering. Non-Euclidian space. Statistical shape analysis. Parallel computing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE SÍMBOLOS

i Número de observações

k Número de *landmarks*

m Dimensão do espaço

$\boldsymbol{X}$ Matriz de configuração

$\boldsymbol{X}_H$ Matriz de configuração centrada

$\boldsymbol{Z}$ Matriz de pré-forma

$[\boldsymbol{X}]$ Matriz de forma

S_m^k Espaço de pré-forma

$\sum_m^k$ Espaço de forma

$\boldsymbol{\Gamma}$ Matriz de rotação/reflexão

$S(\boldsymbol{X})$ Tamanho do centroide

$\boldsymbol{H}^F$ Matriz de Helmert

$\boldsymbol{H}$ Submatriz de Helmert

$\boldsymbol{M}$ Variedade diferenciável

$T_p(\boldsymbol{M})$ Espaço tangente de $\boldsymbol{M}$ no ponto p

$d_F(\boldsymbol{X}_1, \boldsymbol{X}_2)$ Distância total de *Procrustes*

$\rho(\boldsymbol{X}_1, \boldsymbol{X}_2)$ Distância Riemanniana

$P_m(k-1)$ Espaço das matrizes simétricas de dimensão $k-1$

$\phi(\boldsymbol{\Xi})$ Forma média Fi

K Número de grupos no agrupamento

$W(C_k)$ Função objetivo

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Neste estudo, trabalhamos com o problema de agrupamento de dados de formas tridimensionais. Em particular, utilizando o algoritmo de aprendizado de máquina não-supervisionado *K-means* proposto por (LLOYD, 1982) numa versão adaptada para o contexto da Análise Estatística de Formas. O desenvolvimento do algoritmo *K-means* no contexto da Estatística de Formas já é disseminado e possui algumas aplicações para dados bidimensionais, a mencionar (AMARAL *et al.*, 2010; GEORGESCU, 2009). Por sua vez, a modelagem de dados de formas tridimensionais utilizando o algoritmo *K-means* é única (VINUé; SIMO; ALEMANY, 2014). Quando passamos a lidar com dados de formas em três dimensões, isto é, os objetos de estudo sendo considerados como subconjuntos do $\mathbb{R}^3$ (SMALL, 1996, página 29), nos deparamos com algumas mudanças analíticas que modificam a maneira como estrutura-se a lógica e performance computacional do algoritmo. A principal mudança é a falta de um estimador para a forma média por *Procrustes* (DRYDEN, 2016, página 136) de forma fechada. Ou seja, a estrutura da forma média precisaria ser encontrada de uma maneira iterativa.

Uma alternativa para a estimação da forma média, para os casos em que a dimensão $m \geq 3$, foi proposta em (DRYDEN *et al.*, 2008), a qual possui forma fechada e, por isso, sugere uma melhora na performance computacional no processo de estimação, em comparação com o método de *Procrustes*. Esta nova definição de forma média, chamada de forma média ϕ, é baseada no método de escalonamento multidimensional, comumente referido como MDS, sigla em inglês para *Multidimensional Scaling* (MARDIA J. T. KENT, 1995, página 394). É proposto, nesse trabalho, a utilização da forma média ϕ no algoritmo *K-means* para a atualização dos centroides durante o processamento com a disposição de melhorar a performance computacional do algoritmo. Uma vez que a estimação da forma média por métodos iterativos pode consumir um longo período para convergir quando consideradas grandes amostras, ou um grande número de *landmarks*, que é o nosso caso nesse trabalho.

Também é considerado teste de hipóteses acerca da forma média dos grupos encontrados como resultado do agrupamento. O algoritmo em questão é utilizado como um classificador de formas e, portanto, a aplicação dele deve conduzir a uma separação

das observações em grupos nos quais as intra-observações são mais similares do que as inter-observações. É proposto em (DRYDEN *et al.*, 2008) teste de hipóteses para a forma média ϕ no caso de dados de formas em que $m \geq 3$. O teste apresentado é uma extensão das ideias propostas por (AMARAL; DRYDEN; WOOD, 2007).

Como uma aplicação, é considerado o problema no campo da neurociência de classificar córtex cerebrais de pacientes com o diagnóstico de esquizofrenia. É constatado que pacientes com o diagnóstico da doença possuem uma assimetria, chamada de torque cerebral, menor no córtex (BILDER *et al.*, 1994; MACKAY *et al.*, 2003). Este problema já foi discutido em (BRIGNELL, 2007; BRIGNELL *et al.*, 2010) e (DRYDEN *et al.*, 2008) que, em particular, visitou esse conjunto de dados para exemplificar a utilização da forma média ϕ proposta.

Neste trabalho, consideramos o livro proposto por (DRYDEN, 2016) como material de apoio, para as definições fundamentais. O pacote **shapes** (DRYDEN, 2018) é utilizado em todas as análises computacionais, simulações e resultados numéricos. Para o algoritmo *K-means* é utilizada uma versão modificada da disponível no pacote **anthropometry** (VINUé, 2017). Além disso, consideramos os pacotes **parallel** (R Core Team, 2020), **foreach** e **doSNOW** (Microsoft; WESTON, 2017) para a execução do algoritmo em processos paralelizados. Todos os pacotes são baseados na linguagem de programação R (R Core Team, 2019).

1.2 OBJETIVOS

Temos os seguintes objetivos:

- Revisar a teoria e métodos da Análise Estatística de Formas. Em particular, uma introdução de conceitos fundamentais e do processo de estimação da forma média;
- Revisar métodos de implementação do algoritmo *K-means* e a literatura que se refere à sua utilização no contexto de formas;
- Propor uma adaptação do algoritmo *K-means*, utilizando a forma média ϕ.

1.3 ESTRUTURA DO TRABALHO

O trabalho tem a seguinte estrutura:

- No **Capítulo 2**, revisamos conceitos fundamentais e literaturas clássicas da Análise Estatística de Formas, com objetivo de esclarecer a usabilidade de seus métodos e

as mudanças tocantes à mudança de espaço e perspectiva presentes no seu domínio. Explorando, também, o processo de estimação da forma média e algumas variações para o caso em que $m = 3$;

- No **Capítulo 3**, primeiro, revisamos o algoritmo *K-means* e algumas de suas adaptações. Em particular, focamos na implementação do algoritmo no contexto de formas e na avaliação de sua performance. Depois, usamos o material discutido no Capítulo 2 para propor uma adaptação ao algoritmo utilizando a forma média ϕ;
- No **Capítulo 4**, consideramos uma aplicação dos desenvolvimentos apresentados nos capítulos anteriores, na área da neurociência, que tem como objetivo agrupar a forma do córtex cerebral de pacientes com esquizofrenia dentre pacientes com e sem o diagnóstico.

2 ANÁLISE ESTATÍSTICA DE FORMAS

No nosso cotidiano, comumente usamos a palavra forma para descrever a aparência de algum objeto em relação a sua estrutura física. Essa ideia acerca da definição da palavra forma é bastante consolidada dentro do pensamento humano. Por exemplo, quando queremos dizer que dois ou mais objetos são similares ou parecidos, em geral, estamos nos apegando a essa noção de similaridade na estrutura física dos objetos. Podemos pensar, de uma maneira mais refinada, que cada objeto possui um conjunto de dados geométricos associados à ele e nosso interesse é fazer uma comparação das geometrias dos objetos, buscando entender sua variabilidade. Ou seja, a partir da geometria de um objeto podermos quantificar uma similaridade com outro.

2.1 FUNDAMENTOS

Entretanto, essa comparação informal está associada à subjetividade do observador. Desse modo, consideramos a definição de forma dada em (KENDALL, 1977).

Definição 2.1. ***Forma*** *é toda informação geométrica que permanece quando os efeitos de locação, escala e rotação são removidos de um objeto.*

Segundo (DRYDEN, 2016, Capítulo 1), a forma de um objeto é invariante sob transformações de similaridade euclidiana de translação, escala e rotação. Isso quer dizer que após os efeitos mencionados serem retirados, temos a forma do objeto sem nenhuma perda de informação. Logo, dois objetos possuem a mesma forma se eles podem ser transladados, redimensionados e rotacionados um para o outro de modo que eles se ajustem, isto é, se eles são semelhantes. Outra definição fundamental que temos no contexto da estatística de formas é a de tamanho-e-forma. Às vezes, existe um desejo de comparar não somente a forma de certos objetos, como também a escala, tamanho dos mesmos.

Definição 2.2. ***Tamanho-e-forma*** *é toda informação geométrica que permanece quando os efeitos de locação e rotação são removidos de um objeto.*

Da mesma maneira como discutimos na Definição 2.1, agora, dois objetos possuem o mesmo tamanho-e-forma se eles podem ser transladados e rotacionados um para o outro de modo que eles se ajustem. As definições anteriores se referiam, de

maneira figurativa, a dois objetos para descrever os conceitos de forma do ponto de vista matemático. Mas, em geral, temos um conjunto de objetos e desejamos quantificar e mensurar a variabilidade de suas formas. O primeiro trabalho de investigação da análise de formas, do ponto de vista geométrico, foi realizado por (THOMPSON, 1917), com aplicações no campo biológico.

Discutimos no início deste capítulo, que comumente usamos a definição de forma em uma associação a estrutura física de um objeto. Mesmo após definirmos o conceito de um modo mais refinado, ainda resta um questionamento: como podemos descrever de forma numérica esta estrutura que entendemos de maneira visual? Com o intuito de resolver esse problema, surge a definição de *landmark*.

Definição 2.3. *Um* ***Landmark*** *é um ponto de correspondência em cada objeto que coincide entre e dentro de populações.*

A escolha de um conjunto finito de pontos no objeto deve ser feita de modo que esses pontos consigam resumir as principais informações geométricas. Existem três tipos principais de *landmarks*: científico, matemático e pseudo. A seguir, temos uma definição para cada um desses tipos.

- **Landmark científico** é um ponto assinalado por um especialista que corresponde, em algum sentido lógico, a um significado científico associado ao objeto de estudo. Esse *landmark* também é conhecido por *landmark* biológico, nome utilizado em aplicações na área da biologia. Por exemplo, o canto do olho de algum animal.
- **Landmarks matemáticos** são pontos localizados no objeto de acordo com alguma propriedade matemática da figura. Por exemplo, no ponto de alta curvatura.
- **Pseudo Landmark** são pontos construídos em um objeto, localizados ao redor do contorno ou entre *landmarks* científicos ou matemáticos. Curvas contínuas podem ser aproximadas por um grande número de pseudo-*landmarks* ao longo da curva.

Na biologia, ainda existem mais 3 tipos de *landmarks* que são caracterizados por alguns contextos particulares (BOOKSTEIN, 1992). Na aplicação do Capítulo 4, consideramos pseudo-*landmarks* para a caracterização da superfície cerebral em questão. Como os primeiros estudos utilizando a análise de formas foram feitos na área da biologia, boa parte dos *landmarks* considerados, ao longo dos anos, foram científicos. Entretanto, com o avanço das tecnologias de escaneamento, o uso de pseudo-*landmarks* para aproximar as superfícies dos objetos de estudo está em uma crescente, como discutido em (BRIGNELL

et al., 2010). Também existem trabalhos em análise de formas de superfície de objetos (COSTA; CESAR, 2000) com uma abordagem diferente da definida a partir de (KENDALL, 1977).

Podemos nos perguntar, ainda, como associar dois ou mais *landmarks* em diferentes objetos? Como podemos saber que esses *landmarks*, em cada um dos objetos, são equivalentes para comparação? Podemos estabelecer essa relação a partir da seguinte definição.

Definição 2.4. *Um* ***rótulo*** *é um nome ou número associado a um landmark, e identifica qual par de landmarks são correspondentes quando comparados dois objetos. Landmarks com esses rótulos associados são chamados de* ***landmarks rotulados****.*

Dois exemplos podem ser encontrados em (AMARAL *et al.*, 2010) e (VINUé; SIMO; ALEMANY, 2014). Em (AMARAL *et al.*, 2010), os *landmarks* possuem um número associado que podemos utilizar para a criação de pares entre dois objetos. Já em (VINUé; SIMO; ALEMANY, 2014), os *landmarks* estão associados às partes do corpo nas quais eles foram delimitados e coletados.

Anteriormente, discutimos sobre como, subjetivamente, associamos a definição da forma de um objeto com a sua estrutura física. As definições nos apresentaram o conceito de *landmark* e como ele é utilizado para representar pontos principais da característica física dos objetos em estudo, sejam definidos com um intuito mais anatômico, como os *landmarks* científicos, ou com o objetivo de captura de uma superfície contínua, como os pseudo-*landmarks*. A seguir, temos a definição que formaliza o uso de um conjunto de *landmarks* para descrever a forma de um objeto.

Definição 2.5. *A* ***configuração*** *é o conjunto de landmarks em um particular objeto. A* ***matriz de configuração*** $\boldsymbol{X}$ *é a $k \times m$ matriz de coordenadas cartesianas dos k landmarks em m dimensões. O* ***espaço de configuração*** *é o espaço das coordenadas de todos os landmarks.*

Na aplicação do Capítulo 4, consideramos $k = 62.501$ em $m = 3$ dimensões e o espaço de configuração é o espaço real das matrizes $k \times m$ ($\mathbb{R}^{km}$). É a partir da matriz de configuração que retiramos os efeitos de locação, escala e rotação para chegarmos até a forma, propriamente dita, do objeto. O primeiro efeito que desejamos retirar da matriz de configuração é o efeito de locação, em outras palavras, desejamos centralizar o

objeto no espaço no qual ele se encontra. Para isso, consideramos a submatriz de Helmert (KENDALL, 1984) para remover a locação. A submatriz $\boldsymbol{H}$ é a $(k-1)\times k$ matriz de Helmert sem a primeira linha. A matriz completa $\boldsymbol{H}^F$, a qual é comumente utilizada na Estatística, é uma matriz ortognal quadrada $k\times k$ com os elementos da sua primeira linha iguais a $1/\sqrt{k}$, e as linhas restantes são ortogonais a primeira. A primeira linha de $\boldsymbol{H}^F$ é descartada então a transformação $\boldsymbol{HX}$ não depende da locação da configuração original, como discutido em (DRYDEN, 2016, página 49).

Definição 2.6. *A j-ésima linha da* ***submatriz de Helmert H*** *é dada por*

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

e a j-ésima linha consiste de h_j repetida j vezes, seguida por $-jh_j$ e depois $k-j-1$ zeros, $j=1,\ldots,k-1$.

Para $k=3$, por exemplo, a matriz completa de Helmert é dada por

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

e a submatriz de Helmert

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

Considere

$$\boldsymbol{X_H}=\boldsymbol{HX}\in\mathbb{R}^{(k-1)m}, \tag{2.1}$$

como sendo as coordenadas Helmertizadas dos *landmarks*. A partir dessa transformação, utilizando a matriz $\boldsymbol{H}$, temos uma matriz de configuração, agora, centrada $\boldsymbol{X_H}$. Após retirar o efeito de locação, desejamos retirar o efeito de escala do objeto. Com isso em mente, considere a seguinte definição.

Definição 2.7. *O* ***tamanho do centroide*** *é dado por*

$$S(\boldsymbol{X})=\|\boldsymbol{CX}\|=\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}(\boldsymbol{X}_{ij}-\bar{\boldsymbol{X}}_j)^2},\quad \boldsymbol{X}\in\mathbb{R}^{km},$$

em que $\boldsymbol{X}_{ij}$ é a $(i,j)-$ésima entrada de $\boldsymbol{X}$, a média aritmética na $j-$ésima dimensão é $\bar{\boldsymbol{X}}_j = 1/k \sum_{i=1}^{k} \boldsymbol{X}_{ij}$,

$$\boldsymbol{C} = \boldsymbol{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T \tag{2.2}$$

é a matriz de centralização, e

$$\|\boldsymbol{X}\| = \sqrt{tr(\boldsymbol{X}^T\boldsymbol{X})}$$

é a norma euclidiana, $\boldsymbol{I}_k$ é a $k \times k$ matriz identidade e $\mathbf{1}_k$ é o $k \times 1$ vetor de uns. Note que $\boldsymbol{C}$ também pode ser usada, como uma alternativa, para a retirada da locação do objeto, de modo que

$$\boldsymbol{X}_{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{X}, \tag{2.3}$$

são as coordenadas centradas dos *landmarks*. Temos a seguinte relação entre ambas as alternativas de centralização

$$\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T = \boldsymbol{C}$$

e

$$\boldsymbol{H}^T\boldsymbol{X}_H = \boldsymbol{H}^T\boldsymbol{H}\boldsymbol{X} = \boldsymbol{C}\boldsymbol{X}.$$

Com o objetivo de retirar o efeito de escala, nós padronizamos dividindo pela noção de tamanho da Definição 2.7.

$$\|\boldsymbol{X}_H\| = \sqrt{tr(\boldsymbol{X}^T\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{X})} = \sqrt{tr(\boldsymbol{X}^T\boldsymbol{C}\boldsymbol{X})} = \|\boldsymbol{C}\boldsymbol{X}\| = S(\boldsymbol{X}), \tag{2.4}$$

em que $\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{C}$ é idempotente. Como observado em (DRYDEN, 2016, página 64), $S(\boldsymbol{X}) > 0$ porque não permitimos um conjunto de *landmarks* completamente coincidentes.

Na Figura 1, temos um exemplo do cálculo do tamanho do centroide, a partir *boxplots* do tamanho do centroide dos seis grupos do conjunto de dados dos crânios de

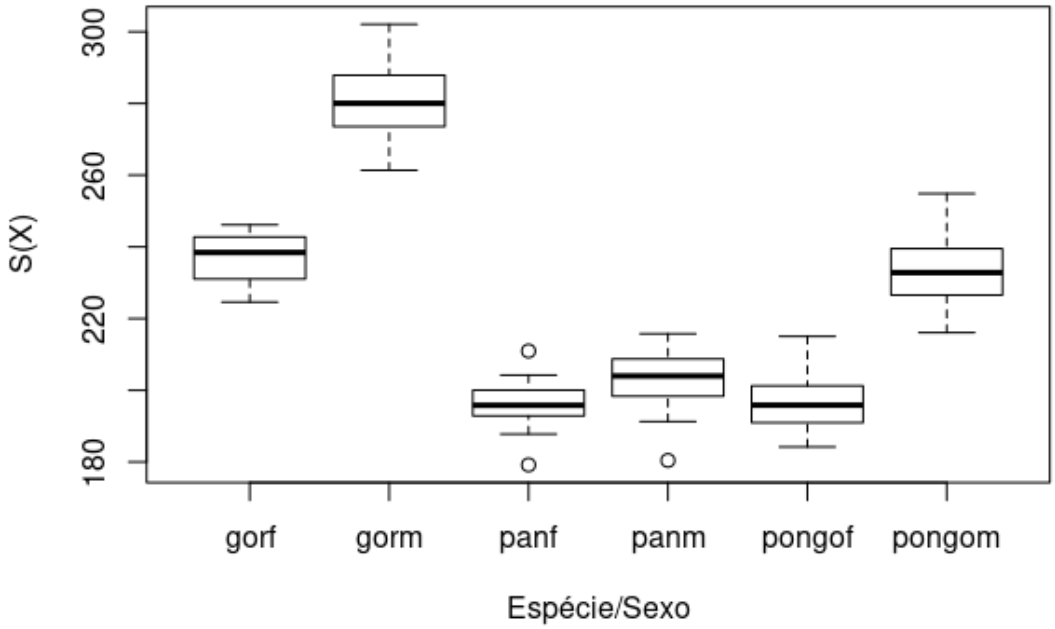


Figura 1 – Boxplots do tamanho dos centroides para o conjunto de dados dos macacos

gorilas (DRYDEN, 2016). É visível que existem diferenças no tamanho do centroide entre gorilas e orangotangos (pongo), com todos os machos sendo maiores do que as fêmeas. Para os chimpanzés (pan), há uma sobreposição com os orangotangos que são fêmeas.

2.1.1 Espaços de Formas

Quando retiramos o efeito de escala de uma matriz de configuração, temos o que é chamado pela literatura de pré-forma.

Definição 2.8. *A **pré-forma** de uma matriz de configuração é dada por*

$$\boldsymbol{Z} = \frac{\boldsymbol{X}_H}{\|\boldsymbol{X}_H\|} = \frac{\boldsymbol{HX}}{\|\boldsymbol{HX}\|},$$

a qual é invariável sob a translação e o dimensionamento da configuração original.

Podemos utilizar também a matriz de centralização $\boldsymbol{C}$ para uma definição alternativa à 2.8. Entretanto, o uso de $\boldsymbol{Z}$ nos traz a vantagem do número de linhas ser inferior ao definido em $\boldsymbol{C}$. Nós usamos a notação S_m^k para denotar o espaço de pré-forma de k pontos em m dimensões.

Definição 2.9. *O **espaço de pré-forma** é o espaço de todas as pré-formas. Formalmente, o espaço de pré-forma S_m^k é o espaço em órbita das configurações não coincidentes do conjunto de pontos k em $\mathbb{R}^m$ sob os efeitos de translação e escala.*

O espaço de pré-forma $S_m^k := S^{(k-1)m-1}$ é uma hiperesfera de raio unitário em $(k-1)m$ dimensões reais, uma vez que $\|\boldsymbol{Z}\| = 1$. Na seção 2.2.1, são abordados conceitos acerca de Geometria Diferencial e a relação com Análise Estatística de Formas. Após retirar os efeitos de locação e escala, nos resta remover o efeito de rotação da configuração do objeto para, enfim, obter a forma (Definição 2.1). Para retirar o efeito de rotação de uma configuração, consideramos uma matriz de rotação $\boldsymbol{\Gamma}$.

Definição 2.10. *Uma **matriz de rotação** $m \times m$ satisfaz $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{I}_m$ e det$(\boldsymbol{\Gamma}) = +1$. Uma matriz de rotação também é conhecida como uma matriz ortogonal especial, a qual é uma matriz ortognal com determinante +1. O conjunto de todas as matrizes de rotação $m \times m$ é conhecido como o grupo ortogonal especial SO(m).*

Dada a Definição 2.10, considere abaixo.

Definição 2.11. *A **forma** de uma matriz de configuração $\boldsymbol{X}$ é toda informação geométrica acerca de $\boldsymbol{X}$ que é invariante sob locação, escala e rotação. A forma pode ser representada pelo conjunto $[\boldsymbol{X}]$ dado por*

$$[\boldsymbol{X}] = \{\boldsymbol{Z}\boldsymbol{\Gamma} : \boldsymbol{\Gamma} \in SO(m)\},$$

em que SO(m) é o grupo ortogonal especial de rotações e $\boldsymbol{Z}$ é a pré-forma de $\boldsymbol{X}$.

Nós usamos a notação Σ_m^k para denotar o espaço de forma de k pontos em m dimensões.

Definição 2.12. *O **espaço de forma** é o espaço de todas as formas. Formalmente, o espaço de forma Σ_m^k é o espaço em órbita das configurações não conincidentes do conjunto de pontos k em $\mathbb{R}^m$ sob os efeitos de translação, rotação e escala.*

A dimensão do espaço de forma é

$$q = km - m - 1 - \frac{m(m-1)}{2},$$

e isso pode ser visto de modo que, inicialmente, tínhamos $k \times m$ coordenadas e, então, perdemos m dimensões para locação, uma dimensão para a escala e $\frac{1}{2}m(m-1)$ para rotação.

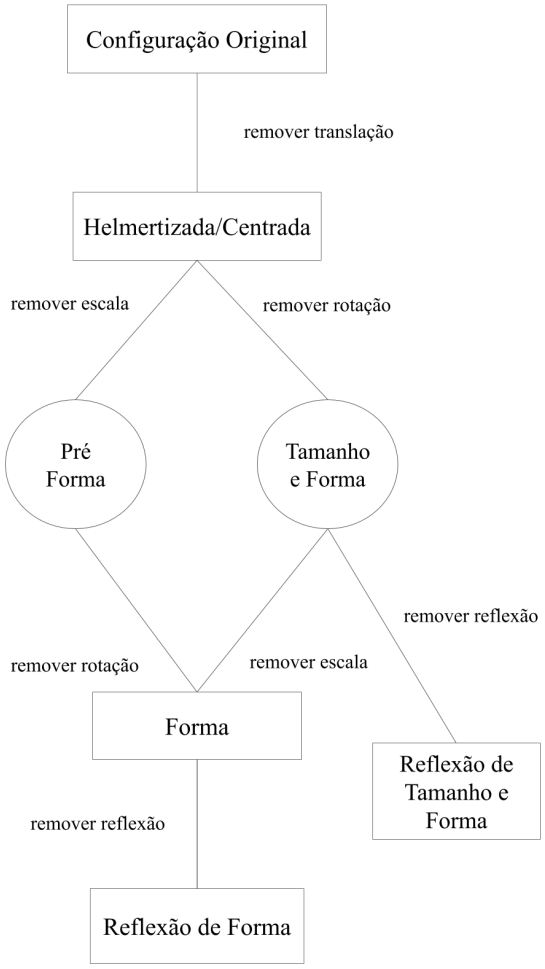


Figura 2 – A hierarquia dos vários espaços (DRYDEN, 2016)

Outro tipo de efeito que também podemos incluir para ser retirado é o de reflexão para forma (DRYDEN, 2016, página 67), representado pelo conjunto $[\boldsymbol{X}]_R$. A reflexão pode ser obtida multiplicando um dos eixos das coordenadas por -1. Rotações e reflexões podem ser representadas por uma matriz ortogonal de uma maneira conjunta.

Definição 2.13. *Uma* ***matriz ortogonal*** $m \times m$ *satisfaz* $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{I}_m$ *e* $det(\boldsymbol{\Gamma}) = \pm 1$. *O conjunto de todas as matrizes ortogonais* $m \times m$ *é conhecido como grupo ortogonal* $O(m)$.

O grupo ortogonal inclui rotações (determinante +1) e rotações/reflexões (determinante -1). Em (DRYDEN, 2016, página 67), define-se a forma de reflexão de modo análogo à Definição 2.1 mas, agora, com uma matriz $\boldsymbol{R}$ que pertence ao grupo ortogonal $O(m)$.

A Figura 2 exemplifica a hierarquia dos espaços e as ações que aplicamos para transitar entre eles.

2.2 DISTÂNCIAS DO ESPAÇO DE FORMAS

2.2.1 Variedades Diferenciáveis

Na seção anterior, foi mencionado o fato do espaço de pré-forma S_m^k ser uma hiperesfera de raio unitário em $(k-1)m$ dimensões reais. Um conceito chave na interpretação e construção de medidas de distância, por exemplo, dentro desses espaços é o de variedade. (SMALL, 1996, página 38) comenta que uma variedade é a generalização do nosso entendimento de uma superfície curvada em 3 dimensões. No nosso caso, estamos interessados em variedades diferenciáveis (LIMA, 2015, apêndice A), em particular, as variedades de Riemann (DRYDEN, 2016, página 59). Um toro e uma esfera são exemplos comuns de variedades diferenciáveis. Isso quer dizer que as matrizes pertencentes ao espaço de pré-forma S_m^k podem ser consideradas variedades diferenciáveis. Uma variedade também pode ser definida como um espaço topológico (LIMA, 2015, página 99) que pode ser visto, localmente, como um espaço Euclidiano.

Imaginemos que uma criatura pequena que vive em duas dimensões, será colocada na superfície de um toro. Essa criatura terá sérios problemas em distinguir se a superfície na qual foi alocada é um toro, uma esfera ou um plano, por exemplo. Isso acontece porque superfícies curvadas parecem, aproximadamente, planas quando observadas sobre uma região pequena. A vizinhança imediata da criatura fornece informações locais sobre a superfície, mas pouco sobre as propriedades globais da superfície que, por sua vez, a distinguem das esferas e dos planos. Para encontrar informações globais, a criatura necessitaria percorrer ambas as superfícies para verificar ângulos e distâncias. Caso não seja possível uma inspeção ao longo da superfície por completo, toda informação e examinação dos aspectos locais da superfície não iriam ajudar na caracterização da curvatura da superfície. É isso que nós queremos dizer quando falamos que uma variedade diferencial assemelha-se, localmente, ao $\mathbb{R}^p$, como discutido em (SMALL, 1996, capítulo 2.2). A falta de entendimento das características globais associadas a uma superfície em questão, podem levar a inferências errôneas quando apenas observadas informações e características locais. Na literatura romântica, (ABBOTT, 1884) utiliza os elementos geométricos para realizar uma crítica social à sociedade da época e também ao entendimento que os seres criados

por ele possuem acerca do espaço em que vivem. No contexto da Análise Estatística de Formas, o entendimento do espaço de pré-forma S_m^k como uma variedade de Riemann é essencial para definirmos medidas de distância entre dois pontos, por exemplo, numa superfície que não é um plano Euclidiano.

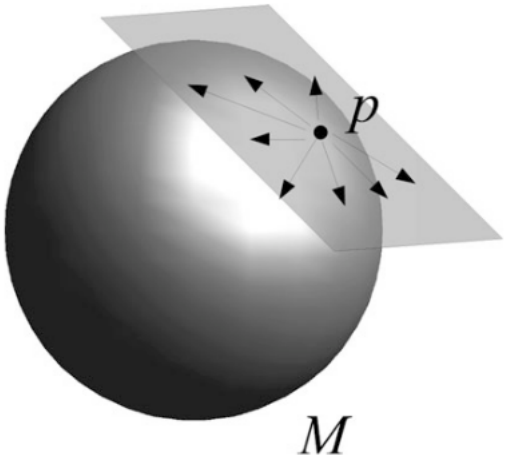


Figura 3 – Espaço tangente de $\boldsymbol{M}$ em p (A.; KLASSEN, 2016)

Primeiro, consideramos espaços tangentes para uma variedade geral $\boldsymbol{M}$. Considere uma curva diferenciável em $\boldsymbol{M}$ dada por $\gamma(t) \in \boldsymbol{M}, \boldsymbol{t} \in \mathbb{R}$ com $\gamma(0) = p$. O vetor tangente no ponto p é dado por

$$\gamma'(0) = \lim_{t \longrightarrow 0} \frac{\mathrm{d}\gamma}{\mathrm{d}t},$$

e o vetor unitário é $\xi = \gamma'(0)/\|\gamma'(0)\|$. O conjunto de todos os vetores tangentes para todas as curvas que passam através do ponto p é chamado de **espaço tangente** de $\boldsymbol{M}$ em p, denotado por $T_p(\boldsymbol{M})$. Na Figura 3 temos uma exemplificação de um possível $T_p(\boldsymbol{M})$. Uma variedade de Riemann $\boldsymbol{M}$ é uma variedade conectada que possui um produto interno positivo-definido em cada espaço tangente $T_p(\boldsymbol{M})$, tal que a escolha varia suavemente de ponto para ponto. Considere $g = g_{ij}$ como o tensor (GOODFELLOW; BENGIO; COURVILLE, 2016, página 31) positivo-definido que define o produto interno em cada espaço tangente dado um sistema de coordenadas. De modo que, tendo as coordenadas $(x_1, \ldots, x_n)$, a métrica nesse espaço é

$$ds^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} dx_i dx_j.$$

A distância de Riemann entre dois pontos numa variedade de Riemann é dada pelo comprimento de arco que minimiza o geodésico entre dois pontos, em que o comprimento de uma curva parametrizada $\gamma(t) \in [a, b]$ é definido como

$$L = \int_a^b \|\gamma'(t)\|_g \mathrm{d}t. \tag{2.5}$$

Podemos notar a associação dessa definição com o direcionamento que usamos na disciplina de Cálculo Diferencial e Integral para o cálculo da integral de linha relativa ao comprimento de arco (GUIDORIZZI, 2002).

2.2.2 Distância de Procrustes

Considere duas matrizes de configurações de k pontos em m dimensões, $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$, com as pré-formas $\boldsymbol{Z}_1$ e $\boldsymbol{Z}_2$. Temos a seguir, as definições das distâncias comumente utilizadas no contexto da Análise Estatística de Formas.

Definição 2.14. *A distância total de Procrustes entre $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ é*

$$d_F(\boldsymbol{X}_1, \boldsymbol{X}_2) = \inf_{\boldsymbol{\Gamma} \in SO(m), \beta \in \mathbb{R}^+} \|\boldsymbol{Z}_2 - \beta \boldsymbol{Z}_1 \boldsymbol{\Gamma}\|,$$

em que $\boldsymbol{Z}_r = \boldsymbol{H}\boldsymbol{X}_r / \|\boldsymbol{H}\boldsymbol{X}_r\|, r = 1, 2.$

Resultado 2.1. *A distância total de Procrustes é*

$$d_F(\boldsymbol{X}_1, \boldsymbol{X}_2) = \left\{ 1 - \left(\sum_{i=1}^{m} \lambda_i \right)^2 \right\}^{1/2},$$

em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m-1} \geq |\lambda_m|$ são as raízes quadradas dos autovalores de $\boldsymbol{Z}_1^T \boldsymbol{Z}_2 \boldsymbol{Z}_2^T \boldsymbol{Z}_1$ e o menor valor λ_m é o negativo da raíz quadrada se e somente se $det(\boldsymbol{Z}_1^T \boldsymbol{Z}_2) < 0$.

Em (DRYDEN, 2016, capítulo 8), é discutido a distância de *Procrustes* para o caso de dados bidimensionais, os quais passam a ser considerados como um vetor complexo de tamanho k, e não uma matriz de configuração real de dimensão $k \times m$. Mesmo considerando o caso em que $m = 2$, essa seção é bastante útil, pois ela permite que visualizemos, de um modo mais simples, que a distância de *Procrustes* é um resíduo "studentizado". Esse resíduo é obtido a partir de um ajuste de uma regressão linear complexa de w em y, em que w e y são vetores complexos de tamanho k.

2.2.3 Distância Riemanniana

Resultado 2.2. *A distância Riemanniana ρ é*

$$\rho(\boldsymbol{X}_1, \boldsymbol{X}_2) = arccos\left(\sum_{i=1}^{m} \lambda_i\right),$$

em que o autovalor $\lambda_i, i = 1, \ldots, m$ são definidos no Resultado 2.1.

As provas para os resultados apresentados nessa seção podem ser encontrados em (DRYDEN, 2016, capítulo 4). A seguir, temos uma tabela que resume as medidas de distância no espaço de formas e os respectivos intervalos nos quais assumem valor.

Distância	Notação	Fórmula	Suporte
Distância Total de Procrustes	d_F	$\{1 - (\sum_{i=1}^{m} \lambda_i)^2\}^{1/2}$	$0 \leq d_F \leq 1$
Distância Riemanniana	ρ	$\arccos(\sum_{i=1}^{m} \lambda_i)$	$0 \leq \rho \leq \pi/2$

Tabela 1 – Distâncias no espaço de formas Σ_m^k

Temos a seguinte relação entre as distâncias definidas anteriormente:

$$d_F(\boldsymbol{X}_1, \boldsymbol{X}_2) = sen(\rho). \tag{2.6}$$

Como pontuado em (VINUé; SIMO; ALEMANY, 2014, página 108), a distância Riemanniana ρ é uma métrica Riemmaniana (ALEXANDRINO, 2018; LIMA, 2015, página 50), mas a distância total de Procrustes d_F não é uma métrica Riemanniana no espaço de forma. Entretanto, as duas distâncias são topologicamente equivalentes. Ainda, para formas que estão próximas, há pouca diferença entre as distâncias, de modo que

$$\rho = d_F + O(d_F^3).$$

2.3 FORMA MÉDIA

Em geral, como discutido nas seções anteriores deste capítulo, nosso interesse é mensurar a variabilidade entre dois ou mais objetos em relação as respectivas formas. O conceito de forma média desempenha um papel importante para a análise da variabilidade de formas.

2.3.1 Forma Média de Procrustes

Considere uma amostra de configurações $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$.

Definição 2.15. *A **forma média amostral total de Procrustes** é*

$$\hat{\boldsymbol{\mu}}_F = arg\ inf_{\mu} \frac{1}{n} \sum_{i=1}^{n} d_F^2(\boldsymbol{X}_i, \boldsymbol{\mu}).$$

A forma média da Definição 2.15 é considerada uma média extrínseca. Isso quer dizer que a distância utilizada para o seu cálculo não é o menor comprimento de arco em uma variedade, como já comentado neste capítulo. Para dados bidimensionais, existe um autovetor de solução explícita do problema de otmização na definição da forma média de *Procrustes* (DRYDEN, 2016, página 178). Mas para $m \geq 3$, um processo iterativo deve ser usado para obter a forma média. Considere o seguinte algoritmo para a obtenção de tal quantidade.

Algoritmo 1: Algoritmo para o cálculo de $\hat{\boldsymbol{\mu}}_F$

---

1. **Translações**. Centralize as configurações para remover a locação. Inicialmente, considere

$$\boldsymbol{X}_i^P = \boldsymbol{C}\boldsymbol{X}_i, i = 1, \ldots, n,$$

em que $\boldsymbol{C}$ é a matriz de centralização da Equação 2.2.

2. **Rotações**. Calcule $G = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|\boldsymbol{X}_i^P - \boldsymbol{X}_j^P\|^2$.
Para a i-ésima configuração, seja

$$\bar{\boldsymbol{X}}_{(i)} = \frac{1}{n-1} \sum_{j \neq i} \boldsymbol{X}_j^P.$$

Otimize $\|\bar{\boldsymbol{X}}_{(i)} - \boldsymbol{X}_i^P \boldsymbol{\Gamma}\|^2$ por rotações. Considere $\boldsymbol{X}_i^P = \boldsymbol{X}_i^P \hat{\Gamma}$ como a matriz de rotação ótima. Repita para todo i. Calcule o novo valor de G.
O processo é repetido até G não poder ser mais reduzido.

3. **Escala**. Para a i-ésima configuração, calcule

$$\hat{\boldsymbol{\beta}}_i = \left(\frac{\sum_{k=1}^{n} \|\boldsymbol{X}_k^P\|^2}{\|\boldsymbol{X}_i^P\|^2} \right)^{1/2} \phi,$$

em que ϕ é o i-ésimo componente do autovetor ϕ correspondente ao maior autovalor da matriz de correlação Φ de $vec\{\boldsymbol{X}_i^P\}$.
Seja $\boldsymbol{X}_i^P = \hat{\boldsymbol{\beta}}_i \boldsymbol{X}_i^P$. Repita para todo i. Calcule o novo valor de G.

4. Repita os passos 2 e 3 até G não pode ser mais reduzido.

5. $[\hat{\boldsymbol{\mu}}] = \frac{1}{n} \sum_i^n \boldsymbol{X}_i^P$.

2.3.2 Forma Média ϕ

A obtenção da forma média, em $m = 3$, pelo método de *Procrustes*, definido no Algoritmo 1, é um método iterativo. (DRYDEN *et al.*, 2008) introduziu a forma média ϕ, como uma alternativa aos métodos iterativos para inferências de formas quando $m \geq 3$. Essa forma média é particularmente útil quando o tamanho da amostra, n, ou o número de *landmarks*, k, é grande.

Dado uma pré-forma $\boldsymbol{Z} \in S_m^k$, trabalhos com $\boldsymbol{Z}\boldsymbol{Z}^T$, a qual é invariante em relação à rotação e reflexão. Podemos observar isso, diretamente, uma vez que para uma matriz $\boldsymbol{R}$ pertencente ao grupo ortogonal $O(m)$, $(\boldsymbol{Z}\boldsymbol{R})(\boldsymbol{Z}\boldsymbol{R})^T = \boldsymbol{Z}\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{Z}^T = \boldsymbol{Z}\boldsymbol{Z}^T$, pela Definição 2.13. Pode-se facilmente reverter, se necessário, a forma quadrática $\boldsymbol{Z}\boldsymbol{Z}^T$ para um representante em $[\boldsymbol{X}]_R$, usando o teorema da decomposição espectral: $\boldsymbol{Z}\boldsymbol{Z}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \sum_{i=1}^m \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$, em que $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m, 0, \ldots, 0)$ é a matriz diagonal dos autovalores de $\boldsymbol{Z}\boldsymbol{Z}^T$ em ordem descendente, e as colunas $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1}$ de $\boldsymbol{U}$ são os autovetores correspondentes. Assim $(\lambda_1^{1/2}\boldsymbol{u}_1^T, \ldots, \lambda_m^{1/2}\boldsymbol{u}_m^T)^T$ é uma pré-forma representativa do conjunto $[\boldsymbol{X}]_R$.

Considere o espaço $P_m(k-1)$ das matrizes simétricas $(k-1)\times(k-1)$ diferentes de zero, de posto igual, no máximo, a m e traço igual a 1. Assim,

$$P_m(k-1) = \{\boldsymbol{P} \in P(k-1) | 1 \leq \text{posto}(\boldsymbol{P}) \leq m,\ \text{tr}(\boldsymbol{P}) = 1\}, \tag{2.7}$$

em que $P(k)$ é o espaço das matrizes reais simétricas positivas semi-definidas, de dimensão k. Para uma pré-forma aleatória $\boldsymbol{Z} \in S_m^k$ com uma distribuição arbitrária da população F, temos

$$\boldsymbol{\Xi}(F) = E_F(\boldsymbol{Z}\boldsymbol{Z}^T) = \int_{\boldsymbol{Z}\in S_m^k} \boldsymbol{Z}\boldsymbol{Z}^T \mathrm{d}F(\boldsymbol{Z}), \tag{2.8}$$

como a esperança de $\boldsymbol{Z}\boldsymbol{Z}^T$ com respeito a F. Suponha $\boldsymbol{\Xi} = \boldsymbol{\Xi}(F)$ possui decomposição espectral $\boldsymbol{\Xi} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, em que $\Delta = \text{diag}(\delta_1, \ldots, \delta_{k-1})$ consiste dos autovalores ordenados $\delta_1 \geq \cdots \geq \delta_{k-1} \geq 0$ de $\boldsymbol{\Xi}$ e as colunas de $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-1})$ são os autovetores unitários correspondentes. Com isso, define-se

$$\phi(\boldsymbol{\Xi}) = \frac{1}{\delta_1 + \cdots + \delta_m} \sum_{i=1}^m \delta_i \boldsymbol{u}_i \boldsymbol{u}_i^T. \tag{2.9}$$

Nós nos referimos à $\phi(\boldsymbol{\Xi})$ como forma média ϕ de $\boldsymbol{Z}$, a qual é definida, exclusivamente, desde $\delta_m > \delta_{m+1}$. Dada uma amostra $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, a forma média amostral ϕ é definida por $\phi(\hat{\boldsymbol{\Xi}})$, em que $\hat{\boldsymbol{\Xi}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{Z}\boldsymbol{Z}^T$ é o análogo da amostra de $\boldsymbol{\Xi}$. (PRESTON; WOOD, 2010) resume o procedimento da forma média ϕ da seguinte maneira. Começando com uma matriz de configuração $\boldsymbol{X}$, $k \times m$, depois removendo-se os efeitos de locação, escala e definindo $\boldsymbol{Z}\boldsymbol{Z}^T$ que remove os efeitos de rotação e reflexão mas retém a forma. Conclui-se que existe um mapeamento um-a-um entre as possíveis formas de uma matriz de configuração $\boldsymbol{X}$ e os elementos do conjunto $P_m(k-1)$.

3 ALGORITMO *K-MEANS*

A área de aprendizado de máquina é divida em duas partes: aprendizado supervisionado e aprendizado não supervisionado, como descrito em (JAMES *et al.*, 2014). Dentro da sub-divisão de aprendizado supervisionado, temos o seguinte cenário: para cada observação das variáveis explicativas $x_i, i = 1, \ldots, n$, há uma variável resposta associada y_i. Desejamos ajustar um modelo que relacione a resposta à preditores, com o objetivo de prever, com certa precisão, a resposta para futuras observações ou melhor compreensão da relação entre a resposta e os preditores. Já na sub-divisão de aprendizado não-supervisionado, temos uma situação mais desafiadora, na qual para toda observação $i = 1, \ldots, n$, nós observamos um vetor de valores x_i, mas que não está associado a uma resposta y_i. Dessa forma, nós estamos vendados quando lidamos com problemas dessa natureza. Essa situação é chamada de não-supervisionada porque nos falta uma variável resposta capaz de supervisionar nossa análise. (JAMES *et al.*, 2014, página 373) pontua os principais métodos utilizados dentro do aprendizado não-supervisionado e as suas características.

Dentro desse cenário não-supervisionado, não estamos interessados em predição, uma vez que não temos uma variável resposta associada. O objetivo é descobrir padrões e informações que as variáveis observadas $x_1, x_2, \ldots, x_p$ possuem. No nosso caso, nós pretendemos atacar o problema de descobrir subgrupos das observações, a partir das variáveis de interesse. Para esse problema, consideramos o algoritmo *K-means*.

3.1 VISÃO GERAL

O algoritmo *K-means* é uma vertente para realizar o agrupamento dos dados, ou seja, o particionamento de um conjunto de dados em K conjuntos distintos. Para proceder com o agrupamento, deve-se especificar o número de grupos K, então o algoritmo irá assinalar cada observação para algum dos K grupos definidos. Como descrito em (HASTIE; TIBSHIRANI; FRIEDMAN, 2001), o algoritmo de agrupamento *K-means* resulta de um procedimento simples, a partir de um problema de otimização. Considere $C_1, \ldots, C_k$ como os conjuntos que possuem as observações assinaladas aos grupos, respectivamente. Esses conjuntos satisfazem duas propriedades:

1. $C_1 \cup C_2 \cup \cdots \cup C_k = 1, \ldots, n$. Ou seja, cada observação pertence a pelo menos um dos K grupos.

2. $C_k \cap C_{k'} = \emptyset \ \forall k \neq k'$. Ou seja, os grupos não possuem sobreposição. Logo, nenhuma observação pertence a mais de um grupo.

Por exemplo, se a i-ésima observação está no k-ésimo grupo, então $i \in C_k$. O objetivo do *K-means* é realizar uma separação de modo que a variabilidade entre as observações de um mesmo grupo seja a menor possível. A variação intra-grupo, para um grupo C_k, é uma medida $W(C_k)$ que representa, numericamente, o quanto as observações dentro daquele grupo se diferenciam entre si. Assim, como mencionado no início deste capítulo, nós desejamos resolver o seguinte problema de otimização

$$\underset{C_i, i=1,\ldots,K}{\textbf{minimizar}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}. \tag{3.1}$$

Esta fórmula define o direcionamento de particionar as observações em K grupos, tal que a variação total intra-grupo, somado todos os grupos, seja a menor possível. Quando lidamos com o problema de agrupar observações baseando-se em alguma medida, existem dois tipos, usualmente, de medidas para isso: medidas de similaridade e medidas de dissimilaridade. (GOSHTASBY, 2012, capítulo 2) define que uma medida de similaridade S é considerada uma métrica se produzir um valor maior à medida que a dependência entre os valores correspondentes das observações aumenta. Já uma medida de dissimilaridade S é uma métrica se produzir um valor menor à medida que a dependência entre os valores correspondentes das observações diminui. Esses valores que nos referimos podem ser o valor do nível de cinza dos *pixels* das imagens comparadas. Como, também, o preço de dois imóveis localizados em uma cidade. Temos, como exemplo de uma medida de similaridade, o Coeficiente de Correlação de Pearson (GOSHTASBY, 2012, página 9) e como uma medida de dissimilaridade, o quadrado da distância Euclidiana (GOSHTASBY, 2012, página 34).

Para definir a variação $W(C_k)$ na Equação 3.1, consideramos o quadrado da distância Euclidiana, que é uma medida de dissimilaridade. Logo, quanto maior for o valor da distância entre duas observações, menos similares elas são entre si. Dessa forma, considere

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{3.2}$$

em que $|C_k|$ denota o número de observações no k-ésimo grupo. A variação intra-grupo para

o k-ésimo grupo é a soma de todos os quadrados da distância Euclidiana das observações dois-a-dois do k-ésimo grupo, controlada pelo número de observações no grupo. A partir das Equações 3.1 e 3.2, o problema de otimização de agrupamento do *K-means* pode ser definido como

$$\underset{C_1,\ldots,C_k}{\text{minimizar}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \tag{3.3}$$

Temos a seguir o algoritmo para o ajuste do *K-means*,

Algoritmo 2: Algoritmo para o ajuste do *K-means*

1. Escolher aleatoriamente K observações para serem os centroides iniciais dos K grupos.
2. Itere até as observações não mudarem mais de grupos ou ao atingir uma cota inferior de erro:
   (a) Assinale cada observação ao grupo no qual o centroide é o mais próximo (o sentido de mais próximo é definido usando a distância Euclidiana).
   (b) Para cada um dos K grupos, calcule o seu centroide. O centroide do k-ésimo grupo é o vetor de médias das p variáveis de cada observação no grupo k.

(JAMES *et al.*, 2014) descreve a convergência do algoritmo, de modo que a função objetivo nunca irá aumentar seu valor dentro das iterações. Ainda, como o algoritmo *K-means* determina um ponto ótimo local ao invés de um global, os resultados obtidos irão depender do chute inicial dado. Ou seja, o resultado da separação depende de quais K observações foram, aleatoriamente, selecionadas para serem os centroides iniciais dos K grupos. Com isso em mente, o Algoritmo 2 deve ser executado várias vezes, cada uma delas obtendo, possivelmente, um chute inicial diferente para os centroides iniciais. Dessa maneira, nós selecionamos a iteração que obteve o menor valor em relação à função objetivo da Equação 3.3.

3.2 NO CONTEXTO DE FORMAS

O desenvolvimento do algoritmo *K-means* na área da Análise Estatística de Formas é, de certo modo, recente. Os primeiros trabalhos e desenvolvimentos foram propostos por (GEORGESCU, 2009; AMARAL *et al.*, 2010). Em (GEORGESCU, 2009), o autor usa um tipo de algoritmo *K-means* para agrupar formas *fuzzy*. Já em (AMARAL *et al.*, 2010), o autor adapta o algoritmo *K-means* de Hartigan-Wong. Ambos

os trabalhos, utilizam dados de formas bidimensionais e com um número de *landmarks* $k = 50, 11$, respectivamente. Para dados de formas tridimensionais, existe apenas o trabalho desenvolvido em (VINUé; SIMO; ALEMANY, 2014). Esse trabalho é o pioneiro no desenvolvimento do algoritmo *K-means* para formas em 3 dimensões. Os autores realizaram uma aplicação com dados reais das medidas antropométricas de um conjunto de mulheres na Espanha.

Para os dados de forma, devemos fazer apenas algumas alterações no Algoritmo 2. As observações, agora, consistem em uma matriz de configuração $\boldsymbol{X}$ de dimensão $k \times 3$. No processo iterativo, realiza-se o cálculo de uma medida de dissimilaridade, além de calcular a forma média para atualizar o centroide a cada iteração. Para este contexto, o modo como calculamos essas quantidades devem mudar. Para o cálculo da distância, usamos a distância Riemanniana, definida no Resultado 2.2. E para a atualização dos centroides, usaremos a forma média de *Procrustes*, Definição 2.15, e a forma média ϕ, Equação 2.9. Sendo assim, podemos considerar a seguinte modificação para o algoritmo

Algoritmo 3: Algoritmo para o ajuste do *K-means* no contexto de formas

1. Escolher aleatoriamente K matrizes de configuração (observações) $\boldsymbol{X}_i$ para serem os centroides iniciais dos K grupos.
2. Itere até as observações não mudarem mais de grupos ou ao atingir uma cota inferior de erro:
   (a) Assinale cada observação ao grupo no qual o centroide é o mais próximo (o sentido de mais próximo é definido usando a distância Riemanniana ρ).
   (b) Para cada um dos K grupos, calcule o seu centroide. O centroide do k-ésimo grupo é a forma média das matrizes de configuração do grupo k.

Apesar do algoritmo proposto em (HARTIGAN; WONG, 1979) apresentar uma versão mais eficiente do *K-means*, (VINUé; SIMO; ALEMANY, 2014) apresentou resultados que demonstram que, para um tamanho de amostra pequeno, o tempo computacional do algoritmo pela versão Hartigan-Wong foi um pouco maior do que o de Lloyd. Entretanto, a versão de Hartigan-Wong teve uma taxa de alocação melhor do que a versão do Lloyd, considerando $k = 34$. Quando o tamanho da amostra aumentou, o algoritmo de Hartigan-Wong teve um aumento do tempo computacional e uma piora da taxa de alocação. Concluindo-se que para o caso de formas tridimensionais, com um número grande de *landmarks* e um tamanho de amostra médio ou grande (em relação às aplicações comumente utilizadas), a versão de Hartigan-Wong será inoperável computacionalmente. Desse modo, consideramos a versão de Lloyd (Algoritmo 3) para a realização do agrupa-

mento. No Capítulo 4, comparamos a performance do algoritmo *K-means* no contexto de formas em relação a escolha da forma média para o cálculo dos centroides.

4 RESULTADOS NUMÉRICOS

Neste capítulo, nós utilizamos os métodos discorridos nos capítulos anteriores em uma aplicação com dados reais. O conjunto de dados em questão (DRYDEN *et al.*, 2008) foi disponibilizado por Ian Dryden, professor de Estatística da Universidade de Nottingham. O conjunto de dados consiste em um número grande de pseudo-*landmarks*, $k = 62.501$, localizados na superfície cerebral com $m = 3$. No total, são 74 matrizes de configuração $\boldsymbol{X}_i, i = 1, \ldots, 74$. A Figura 4 mostra a superfície cerebral de um dos pacientes com o número total de *landmarks*, $k = 62.501$. Além das matrizes de configuração, os dados possuem as seguintes variáveis:

- Diagnóstico: variável indicadora com dois níveis para indicar o diagnóstico de esquizofrenia;
- Destreza: variável indicadora com dois níveis para indicar se o paciente é destro ou canhoto;
- Sexo: variável indicadora com dois níveis para indicar o sexo do paciente;
- Idade: variável numérica associada aos anos de vida do paciente.

Dentre os 74 pacientes, temos a seguinte divisão:44 são do grupo controle e 30 possuem o diagnóstico de esquizofrenia. Os pacientes com o diagnóstico da doença possuem uma assimetria que é chamada de torque. (BILDER *et al.*, 1994) e (MACKAY *et al.*, 2003) estudaram este tipo de assimetria e puderam constatar que existe diferença significativa na forma do córtex cerebral, entre os pacientes do grupo controle e esquizofrenia.

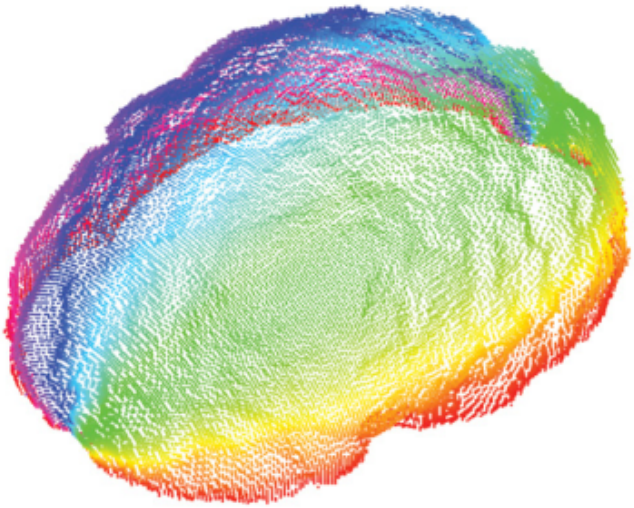


Figura 4 – Superfície cerebral representada por 62.501 *landmarks* (DRYDEN, 2016)

4.1 CÁLCULO DA FORMA MÉDIA

Na seção 2.3, foram discutidos os direcionamentos para o cálculo da forma média pelo método de *Procrustes* (Definição 2.15) e a forma média ϕ (Equação 2.9). Como apresentado, a forma média ϕ possui um estimador de forma fechada, de modo que não é necessário o uso de métodos iterativos para sua computação. (DRYDEN *et al.*, 2008) discorre este resultado como uma alternativa ao método de *Procrustes*, que é, por sua vez, um método iterativo. Nesta seção, realizamos um *benchmark* para comparar a performance, em relação ao tempo de processamento computacional, do cálculo da forma média utilizando os dois métodos mencionados.

Considere múltiplas amostragens sistemáticas (COCHRAN, 1977) ao conjunto de dados dos córtex cerebrais, com $k = 1.300$, $k = 1.000$, $k = 700$, $k = 400$, $k = 100$ e $k = 40$. Para cada valor que k assume, foram realizadas 100 repetições. A Figura 5 nos mostra o tempo médio, em segundos, do cálculo da forma média por ambos métodos e a diferença no tempo de processamento é nítida. A forma média ϕ possui um tempo de execução menor do que a forma média de *Procrustes*. A variação dos valores de k nos ajuda a visualizar o cenário de ganho do desempenho computacional na escolha da forma média ϕ. Para o caso em que $k = 1.300$, o ganho de performance computacional foi de, aproximadamente, 77%. E, para o caso em que $k = 40$, foi de, aproximadamente, 87%. Aqui, o uso da amostragem sistemática tem como objetivo ajudar na comparação do tempo de processamento da forma média pelos diferentes métodos. Com isso, conseguimos visualizar a performance em diferentes cenários em relação à dimensão das matrizes de configuração.

Como discutido em (DRYDEN *et al.*, 2008), a distância Riemanniana (Resultado 2.2) entre a forma média de todas as observações calculada pelo método de *Procrustes* e forma média ϕ foi igual a 0,00023. Para o grupo de controle, essa diferença foi igual a 0,00023, e 0,00024 para o grupo de pacientes com a doença. Desse modo, como a diferença é pequena, não há diferenças visíveis entre as formas médias calculadas por cada método.

(a) (b)

(c) (d)

(e) (f)

Figura 5 – Simulação do tempo de processamento da forma média: ϕ e *Procrustes*. (a) $k = 1.300$, (b) $k = 1.000$, (c) $k = 700$, (d) $k = 400$, (e) $k = 100$ e (f) $k = 40$

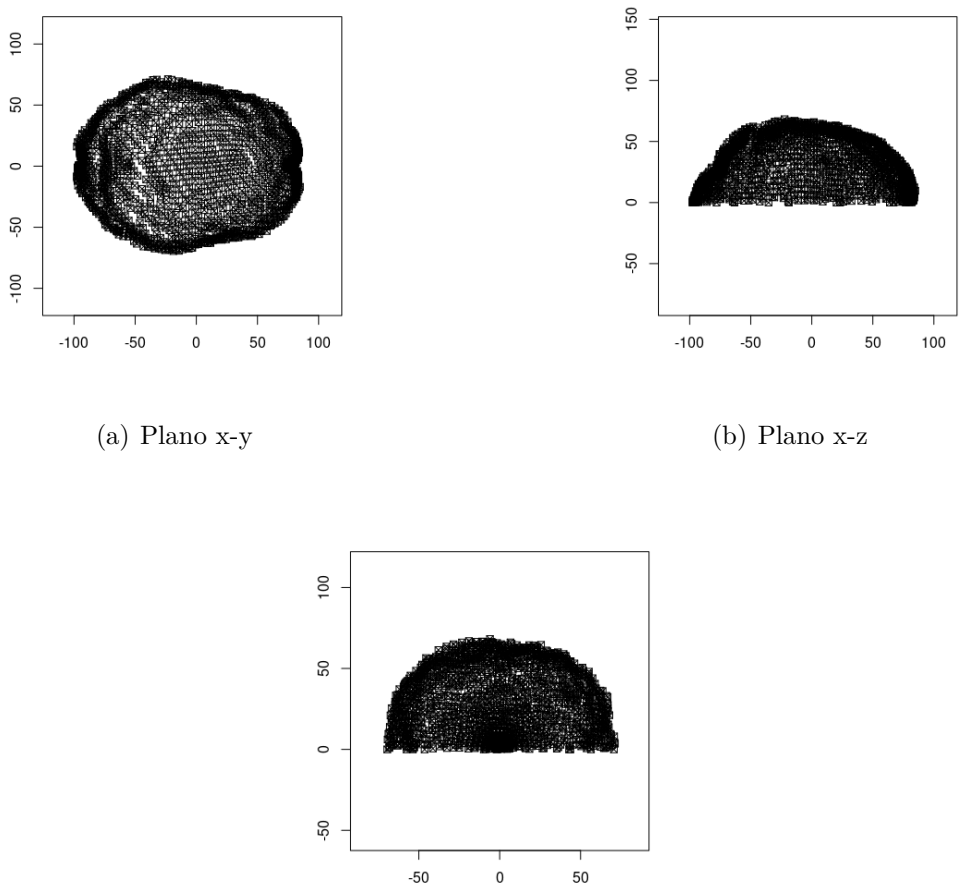
4.2 AJUSTE DO *K-MEANS*

4.2.1 Cenário

Nosso objetivo é agrupar a forma do córtex cerebral dentre os pacientes com e sem o diagnóstico de esquizofrenia. Para a realização do agrupamento, consideramos uma amostra sistemática (COCHRAN, 1977) dos *landmarks* em cada matriz de configuração. Esse direcionamento foi necessário para tornar possível a alocação de recursos computacionais durante os cálculos. Este conjunto de dados possui um número de *landmarks* bastante superior ao número considerado em grande parte das aplicações na Análise Estatística de Formas e, em particular, os trabalhos de (AMARAL *et al.*, 2010; GEORGESCU, 2009; VINUé; SIMO; ALEMANY, 2014) utilizaram 11, 50 e 95 *landmarks*, respectivamente, no desenvolvimento de algoritmos de agrupamento.

Na amostragem sistemática, foram selecionados $k = 1.300$ *landmarks*. Devemos notar que, aqui, os *landmarks* em questão são rotulados, como descrito na Definição 2.4. Este fato é relevante, pois sem ele não seríamos capazes de realizar a comparação entre as matrizes de configuração durante o agrupamento. Uma vez que não seria, digamos, consistente, comparar *landmarks* que não são correspondentes. Na seção anterior, como não tínhamos, estritamente, uma necessidade de correspondência, esse fator não seria necessariamente um problema. Visto que ainda seríamos capazes de identificar o comportamento dos métodos quando avaliados sob dimensões das matrizes de configuração diferentes. Na Figura 6, temos a visualização de um dos córtex cerebrais resultantes após a amostragem. Pode-se visualizar o objeto pela perspectiva dos três possíveis planos no $\mathbb{R}^2$.

O agrupamento *K-means* foi executado considerando o particionamento das observações em 2 grupos, com um número de inicializações aleatórias igual a 100 e um número iterações igual a 10, para cada inicialização. A máquina utilizada possuía 16GB de memória RAM e um processador com 8 núcleos. O algoritmo foi adaptado, de modo que as inicializações aleatórias foram divididas entre os núcleos para serem executadas em paralelo, a partir dos pacotes **parallel** (R Core Team, 2020), **foreach** e **doSNOW** (Microsoft; WESTON, 2017). O advento da paralelização foi possível, aqui, pois temos o cenário no qual cada inicialização aleatória não depende das outras para seu processamento. Sendo assim, executamos as inicializações dessa maneira e, ao término do processamento, buscamos identificar qual obteve o melhor desempenho de acordo com os requisitos



(a) Plano x-y

(b) Plano x-z

(c) Plano y-z

Figura 6 – Superfície cerebral representada por 1.300 *landmarks* com os planos das 3 dimensões do espaço

definidos.

O algoritmo foi executado considerando tanto a forma média ϕ, quanto o método de *Procrustes*, para o cálculo dos centroides dos grupos durante o processamento. Como resultado, pôde-se observar os seguintes tempos de processamento: 1:35, uma hora e trinta e cinco minutos e 12:50, doze horas e cinquenta minutos, para o caso da forma média ϕ e *Procrustes*, respectivamente. O algoritmo executou mais rápido quando utilizamos a forma média ϕ para o cálculo dos centroides. Ou seja, o desempenho que observamos na seção anterior, foi observado também no algoritmo *K-means*. Os chutes iniciais para os casos ótimos foram (4, 37) e (41, 56), para ϕ e *Procrustes*, respectivamente. Mesmo assim, ao término do processamento, ambos os cenários convergiram para os mesmos grupos, com 37 observações cada. Como comentado anteriormente, a diferença numérica entre o cálculo da forma média pelos métodos é pequena. Consequentemente, a similaridade das

convergências tende a ser alta.

4.2.2 Qualidade do Agrupamento

Embora o algoritmo *K-means* esteja inserido na classe de métodos não-supervisionados, considere a Tabela 2 como uma matriz de confusão do agrupamento em relação ao diagnóstico do paciente. Essa matriz nos permite visualizar indícios da existência de uma associação entre o valor verdadeiro da variável, neste caso, temos o diagnóstico da doença, com os grupos que o algoritmo determinou. O mapeamento ideal, em relação à classificação, é um cenário em que apenas a diagonal principal dessa matriz assume valores não nulos. De modo que, cada grupo determinado pelo algoritmo possuiria todos os elementos de um certo nível da variável de interesse. Dado os valores na Tabela 2, não há diferenças visíveis na classificação em relação ao agrupamento.

Tabela 2 – Matriz de confusão do agrupamento

Diagnosticado	**Agrupamento**	
	Grupo 1	Grupo 2
Sim	15	16
Não	22	21

Também, consideramos o uso do índice de Rand ajustado (HUBERT; ARABIE, 1985) para comparar as partições. Sejam $U = \{u_1, \ldots, u_i, \ldots, u_R\}$ e $V = \{v_1, \ldots, v_j, \ldots, v_C\}$ duas partições de um mesmo conjunto de dados, possuindo R e C grupos, respectivamente. O índice ajustado de Rand é

$$CR = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}\Big[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\Big] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}, \tag{4.1}$$

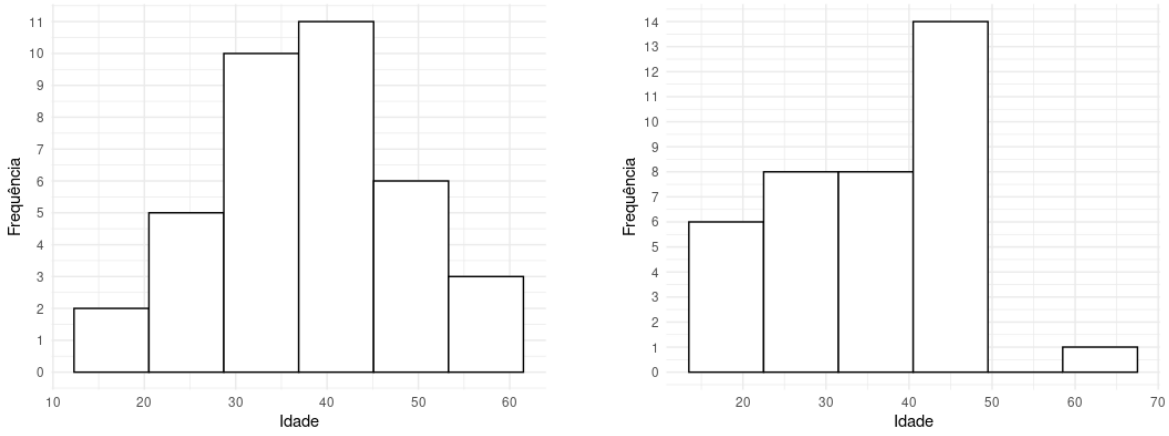
em que $\binom{n}{2} = \frac{n(n-1)}{2}$, n_{ij} representa o número de objetos que estão no grupo u_i e v_i, $n_{i.}$ e $n_{.j}$ indicam o número de objetos no grupo u_i e v_j, respectivamente. O total de objetos no conjunto de dados é expressado por n. O índice de Rand ajustado mede a semelhança entre uma partição *a priori* e a partição obtida através de algoritmos de agrupamento. O índice assume valores no intervalo $[-1, 1]$, em que 1 indica o ajuste perfeito entre as partições, enquanto que 0, ou valores negativos, correspondem a um ajuste encontrado por acaso. A Tabela 3 sumariza os valores para o índice de Rand ajustado, para a partição

conhecida acerca do diagnóstico, como, também, para as variáveis de destreza e sexo do indivíduo. Como pode-se observar, os valores do índice estão próximos de zero, não fornecendo indícios de semelhança entre a partição resultante do agrupamento e as outras já conhecidas. O agrupamento resultante também é analisado pela perspectiva da variável idade e não foram

Tabela 3 – Índice de Rand ajustado

Partição *a priori*	**Valor**
Diagnóstico	-0.0127
Destreza	0.0001
Sexo	-0.0055

encontradas evidências na diferença da idade média entre um grupo e outro, a partir da realização de um teste-t de Welch. Na Figura 7, temos os histogramas da idade para os indivíduos de cada grupo. Pode-se observar que a distribuição do grupo à esquerda é mais simétrico do que o outro, possuindo, até, mais pacientes com idade superior a 50 anos. As médias de idade dos grupos foram 37,5 e 35,4 anos.



(a) Grupo 1 (b) Grupo 2

Figura 7 – Histogramas da idade dos indivíduos por grupo

Finalmente, realizamos um teste de hipótese para examinar se a forma média é a mesma em cada grupo. O valor correspondente da estatística F com base no teste de Goodall de duas amostras, conforme descrito em (DRYDEN, 2016, página 199), é, aproximadamente, 12,8. Aplicando um teste de permutação, usando essa estatística, temos um p-valor igual a 0,01, aproximadamente. Consequentemente, há evidências de diferença na forma média entre os dois grupos ajustados pelo *K-means*.

5 CONCLUSÃO

5.1 VISÃO GERAL DOS RESULTADOS E DISCUSSÃO

Neste trabalho, estudamos o problema de agrupamento de formas tridimensionais utilizando o algoritmo *K-means*. Propondo uma adaptação a tal algoritmo com o objetivo de reduzir o custo computacional de sua execução. Dentro do cenário em que $m = 3$, a performance computacional do cálculo da forma média ϕ foi superior ao método de *Procrustes*. Caracterizando uma alternativa ao método de *Procrustes*, quando inserido em cenários em que o tamanho da amostra ou número de *landmarks* é grande.

Por conta de limitações computacionais, não foi possível a execução do algoritmo de agrupamento proposto utilizando a totalidade de *landmarks* no conjunto de dados. Desse modo, uma amostra sistemática foi considerada para o agrupamento. O procedimento de agrupamento para esse conjunto de dados não foi considerado satisfatório pela perspectiva de classificação do diagnóstico de esquizofrenia dos pacientes. Mesmo assim, o teste realizado sugere que a média dos dois grupos possuem formas diferentes.

O trabalho levou às seguintes contribuições:

- Análise de performance computacional para o cálculo da forma média, quando $m = 3$, comparando os métodos de *Procrutes* e a forma média ϕ;
- Proposição de uma versão otimizada do algoritmo *K-means* para o contexto de formas tridimensionais;
- Desenvolvimento do computacional do algoritmo *K-means* otimizado com aplicação ao conjunto de dados da superfície cerebral em 3 dimensões.

Na seção A, pode ser encontrada a implementação computacional do algoritmo de agrupamento considerado neste trabalho.

REFERÊNCIAS

A., S.; KLASSEN, E. P. *Functional and Shape Data Analysis*. [S.l.]: Springer, 2016.

ABBOTT, E. A. **Flatland: A Romance of Many Dimensions**. [S.l.]: Seeley Co., 1884.

ALEXANDRINO, M. **Introdução a Geometria Riemanniana**. [S.l.]: USP, 2018.

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**, Taylor Francis, v. 39, n. 5, p. 1016–1026, 2010.

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

BILDER, R. M.; WU, H.; BOGERTS, B.; DEGREEF, G.; ASHTARI, M.; ALVIR, J. M. J.; SNYDER, P. J.; LIEBERMAN, J. Absence of regional hemispheric volume asymmetries in first-episode schizophrenia. **American Journal of Psychiatry**, v. 151, 1994.

BOOKSTEIN, F. L. **Morphometric Tools for Landmark Data: Geometry and Biology**. [S.l.]: Cambridge University Press, 1992.

BRIGNELL, C. J. Shape analysis and statistical modelling in brain imaging. **PhD thesis**, University of Nottingham, United Kingdom, 2007.

BRIGNELL, C. J.; DRYDEN, I. L.; GATTONE, S. A.; PARK, B.; LEASK, S.; BROWNE, W. J.; FLYNN, S. Surface shape analysis with an application to brain surface asymmetry in schizophrenia. **Biostatistics**, v. 11, n. 4, p. 609–630, 03 2010.

COCHRAN, W. G. **Sampling Techniques, 3rd Edition.** [S.l.]: John Wiley, 1977.

COSTA, L. d. F. D.; CESAR, R. M. **Shape Analysis and Classification: Theory and Practice**. 1st. ed. USA: CRC Press, Inc., 2000.

DRYDEN, I. L. **shapes package**. Vienna, Austria, 2018. Contributed package, Version 1.2.4.

DRYDEN, I. L.; KUME, A.; LE, H.; WOOD, A. T. A. A multi-dimensional scaling approach to shape analysis. **Biometrika**, [Oxford University Press, Biometrika Trust], v. 95, n. 4, p. 779–798, 2008.

DRYDEN, K. V. M. I. L. **Statistical Shape Analysis**. [S.l.]: Wiley, 2016.

GEORGESCU, V. Clustering of fuzzy shapes by integrating procrustean metrics and full mean shape estimation into k-means algorithm. In: **IFSA/EUSFLAT Conf.** [S.l.: s.n.], 2009.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016.

GOSHTASBY, A. A. **Image Registration: Principles, Tools and Methods**. [S.l.]: Springer Publishing Company, Incorporated, 2012.

GUIDORIZZI, H. L. **Um Curso de Cálculo: Volume 3**. [S.l.]: LTC, 2002.

HARTIGAN, J. A.; WONG, M. A. A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 28, n. 1, p. 100–108, 1979.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. [S.l.]: Springer New York Inc., 2001.

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

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: With Applications in R**. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

KENDALL, D. G. The diffusion of shape. **Advances in Applied Probability**, Applied Probability Trust, v. 9, n. 3, p. 428–430, 1977.

KENDALL, D. G. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. **Bulletin of the London Mathematical Society**, v. 16, n. 2, p. 81–121, 03 1984.

LIMA, R. F. de. **Topologia e Análise no Espaço Rn**. [S.l.]: SBM, 2015.

LLOYD, S. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, Institute of Electrical and Electronics Engineers (IEEE), v. 28, n. 2, p. 129–137, mar 1982.

MACKAY, C. E.; BARRICK, T. R.; ROBERTS, N.; DELISI L. E. MAES, F.; VANDERMEULEN, D.; CROW, T. J. Application of new image analysis technique to study brain asymmetry in schizophrenia. **Neuroimaging**, v. 124, p. 25–35, 2003.

MARDIA J. T. KENT, J. M. B. K. V. **Multivariate Analysis**. [S.l.]: Academic Press, 1995.

Microsoft; WESTON, S. **foreach: Provides Foreach Looping Construct for R**. [S.l.], 2017. R package version 1.4.4.

PRESTON, S. P.; WOOD, A. T. A. Two-sample bootstrap hypothesis tests for three-dimensional labelled landmark data. **Scandinavian Journal of Statistics**, v. 37, n. 4, p. 568–587, 2010.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019.

R Core Team. **parallel package**. Vienna, Austria, 2020.

SMALL, C. G. **The Statistical Theory of Shape**. [S.l.]: New York, NY: Springer New York, 1996.

THOMPSON, D. W. **On Growth and Form**. [S.l.]: Cambridge University Press, 1917.

VINUé, G. Anthropometry: An R package for analysis of anthropometric data. **Journal of Statistical Software**, v. 77, n. 6, p. 1–39, 2017.

VINUé, G.; SIMO, A.; ALEMANY, S. The k-means algorithm for 3d shapes with an application to apparel design. **Advances in Data Analysis and Classification**, v. 10, p. 1–30, 10 2014.

A CÓDIGOS COMPUTACIONAIS

A seguinte função calcula o agrupamento *K-means* no contexto de formas tridimensionais, utilizando computação paralela.

```
################################################################################
#                                                                              #
# -------          Algoritmo de agrupamento K-means              ------- #
#                                                                              #
#                                                                              #
################################################################################

# ============================================================================ #
#                                                                              #
#                          DETALHES DA FUNÇÃO                                  #
#                                                                              #
#   A função recebe 6 objetos de entrada:                                      #
#                                                                              #
#   1 - Array com as matrizes de configuração                                  #
#   2 - Número de grupos/clusters a serem particionados                        #
#   3 - Número de inicializações aleatórias                                    #
#   4 - Número de iterações do algoritmo                                       #
#   5 - Valor para ser considerado como critério de parada das iterações#
#   6 - Número de processadores para a computação paralela                     #
#           ------- (NESSA ORDEM) --------                                     #
#                                                                              #
#   Obs.: Essa função escreve os resultados em um .txt no diretório            #
#         no qual foi executada                                                #
#                                                                              #
# ============================================================================ #


# Packages ----
library(shapes, warn.conflicts = F)
library(parallel, warn.conflicts = F)
library(foreach, warn.conflicts = F)
library(doParallel, warn.conflicts = F)
library(doSNOW, warn.conflicts = F)
```

```
ShapeKmeansParallel <- function(array3D,
                                numClust = 2,
                                niter = 100,
                                algSteps = 10,
                                stopCr = 0.0001,
                                cores = parallel::detectCores() - 1){

    # Creating parallel threads ----
    cl <- makeCluster(cores)
    registerDoSNOW(cl)

    # Aux. functions ----
    comb <- function(x, ...) {
      lapply(seq_along(x),
             function(i) c(x[[i]], lapply(list(...), function(y) y[[i]])))
    }

    pb <- txtProgressBar(max=niter, style=3)
    progress <- function(n) setTxtProgressBar(pb, n)
    opts <- list(progress=progress)

    # LloydShapesParallel algorithm ----
    set.seed(2020)
    time_iter <- list()
    comp_time <- c()
    list_asig_step <- list()
    list_asig <- list()
    vect_all_rate <- c()
    initials <- list()
    ll <- 1:numClust
    dist <- matrix(0, dim(array3D)[3], numClust)
    time_ini <- Sys.time()
    vopt <- 1e+08
    results <- foreach(iter=1:niter, .packages="shapes", .combine="comb", .multicombine=T,
    .options.snow=opts, .init=list(list(), list(), list(), list(), list())) %dopar% {
      obj <- list()
      meanshapes <- 0
      meanshapes_aux <- 0
      asig <- 0
```

```
mean_sh <- list()
n <- dim(array3D)[3]
initials[[iter]] <- sample(1:n, numClust, replace = FALSE)
meanshapes <- array3D[, , initials[[iter]]]
meanshapes_aux <- array3D[, , initials[[iter]]]
for (step in 1:algSteps) {
  for (h in 1:numClust) {
    dist[, h] = apply(array3D[, , 1:n], 3, riemdist,
                      y = meanshapes[, , h])
  }
  asig = max.col(-dist)
  for (h in 1:numClust) {
    if (table(asig == h)[2] == 1) {
      meanshapes[, , h] = array3D[, , asig == h]
      mean_sh[[step]] <- meanshapes
    }
    else {
      meanshapes[, , h] = procGPA(array3D[, , asig == h],
                                  distances = F,
                                  pcaoutput = F)$mshape
      mean_sh[[step]] <- meanshapes
    }
  }
  obj[[step]] <- c(0)
  for (l in 1:n) {
    obj[[step]] <- obj[[step]] + dist[l, asig[l]]^2
  }
  obj[[step]] <- obj[[step]]/n
  list_asig_step[[step]] <- asig
  if (step > 1) {
    aux <- obj[[step]]
    aux1 <- obj[[step - 1]]
    if (((aux1 - aux)/aux1) < stopCr) {
      break
    }
  }
}
optim_obj <- which.min(unlist(obj))
list(min(unlist(obj)),
```

```
            which.min(unlist(obj)),
            mean_sh[[optim_obj]],
            list_asig_step[[optim_obj]],
            initials[[iter]]
      )
    }
    close(pb)
    stopCluster(cl)
    optim_iter <- which.min(unlist(results[[1]]))
    capture.output(table(results[[4]][optim_iter]), file = "Clustering_results.txt")
    capture.output(results[[4]][optim_iter], file = "Clustering_results.txt", append = T)
    capture.output(results[[1]], file = "Clustering_results.txt", append = T)
    capture.output(results[[1]][optim_iter], file = "Clustering_results.txt", append = T)
    capture.output(Sys.time()-time_ini, file = "Clustering_results.txt", append = T)
    capture.output(results[[5]], file = "Clustering_results.txt", append = T)
    capture.output(results[[5]][optim_iter], file = "Clustering_results.txt", append = T)
}

##############################################################################
```