Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Programa de Pós-Graduação em Estatística

Algoritmos para Determinação do Número de Grupos em Estudos de Formas Planas

Rodrigo Alves de Oliveira

Dissertação de Mestrado

Recife-PE
2016

Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Rodrigo Alves de Oliveira

Algoritmos para Determinação do Número de Grupos em Estudos de Formas Planas

Orientador: Getúlio José Amorim do Amaral
Co-Orientadora: Renata Maria Cardoso Rodrigues de Souza

Área de Concentração: Estatística Aplicada

Dissertação apresentada ao programa de Pós-graduação em Estatística do Departamento de Estatística da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de **Mestre em Estatística**.

Recife-PE
2016

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

O48a Oliveira, Rodrigo Alves de
Algoritmos para determinação do número de grupos em estudos de formas planas / Rodrigo Alves de Oliveira. – 2016.
73 f.: il., fig., tab.

Orientador: Getúlio José Amorim do Amaral.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2016.
Inclui referências e apêndices.

1. Análise multivariada. 2. Estatística aplicada. 3. Análise de agrupamento I. Amaral, Getúlio José Amorim do (orientador). II. Título.

519.535 CDD (23. ed.) UFPE- MEI 2016-020

RODRIGO ALVES DE OLIVEIRA

ALGORITMOS PARA DETERMINAÇÃO DO NÚMERO DE GRUPOS EM ESTUDOS DE 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: 05 de fevereiro de 2016.

BANCA EXAMINADORA

Prof. PhD. Getúlio José Amorim do Amaral
UFPE

Prof. Dr. Abraão David Costa do Nascimento (Examinador Interno)
UFPE

Prof. Dr. Paulo José Duarte Neto (Examinador Externo)
UFRPE

Dedico esse trabalho à Deus.
Meus pais (Jávier e Fernanda).
Aos avós (Assis e Maria)

Agradecimentos

Agradeço a Deus por estar vivo, pelas bençãos e oportunidades que Ele me oferece todos os dias. Por sempre me conceder sabedoria nas escolhas dos melhores caminhos, coragem para acreditar, força para não desistir e proteção para me amparar.
Aos meus pais e avós, pelo apoio à educação desde criança, ao apoio emocional e financeiro durante todo esse período.
À namorada Suzan que entendeu o motivo da ausência e sempre me apoiou nas decisões que podem auxiliar nossa progressão social e cultural.
Aos amigos do colégio em Goiânia, em especial, ao Leonardo e Ramon que durante vários anos discutimos sobre o futuro e os projetos que em breve iremos dar continuidade. Aos amigos da UFG Alex, Felipe, Murilo e Isabel que sempre estavam do lado para enfrentar os momentos mais difíceis da graduação.
Aos familiares e primos.
À turma do Peteco que sempre está pronta pra conversar, divertir e tirar o momento de estresse com muita alegria e felicidade, "sempre".
Aos novos amigos da UFPE, em especial a turma de 2015, Pedro, Hildemar, Telma, Stênio, Jonas; além de Laura, Wanessa, Raquel, Jeny, pela ajuda nos momentos mais complicados (só quem passa sabe), paciência, boa vontade e contribuição pelo todo trabalho e estudos.
Agradeço também ao professor Getúlio Amaral de Amorim, pela orientação e paciência. A co-orientadora Renata Maria, pela excelente co-orientação.
Aos professores: Francisco Cribari, Gauss Cordeiro e Leandro Chaves Rêgo pela paciência, competência e ensinamentos.
À secretária Valéria e Elaine pela competência e dedicação perante aos professores e alunos de pós-graduação.
À CAPES pelo suporte Financeiro.

Se você não está disposto a arriscar, esteja disposto a uma vida comum.

Jim Rohn, empreendedor

Não tente ser uma pessoa de sucesso. Em vez disso, seja uma pessoa de valor.

Albert Einstein, físico

Há dois tipos de pessoa que vão te dizer que você não pode fazer a diferença neste mundo: as que têm medo de tentar e as que têm medo de que você se dê bem.

Ray Goforth, executivo

Resumo

Análise de formas planas é uma área de conhecimento bastante útil e sólida para lidar com estudos de estruturas de objetos e informação geométrica. A fim de descrever objetos bidimensionais é necessário especificar um sistema de coordenadas a qual deve ser invariante sob locação, escala e rotação da configuração tal como as coordenadas de Kendall. E uma versão linearizada do espaço de formas são as coordenadas tangentes, esta pertence ao espaço Euclidiano, portanto, toda literatura de análise multivariada pode ser utilizada. Em diversas ocasiões é necessário agrupar conjuntos de dados de tal maneira que se tenha grupos com características mais homogêneos entre si. Para tanto Amaral et al. (2010a) desenvolveu o algoritmo K-médias para lidar com análise de formas. Devido as desvantagens deste algoritmo, Jayasumana et al. (2013) propôs o algoritmo *Kernel K-médias*. Estes dois algoritmos dependem da escolha do número de grupos, K. E para o segundo, deve-se estimar o parâmetro de largura de banda. Em situações em que não se conhecem os rótulos dos grupos, a escolha de um valor apropriado para K é difícil. Para resolver esse desafio, medidas de validade tentam determinar como precisamente se retratam os grupos dos dados. No entanto, diversas medidas de validade surgem, e diferentes medidas geralmente produzem resultados discrepantes. Esta dissertação introduz métodos para computar o número de grupos em um determinado conjunto de dados que lidam com a natureza das estruturas planas. Os métodos propostos são baseados nas medidas de validade Silhoueta, Davies-Bouldin e os Resíduos Procrustes. Gerou-se amostras de duas populações da distribuição Bingham complexa a qual possui suporte na esfera unitária; e também amostras de duas populações com espaço nos marcos. Considera-se vários cenários com alta e baixa concentração dos dados. Percebe-se que os índices para coordenadas tangentes encontram corretamente o número de grupos para dados de alta concentração assim como os índices modificados para coordenadas de Kendall. Já em situações com baixa concentração os índices para coordenadas tangentes não funcionam bem, portanto, não identificam o número correto de grupos, ao contrário, os índices com natureza própria de formas planas conseguem estimar o verdadeiro número de grupos para os dados simulados. Os índices mais apropriados são o Procruste Residual e o Davies-Bouldin ajustado pela segunda vez. Análise de dados reais mostra

que os índices existentes para coordenadas tangentes e os índices modificados para coordenadas de Kendall estimam o número correto de grupos.

Palavras-chave: análise de formas. análise de agrupamentos. validação interna. K-means. Kernel K-means

Abstract

Statistical Shape Analysis is a useful and solid area of knowledge for deal objects structures study and geometrical information. In order to describe two-dimensional objects you must specify a coordinate system which must be filter out translation, rotation and scale information of the setting as the Kendall coordinates. One linearized version of the shape space in the vicinity of a particular point of shape space is the tangent coordinates, that belongs to the Euclidian space, so all multivariate analysis may be used. On several occasions it is necessary to group data sets in such a way that it has groups with more homogeneous characteristics together. Therefore, Amaral et al. (2010a) developed the K-means algorithm to deal with shape analysis. Because of the disadvantages of this algorithm, Jayasumana et al. (2013) proposed Kernel K-means algoritm. These two algorithms depends on the choice of the number of groups, K. And for second, to estimate the bandwidth parameter. In situations in which there is no known labels groups, the choice of an appropriate value for K is difficult. To overcome this challenge, validity measures attempt to determine how accurately the clusters reflect the data. However, numerous validity measures proliferate, and different measures often produce disparate results. This paper introduces methods to compute the number of groups in a given data set that deal with the nature of the planar shapes. The proposed methods are based on the validity of measures Silhoutte, Davies-Bouldin and Procrustes Residuals. Samples were generated from two populations of complex Bingham distribution which is supported on the unit sphere; and also samples of two populatoin with space in the landmarks. Considered some scenarios with high and low concentration of data. It is noticed that the contents are properly coordinated tangent to the number of groups for high-concentration data, as well as modified indices for Kendall coordinates. Already in situations with low concentration indexes to coordinate tangents do not work well, so do not identify the correct number of groups, by contrast, the indexes with the nature of planar shapes can estimate the true number of groups for the simulated data. The most suitable index are Procrustes Residuals and Davies-Bouldin adapted the second time. Real data analysis shows that the existing index for tangent coordinates and indexes modified to Kendall coordinates estimate the correct number of groups.

Keywords: clustering. planar shapes. K-means. Kernel K-means. internal validation

Lista de Figuras

Lista de Tabelas

Sumário

CAPÍTULO 1

Introdução

1.1 Motivação e Revisão da Literatura

Atualmente, análises de formas (*shape analysis*) é uma área bastante útil e sólida para lidar com estudos de formas de objetos e informação geométrica. Seus métodos têm se mostrado relevantes em vários campos e aplicações tais como biologia, medicina, análise de imagens, arqueologia, geografia, geologia, agricultura, genética, imagens biomédicas, reconhecimento de padrões militares, para mais detalhes veja (DRYDEN e MARDIA, 1998).

O termo "forma" no inglês pode ser "form" ou "shape". A diferença é basicamente, "shape" é a descrição de um objeto em virtude de quantos lados existem e por algum grau em relações aos ângulos. Há uma fronteira clara e uma borda bem definida, como por exemplo, círculos e quadrados. Por outro lado, "form" detalha ainda mais a área delimitada pelas linhas criadas do objeto, por exemplo, cubo, cilindros, esferas.

Avanços na tecnologia tem levado à coletas de rotinas de informações geométricas e o estudo de formas de objeto é cada vez mais importante. Em particular, a localização de pontos em objetos é muitas vezes simples, e a análise de formas lida com esses conjunto de pontos de dados. É necessário uma notação matemática de formas a qual é uma classe equivalente sob certos tipos de transformações de grupos. As transformações mais comuns 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 ocupa em estudar medidas de resumo e comparações de formas de objetos segundo os autores Dryden e Mardia (1998). O primeiro trabalho nesta área foi feito por Kendall (1977). No entanto, foi no trabalho de Kendall (1984) que a análise de formas foi formalizada e foram definidos os conceitos básicos da área. Um dos pontos mais importantes do trabalho de Kendall foi a proposta dos sistemas de coordenadas que tinham a finalidade de obter a forma de um objeto através de pontos dispostos no objeto chamados de marcos anatômicos. Essas coordenadas propostas estão dispostas em um espaço não Euclidiano e, portanto,

é necessário ferramentas próprias para lidar com as análises.

Em diversas ocasiões em análise de formas, é necessário agrupar um conjunto de dados em grupos de tal maneira que se tenha grupos com características mais homogêneas. Por exemplo, quando é necessário detectar o número de diferentes espécies em um conjunto, análise de agrupamento pode ser utilizado para detectar os grupos de cada espécie, (AMARAL et al., 2010a). Dessa forma, muitos pesquisadores têm conduzido estudos e têm estudado a *performance* desses métodos.

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.

Análise de agrupamento é uma área importante da análise multivariada em que não se conhece os rótulos das categorias que denotam a partição a priori dos objetos em uso. A ideia principal é dividir um conjunto de objetos em grupos de tal forma que os objetos dentro de um mesmo grupo possuem características similares, enquanto que os objetos de diferentes grupos possuem características mais distintas. Análise de agrupamento é amplamente utilizado em diversos campos do estudo como pode ser estudado pelos autores Jain e Dubes (1988), Jain; Murty e Flynn (1999), Kaufman e Rousseeuw (2009).

Uma das técnicas mais utilizadas para análise de agrupamento é o algoritmo K-médias, (JAIN, 2010). O algoritmo K-médias é eficiente e efetivo, mas possui algumas deficiências. Dentre elas, o algoritmo gera resultados pobres quando um conjunto de dados é composto por grupos não linearmente separáveis. Isso motivou a evolução de outros métodos de agrupamento que apresentam maior eficiência na separação destes tipos de grupos. Girolami (2002) desenvolveu um algoritmo capaz de produzir separações não lineares entre grupos, transformando o espaço de entradas em um espaço de alta dimensão e então executa o agrupamento neste novo espaço. Este mapeamento para o novo espaço de alta dimensão é realizado através de funções de *Kernel*, (SHAWE-TAYLOR e CRISTIANINI, 2004).

Um problema persistente a esses algoritmos é que os resultados possuem certa dependência de alguns parâmetros, mais significante para a proposta dessa dissertação, o número de grupos "K", e a largura de banda para o *Kernel* K-médias.

O primeiro passo nesses algoritmos é escolher o número de grupos no qual os dados serão divididos. A escolha inicial de K é, em grande parte, uma decisão interpretativa do pesquisador. Execuções sucessivas dos algoritmos pode otimizar a divisão dos dados para algum número de grupos, mas a escolha de um valor inicial de K pressupõe o conhecimento prévio da estrutura de dados. Sem esse prévio conhecimento é difícil saber o número correto de grupos. Em alguns casos simples (dados bidimensionais ou tridimensionais), é fácil notar que tal conjunto de dados possuem grupos bem definidos, mas nem sempre isso acontece. Sendo assim, é necessário estratégias mais eficientes para determinar o número de grupos.

Com esses problemas relacionados a análise de agrupamento não supervisionado, nas últimas décadas estudos foram realizados com o intuito de validar os grupos obtidos pelos algoritmos. Validação externa e validação interna são as duas principais categorias de validação de agrupamentos, (JAIN e DUBES, 1988). A principal diferença é se informação externa, conhecimento a priori dos objetos, é usada ou não para validar os agrupamentos. Ao contrário das

medidas de validação externa (as quais usam informações externas não presente nos dados) as medidas de validação interna fazem uso somente das informações dos dados.

Desde que a validação externa conheça a verdadeira partição e o número de grupos com antecedência, eles são usados principalmente para a escolha de um algoritmo de agrupamento ideal em um conjunto de dados específico. Por outro lado, validação interna pode ser usado para escolher o melhor algoritmo, mas também, determinar o número ótimo de grupos sem alguma informação adicional. Além disso, pode-se estimar por validação cruzada, o valor ideal para a largura de banda do truque do *Kernel* no algoritmo *Kernel* K-médias. Alguns exemplos de validação externa são desenvolvidas por Jain e Dubes (1988), tais como índice de Rand, índice de Rand ajustado e *F-measure*; e exemplos de índice interno são Silhoueta e Davies-Bouldin.

Como já apresentado anteriormente, as técnicas de análises de formas não estão no espaço Euclidiano. Dessa forma, os índices de validação existentes na literatura não podem ser utilizadas. Para contornar tal problema, essa dissertação está disposta a propor índices internos de validação para as coordenadas de Kendall. Portanto, propôs-se um novo índice que considera os conceitos básicos de análise de formas. E quatro índices modificados que são versões de índices existentes na literatura para tratar dados de formas. Essas versões são baseadas nos seguintes índices da literatura de agrupamento: Silhoueta (ROUSSEEUW, 1987) e Davies-Bouldin (DAVIES e BOULDIN, 1979); bem como suas versões kernelizadas (CAMPS-VALLS, 2006) para índice Silhoueta e (NASSER; HÉBERT e HAMAD, 2007) para índice Davies-Bouldin.

Valida-se o índice interno proposto e os modificados através de estudos de simulação no espaço dos marcos anatômicos e no espaço das pré-formas (distribuição Bingham), e de dados reais para análise de formas.

As plataformas computacionais utilizadas no desenvolvimento deste projeto foram o LaTeX para editar o projeto Lamport (1994), Greenberg (2000); e software `R`, R Development Core Team (2009), para gerar os resultados das simulações e dos dados reais.

Por fim, registra-se que foi utilizado um notebook DELL (Intel(R) Core(TM) i7-2640M CPU @ 2.80GHz 8GB Memória(RAM), sistema 64-bit, 452GB e plataforma Windows 7 Home Premium) para realização do projeto. Para rodar as simulações, utilizou-se o servidor do Departamento de Estatística da Universidade Federal de Pernambuco, com a seguinte configuração processador i7 com 8 núcleos, 1TB de disco rígido e 8GB de Memória (RAM) e sistema operacional Linux asimov 3.13.0-37-generic.

1.2 Organização da dissertação

Além do capítulo da introdução, esta dissertação é composta por mais quatro capítulos. No Capítulo 2 apresentamos uma revisão geral sobre *shape analysis* em estruturas planas planas fazendo uso das coordenas de Kendall e análise de agrupamento não supervisionado e os algoritmos K-médias e *Kernel* K-médias para formas planas. No Capítulo 3 apresentamos índices internos de validação existentes para análise de agrupamento em um contexto geral, e em consequência propomos um novo índice próprio para lidar com formas planas. Além disso, propomos índices internos modificados para validar e encontrar o número ideal de grupos a partir

dos resultados dos algoritmos particionais no espaço de entrada e no espaço característico para um determinado conjunto de dados em formas planas no espaço não Euclidiano. Vale notar que esses índices podem ser usados para encontrar a largura de banda para o algoritmo *Kernel* K-médias, através da validação cruzada . No Capítulo 4 é avaliado o desempenho desses índices internos propostos para dados simulados e aplicações em conjunto de dados reais. Finalmente, no Capítulo 5 são apresentadas as conclusões obtidas na dissertação e algumas sugestões para trabalhos futuros.

CAPÍTULO 2

Fundamentação Teórica

2.1 Análises de Formas - *Shape Analysis*

Análise de formas é uma área bastante útil para lidar com formas de objetos e informações geométricas. É um tópico relevante para vários campos de pesquisa e aplicações tais como biologia, medicina, análises de imagem, arqueologia, geografia, geologia, agricultura, genética, análise de imagem biológicas, reconhecimento de alvos militares, para mais detalhes veja os autores Bookstein (1997),Small (2012), Dryden e Mardia (1998), Kendall et al. (2009). Formas são definidas como "toda informação geométrica que permanece quando locação, escala e efeitos de rotação são filtrados de um objeto" Kendall (1977). Não há um modo trivial de representar as formas por análises matemáticas e estatísticas e somente ao longo de quatro décadas que as definições veem sendo estudados e desenvolvidos.

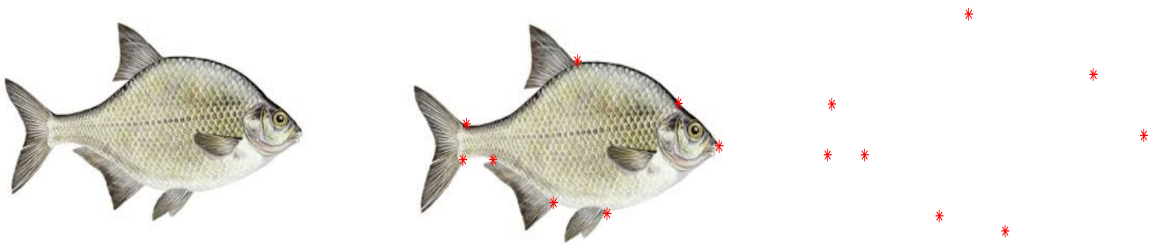
Nas próximas subseções serão especificadas as abordagens feitas para representarem a forma, a configuração matemática, análises e distâncias para lidar com a forma.

2.1.1 Marcos Anatômicos - *Landmarks*

Uma maneira de descrever uma forma é indicar um subconjunto finito de pontos no contorno do objeto. O número de pontos não seguem uma restrição teórica segundo Kendall et al. (2009). Esse número finito de pontos de cada objeto é conhecido como marcos. O marco é a principal fonte dos dados para a descrição dos formas. A posição dos pontos estão associadas à coordenadas cartesianas.

A definição de marcos, segundo Dryden e Mardia (1998), é "um ponto correspondente em cada objeto que corresponde entre e dentro da população", ou seja, todos os indivíduos possuem os mesmos marcos e quantidades (diz-se homólogos). Também é um ponto no espaço bi ou tridimensional que corresponde a posição de um característica particular de um objeto de interesse.

Figura 2.1: Representação da forma e marcos.



(a) Contorno e forma (b) Marcos anatômicos (c) 8 marcos da configuração da forma

Fonte: Elaborada pelo autor.

Em Dryden e Mardia (1998), vê-se que os marcos são divididos em três sub-grupos:

1. Marcos anatômicos: são pontos atribuídos por um perito que correspondem a alguma característica biologicamente significativa.

2. Marcos matemáticos: são pontos alocados em um objeto de acordo com algumas propriedades matemáticas ou geométricas da imagem, por exemplo, pontos de alta curvatura ou ponto extremo, inflexão, máximos, mínimos.

3. Pseudo-Marcos: são pontos construídos em um objeto, localizados no contorno ou entre os marcos anatômicos e/ou matemáticos. Sua utilização é mais para conceituar a forma do objeto.

Na Figura 2.1 pode-se notar a representação do contorno de um tubarão, os marcos anatômicos e a configuração da forma. Quando digitalizado os marcos, os objetos geralmente possuem tamanhos, posições e rotações diferentes, dentro do equipamento de medição. Por isso, é necessário retirar os efeitos de locação, escala e rotação do conjunto de dados em ordem para trazê-los para um tamanho, orientação e posição padronizados antes de uma análise mais aprofundada. Dessa forma, na próxima subseção, mostra-se o passo a passo para filtrar esses efeitos.

2.1.2 Configuração Matemática

A fim de descrever a forma de um objeto é necessário especificar um sistema de coordenadas. Existem vários sistemas de coordenadas, coordenadas de Bookstein, propostas por Bookstein (1984) e Bookstein (1986); coordenadas polares de Kent proposta por Kent (1994), coordenadas de forma Goodall-Mardia QR desenvolvida por Goodall e Mardia (1992) e Goodall e Mardia (1993); e entre outras, as coordenadas de Kendall podem ser vistas em (DRYDEN e MARDIA, 1998). Uma escolha apropriada do sistema de coordenadas para formas deve ser invariante sob locação, escala e rotação da configuração.

Uma configuração é um conjunto de marcos em um objeto particular. Uma configuração matemática X é representada por uma matriz $k \times m$ de coordenadas cartesianas de k marcos

em m dimensões. O espaço da configuração é o espaço de todas coordenadas possíveis dos marcos.

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x_{1,m} \\ x_{2,1} & \dots & x_{2,m} \\ \vdots & \ddots & \vdots \\ x_{k,1} & \dots & x_{k,m} \end{bmatrix} \tag{2.1}$$

Nesta dissertação serão considerados os casos onde $k \geq 3$ e $m = 2$, o que corresponde as formas planas. Assim, a matriz de configuração X da expressão (2.1) resume-se a

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2]^\top = \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{k,1} & x_{k,2} \end{bmatrix} \tag{2.2}$$

Algumas transformações devem ser feitas na matriz X para remover os efeitos de locação, escala e rotação. Para $m = 2$, a configuração matemática deve ser reescrita como um vetor complexo. Defina um vetor complexo $(k \times 1)$ tal que

$$\mathbf{z}^0 = (\mathbf{z}_1^0, \dots, \mathbf{z}_k^0)^\top = (x_{1,1} + ix_{1,2}, \dots, x_{k,1} + ix_{k,2})^\top \tag{2.3}$$

o qual corresponde as coordenadas complexas para os marcos.

Nesta dissertação, será trabalhada as coordenadas de Kendall, o primeiro trabalho feito nessa área foi Kendall (1977), mas em Kendall (1984) que realmente formalizaram e definiram os conceitos básicos. Um dos pontos mais importantes do trabalho de Kendall foi a proposta dos sistemas de coordenadas. Estes sistemas possuem finalidade de obter a forma de um objeto por meio dos pontos dispostos através dos marcos. Para um primeiro passo, deve-se retirar os efeitos da locação. Dessa forma, para remover a locação da forma devemos definir a sub-matriz de Helmert (**H**). A matriz de Helmert ($\mathbf{H}^F$) é uma matriz quadrática ortogonal $k \times k$ com a primeira linha de elementos igual a $1/\sqrt{k}$, então a j-ésima linha possui $(j-1)$ elementos iguais a $-1/\sqrt{j(j-1)}$ seguido por um elemento igual a $(j-1) \times 1/\sqrt{j(j-1)}$ e $(k-j)$ zeros. A sub-matriz de Helmert é a matrix de Helmert sem a primeira linha.

Por exemplo, para $k = 4$ a matriz de Helmert é

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

e a sub-matriz de Helmert será

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

Para remover a locação do vetor complexo $\mathbf{z}^0$, basta multiplicar o vetor pela sub-matriz de Helmert ($\mathbf{H}$) de dimensão $(k-1) \times k$. A configuração Hermertizada é dada por

$$\mathbf{w}_{(k-1\times 1)} = \mathbf{H}_{(k-1\times k)}\mathbf{z}^0_{(k\times 1)} \tag{2.4}$$

onde $\mathbf{w}$ representa a configuração $\mathbf{z}^0$ sem o efeito de locação.

Pode-se reverter de volta aos marcos centrados de um marco Helmertizado por pré-multiplicar por $\mathbf{H}^\top$, como

$$\mathbf{H}^\top\mathbf{H} = \left(I_k - \frac{1}{k}1_k 1_k^\top\right)$$

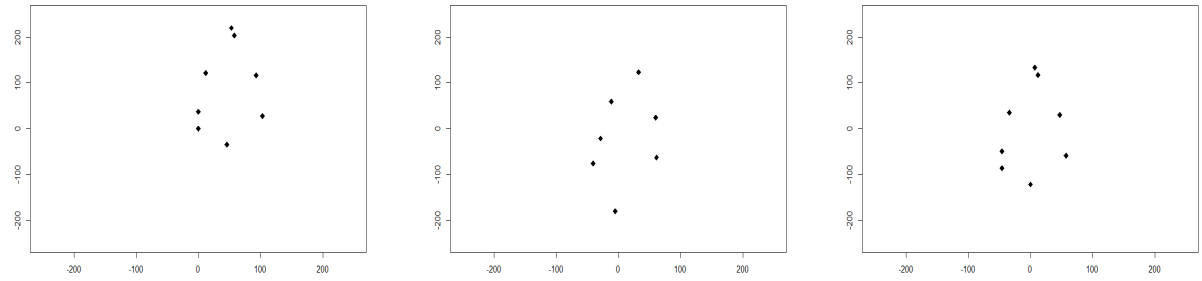
.

Dessa forma, pré-multiplicando o vetor $\mathbf{w}$ por $\mathbf{H}^\top$ obtém-se a configuração centrada

$$\mathbf{H}^\top\mathbf{w} = \mathbf{H}^\top\mathbf{H}\mathbf{z}^0 = \left(I_k - \frac{1}{k}1_k 1_k^\top\right)\mathbf{z}^0 = \mathbf{z}^0 - \frac{1}{k}\sum_{j=1}^{k}\mathbf{z}^0_{(j)}1_k \tag{2.5}$$

Para exemplificar, tem-se a configuração matemática de um indivíduo obtido dos dados de gorilas machos, (DRYDEN e MARDIA, 1998), em que $\mathbf{z}^0 = (53+220i, 46-35i, 0+0i, 0+37i, 12+122i, 58+204i, 93+117i, 103+28i)^\top$. A Figura 2.2 representa as configurações original, Helmertizada e centraliza para marcos de um gorila macho. Vale notar que a configuração Helmertizada, Figura 2.3(b), perde a dimensão original dos dados. Este problema é corrigido com a multiplicação por $\mathbf{H}^\top$.

Figura 2.2: Configurações de um gorila macho.



(a) Configuração original. (b) Configuração Helmertizada. (c) Configuração Centrada.

Fonte: Elaborada pelo autor.

Para remover o efeito de escala deve-se dividir a configuração Helmertizada, obtida na ex-

pressão (2.4), pela sua norma sendo assim,

$$\mathbf{z}_{(k-1\times 1)} = \frac{\mathbf{w}}{|\mathbf{w}|} = \frac{\mathbf{w}}{\sqrt{\mathbf{w}^*\mathbf{w}}} = \frac{\mathbf{H}_{(k-1\times k)}\mathbf{z}^0_{(k\times 1)}}{\sqrt{(\mathbf{H}_{(k-1\times k)}\mathbf{z}^0_{(k\times 1)})^*\mathbf{H}_{(k-1\times k)}\mathbf{z}^0_{(k\times 1)}}} \tag{2.6}$$

onde $\mathbf{w}^*$ é o transposto do conjugado de $\mathbf{w}$ e $|\cdot|$ denota a norma complexa de $\mathbf{w}$. O vetor $\mathbf{z}$, de acordo com Kendall (1984), é chamado de pré-forma da configuração complexa $\mathbf{z}^0$. É importante notar que a pré-forma é uma forma com o efeito de rotação retido.

Devido a importância da pré-forma no estudo das coordenadas de Kendall, alguns conceitos importantes devem ser considerados.

Definição 1 (Pré-forma)**.** *As pré-formas de uma matriz de configuração* $\boldsymbol{X}$*, da Equação* (2.2)*, é dado por*

$$\boldsymbol{z}_{(k-1\times m)} = \frac{\boldsymbol{H}_{(k-1\times k)}\boldsymbol{X}_{(k\times m)}}{|\boldsymbol{HX}|} \tag{2.7}$$

o qual é invariante sob locação e escala da configuração original.

A partir da Equação (2.7) pode-se obter as **pré-formas centralizadas** de forma que

$$\mathbf{z}_{\mathbf{C}_e(k\times m)} = \mathbf{C}_{e\cdot(k\times k)}\mathbf{X}_{(k\times m)}/|\mathbf{C}_e\mathbf{X}|$$

desde que $\mathbf{C}_e = \mathbf{H}^\top\mathbf{H}$. Note que $\mathbf{z}$ é uma matriz $(k-1)\times m$ enquanto que $\mathbf{z}_{\mathbf{C}_e}$ é uma matriz $k\times m$. A vantagem em usar $\mathbf{z}$ é por ser de posto completo e sua dimensão é menor do que de $\mathbf{z}_{\mathbf{C}_e}$. Por outro lado, a vantagem de trabalhar com a pré-forma centralizada $\mathbf{z}_{\mathbf{C}_e}$ é que a representação das coordenadas Cartesianas é coerente com a configuração original, (DRYDEN e MARDIA, 1998).

O espaço das pré-formas é o espaço de todas possíveis pré-formas $\mathbf{z}$. Ou seja, o espaço de todos os possíveis vetores de dimensão $(k-1)$ que não possuem a informação da locação e escala. Para pré-formas planas, este espaço é uma hiperesfera complexa de dimensão $(k-1)$, isto é

$$\mathbb{C}S^{k-2} = \{\mathbf{z} : \mathbf{z}^*\mathbf{z} = 1, \mathbf{z} \in \mathbb{C}^{k-1}\} \tag{2.8}$$

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

Definição 2 (Forma)**.** *A forma de uma matriz de configuração* $\boldsymbol{X}$ *é toda a informação geométrica sobre* $\boldsymbol{X}$ *que é invariante sobre locação, rotação e escala. A forma pode ser representada como*

$$[\boldsymbol{z}] = \{e^{i\theta}\boldsymbol{z} : \theta \in [0, 2\pi)\} \tag{2.9}$$

em que θ *é o grupo especial ortogonal de rotações e* $\boldsymbol{z}$ *é a pré-forma de* $\boldsymbol{X}$*.*

Para $m = 2$ o espaço da forma é o espaço projetivo complexo $\mathbb{C}P^{k-2}$, o espaço de linhas complexas que passam pela origem.

Definição 3 (Ícone)**.** *Um ícone é um membro particular do conjunto de formas* $[\mathbf{z}]$ *o qual é tomado como sendo a representatividade da forma.*

A palavra ícone indica “imagem ou semelhança” e é apropriado como uso para retratar uma imagem representativa de uma classe equivalente da forma o qual possui ‘semelhança’ para outros membros, isto é, os objetos da classe são todos similares. O termo foi usado primeiramente por Goodall (1991). A pré-forma centrada $\mathbf{z}_{\mathbf{C}_e}$ é uma escolha apropriada de ícone. Dessa forma, iremos usar a pré-forma centrada para ter uma representação da configuração original.

2.1.3 Distâncias Procrustes para caso planar

Nesta subseção, mostra-se alguns conceitos básicos para amostras aleatórias de objetos. Alguns desses aspectos de análise de forma são: obter a estimativa da forma média de uma amostra aleatória, o cálculo de distâncias entres formas, os resíduos de cada objeto em relação a um grupo.

Um importante conceito de análise de forma é estimar a forma média de uma amostra aleatória de configurações. Considere $\mathbf{z}_1^0, \ldots, \mathbf{z}_n^0$ como uma amostra aleatória de configurações complexas de uma população de n objetos ou indivíduos o qual $\mathbf{z}_i^0$ foi definido pela Equação (2.3).

De acordo com Kent (1994) obtém-se o seguinte resultado para a estimação da **forma média Procrustes completa** $[\hat{\boldsymbol{\mu}}]$ para formas planas

Resultado 1. *A **forma média Procrustes completa** $[\hat{\boldsymbol{\mu}}]$ pode ser encontrada como o autovetor correspondente ao maior autovalor da **soma quadrática complexa** e **matriz produto***

$$S = \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}_i^* / (\boldsymbol{w}_i^* \boldsymbol{w}_i) = \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^*, \tag{2.10}$$

onde $\boldsymbol{z}_i = \boldsymbol{w}_i / ||\boldsymbol{w}_i||$, $i = 1, \ldots, n$, são as pré-formas.

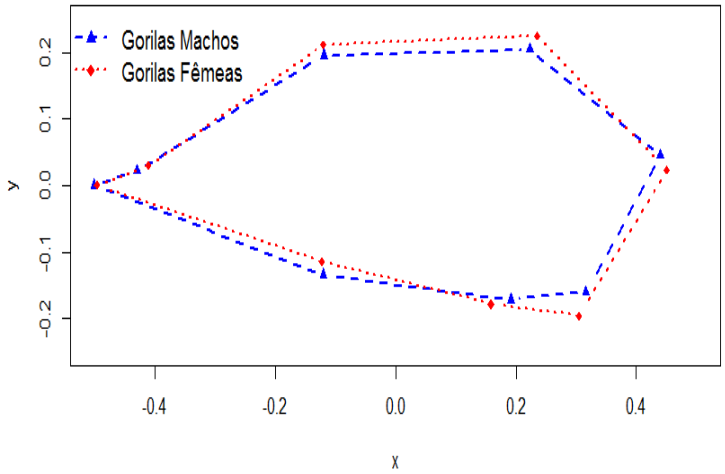
Assim, $\hat{\boldsymbol{\mu}}$ é dado pelo autovetor complexo correspondente ao maior autovalor, ou autovetor dominante de S. O autovetor é único (até uma rotação - todas rotações de $\hat{\boldsymbol{\mu}}$ são também soluções, mas todos estes correspondem a mesma forma), desde que exista um único autovalor maior de S. A forma média de dados dos gorilas 29 machos e 30 fêmeas, (DRYDEN e MARDIA, 1998), são apresentados na Figura 2.3.

A configuração possui uma rotação arbitrária, (DRYDEN e MARDIA, 1998). Assim, é necessário rotacionar todas configurações de tal forma que eles estarão tão próximos quanto possível da forma média amostral. Dessa forma define-se o **ajuste Procrustes completo** ou **coordenadas Procrustes completa** de $w_1, \ldots, w_n$ são

$$\mathbf{w}_i^P = \frac{\mathbf{w}_i^* \hat{\boldsymbol{\mu}} \mathbf{w}_i}{\mathbf{w}_i^* \mathbf{w}_i} = \mathbf{z}_i^* \hat{\boldsymbol{\mu}} \mathbf{z}_i \ , \ i = 1, \ldots, n$$

onde cada $\mathbf{w}_i^P$ é o ajuste Procrustes completo de $\mathbf{w}_i$ em $\hat{\boldsymbol{\mu}}$. A forma média Procrustes completa pode ser obtida por tomar a média aritmética das coordenas Procrustes completa, ou seja, $\frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_i^P$ tem a mesma forma como a forma média Procrustes $[\hat{\boldsymbol{\mu}}]$.

Figura 2.3: A forma média Procrustes completa de gorilas machos e fêmeas.



Fonte: Elaborada pelo autor.

Os resíduos Procrustes são calculados como

$$\mathbf{r}_i = \mathbf{w}_i^P - \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i^P \right) \ , \ i = 1, \ldots, n \tag{2.11}$$

e os resíduos Procrustes são usados para investigar a variabilidade da forma.

Um conceito de distância entre duas formas é necessário para definir completamente o espaço métrico de forma não Euclidiana.

Considere duas matrizes de configuração de k pontos e dimensão $m = 2$, $\mathbf{X}$ e $\mathbf{Y}$, e suas configurações centradas e de tamanho unitário (pré-forma centrada) $\mathbf{z}_x = (\mathbf{z}_{x1}, \ldots, \mathbf{z}_{xk})^\top$ e $\mathbf{z}_y = (\mathbf{z}_{y1}, \ldots, \mathbf{z}_{yk})^\top$, de duas configurações $\mathbf{X}$ e $\mathbf{Y}$ onde $||\mathbf{z}_x|| = 1 = ||\mathbf{z}_y||$ e $\mathbf{z}_x^* \mathbf{1}_k = 0 = \mathbf{z}_y^* \mathbf{1}_k$. Dessa forma, a **distância Procrustes completa** entre duas formas $\mathbf{z}_x$ e $\mathbf{z}_y$

$$d_F^2 = 1 - |\mathbf{z}_x^* \mathbf{z}_y|^2 \tag{2.12}$$

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

Para dados no plano o espaço pré-forma é uma esfera complexa $\mathbb{C}S^{k-2}$ de raio unitário em dimensão complexa $k - 1$ definido na Equação (2.8). O ângulo entre as pré-formas complexas $\mathbf{z}_x$ e $\mathbf{z}_y$ é

$$\rho = \arccos(|\mathbf{z}_x^* \mathbf{z}_y|) \tag{2.13}$$

Essa quantidade também denominada como geodésica, é definida como o caminho mais curto entre $\mathbf{z}_x$ e $\mathbf{z}_y$ na hiperesfera da pré-forma e não é afetada pela rotação, (ZELDITCH; SWIDERSKI e SHEETS, 2012). Consequentemente, pode-se ver explicitamente que a **distância**

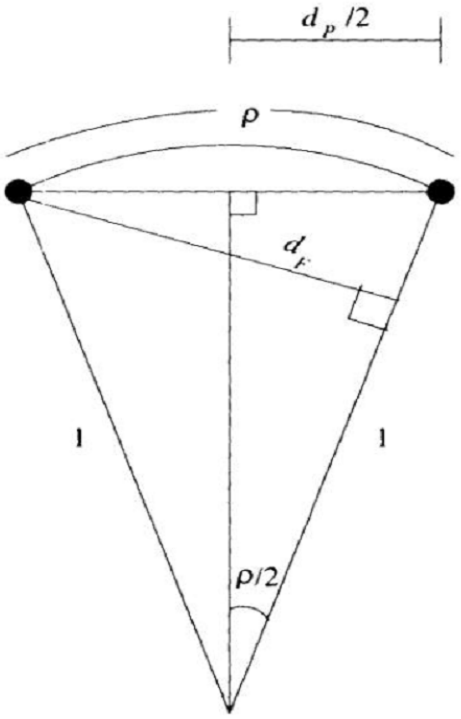
Procrustes ρ é o ângulo entre as pré-formas $\mathbf{z}_x$ e $\mathbf{z}_y$. Também, desde que, o raio da esfera da pré-forma é 1 pode-se considerar ρ ser a distância ótima no círculo na esfera da pré-forma.

A **distância Procrustes parcial** também é invariante quanto a rotação entre $\mathbf{z}_x$ e $\mathbf{z}_y$ e é dada por

$$d_P^2 = 2(1 - |\mathbf{z}_x^* \mathbf{z}_y|) = 2(1 - \cos\rho) \tag{2.14}$$

A Figura 2.4 representa as distâncias (2.12), (2.13) e (2.14) na hiperesfera da pré-forma.

Figura 2.4: Ilustração da relação entre as distâncias d_F, ρ e d_P na esfera da pré-forma.



Fonte: (DRYDEN e MARDIA, 1998).

2.1.4 Coordenadas no Espaço Tangente

O espaço tangente é a versão linearizada do espaço de formas na proximidade de um ponto particular do espaço de forma. Uma das vantagens do espaço tangente é que pode ser usado nas técnicas padrão de multivariada diretamente. Existem vários tipos diferentes de coordenadas no espaço tangente. Aqui nós usaremos as coordenadas tangente Procrustes parcial.

Considere $\mathbf{x}_1, \ldots, \mathbf{x}_n$ uma amostra de configurações, dessa forma as coordenadas tangentes será

$$\mathbf{t}_i = e^{i\hat{\theta}}[I_{k-1} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^*]\mathbf{z}_i, \quad i = 1, \ldots, n$$

onde $\mathbf{z}_i$ é a pré-forma correspondente a configuração $\mathbf{x}_i$ definida em (2.7), $\hat{\theta}$ minimiza $||\hat{\boldsymbol{\mu}} - \mathbf{z}e^{i\hat{\theta}}||^2$ e $||\mathbf{z}|| = \sqrt{\mathbf{z}^*\mathbf{z}}$.

Suponha que $\mathbf{z}_1, \ldots, \mathbf{z}_n$ é uma amostra aleatória de pré-formas e $\mathbf{t}_1, \ldots, \mathbf{t}_n$ suas coordenadas tangentes. Seja $\mathbf{v}_i$ ser um vetor $2k-2$ o qual é obtido por empilhar as coordenadas real e imaginária de cada $\mathbf{t}_i$. Essa operação é representada por $cvec$ onde

$$\mathbf{v}_i = cvec(\mathbf{t}_i) = (\mathbf{Re}(\mathbf{t}_i)^\top, \mathbf{Im}(\mathbf{t}_i)^\top)^\top.$$

Assim o vetor de pré-formas $\mathbf{z}_i \in \mathbb{C}^{k-1}$ é representado nas coordenadas tangentes pelo vetor $\mathbf{v}_i \in \mathbb{R}^{2k-2}$. A distância Euclidiana no espaço tangente para o espaço de formas é uma boa aproximação para situações de alta concentração (DRYDEN e MARDIA, 1998, p. 76) das distâncias Procrustes d_F, ρ e d_P. E assim, pode-se aplicar os métodos multivariados padrões nas coordenadas tangentes.

As coordenadas tangentes parcial possuem grande utilidade nas análises de formas uma vez que pode-se utilizar as diversas técnicas multivariadas no espaço Euclidiano. Entretanto, pesquisadores já mostraram sua ineficiência em dados com baixa concentração. Como por exemplo, os testes de hipóteses das forma médias das pré-formas propostas por Amaral; Dryden e Wood (2007); a região de confiança bootstrap para a forma média planar desenvolvida por Amaral et al. (2010b).

2.1.5 Distribuição Bingham Complexa

A distribuição Bingham Complexa é uma das principais distribuições para análise de formas dos marcos bi-dimensional. Essa distribuição foi introduzida por Kent (1994). Essa é uma distribuição no espaço de vetores unitários complexo, ou equivalentemente, a esfera unitária complexa.

Para $k \geq 2$ seja $\mathbb{C}S^{k-1} = \{\mathbf{z} = (z_1, z_2, \ldots, z_k) : \sum |z_i|^2 = 1\} \subset \mathbb{C}^k$. A distribuição Bingham complexa tem função densidade de probabilidade

$$f(\mathbf{z}) = c(A)^{-1}\exp(\mathbf{z}^* A\mathbf{z}) \tag{2.15}$$

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

Denota-se essa distribuição por $\mathbb{C}B_{k-1}(A)$. Note que em Dryden e Mardia (1998, Cap. 6) desenvolve um papel importante em análise de formas para $k+1$ marcos em $\mathbb{R}^2$. A matriz A é a matriz parâmetro e $\mathrm{diag}(\lambda_1, \ldots \lambda_k)$ denotará seus autovalores.

Para simular a distribuição Bingham complexa, vários métodos foram propostos por Kent; Constable e Er (2004) . Nesta dissertação, será adotado o primeiro método, **Truncação pelo simplex**. Primeiro, gera-se $(k-1)$ exponenciais truncada, $\mathrm{Texp}(\lambda_j)$, pelo método de aceitação

e rejeição e, então, essas variáveis aleatórias são expressas em coordenadas polares para obter uma distribuição Bingham complexa.

O Algoritmo 1 representa os passos para simular a distribuição exponencial truncada multivariada através do método de aceitação e rejeição.

Algoritmo 1 Simulação da distribuição exponencial truncada

Output: Retorna Distribuição exponencial truncada $S_j \sim$Texp$(\lambda_j), \quad j = 1, \ldots, k-1$

1: Simule uma variável aleatória $U_j \sim U[0,1] \quad j = 1, \ldots, k-1$.
2: Seja $S'_j = -(1/\lambda_j)\log(1 - U_j(1 - e^{-\lambda_j}))$, $S' = (S'_1, \ldots, S'_{k-1})^\top$ de modo que S'_j são amostras aleatórias independentes $\mathrm{Texp}(\lambda_j)$.
3: Se $\sum_{j=1}^{k-1} S'_j < 1$, estabelece $S = S'$. Por outro lado, rejeita-se S' e retorne ao passo 1:

O método para simular uma distribuição Bingham complexa usa $(k-1)$ exponenciais truncadas para gerar um vetor k de uma distribuição Bingham complexa. Como descrito por Kent; Constable e Er (2004), sem perda de generalidade, deve-se fazer uma mudança nos autovalores de A de modo que eles são não positivos com o maior deles igual a 0. Sendo assim, é necessário fazer a mudança de tal forma que $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k = 0$ denota os autovalores de $-A$. Escreva $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{k-1})$ para o vetor dos primeiros $k-1$ autovalores. O $\{\lambda_j\}$ pode ser considerado como parâmetro de concentração.

Algoritmo 2 Simulação da Distribuição Bingham complexa

Output: Retorna Distribuição Bingham complexa

1: Gere $S = (S_1, \ldots, S_{k-1})$ onde $S_j \sim$Texp(λ_j) são variáveis aleatórias independentes usando Algoritmo 1.
2: Se $\sum_{j=1}^{k-1} S'_j < 1$, escreva $S_k = 1 - \sum_{j=1}^{k-1} S'_j$. Por outro lado, volte ao passo 1.
3: Gere ângulos independentes $\theta_j \sim U[0, 2\pi), \quad j = 1, \ldots, k$
4: Calcule $z_j = S_j^{1/2}\exp(i\theta_j), \quad j = 1, \ldots, k$

O Algoritmo 2 retorna um vetor k, $\mathbf{z} = (z_1, \ldots, z_k)$ o qual tem um distribuição Bingham complexa. Note que $(S_j^{1/2}, \theta_j)$ são essencialmente coordenadas polares para número complexo z_j.

2.2 Análise de Agrupamento

Agrupamento é uma área de análise multivariada que é usada em diversos campos do estudo, tais como reconhecimento de padrões, ciências sociais, biológicas, médicas entre outros. O objetivo da análise de agrupamentos é obter subgrupos homogêneos chamados de grupos, em que os objetos no mesmo grupo possuem característica mais similares (homogêneos), e os grupos possuem características heterogêneas entre si.

Análise de agrupamento é uma das técnicas mais conhecidas e populares de reconhecimento de padrão; assim, existem muitos modelos de agrupamento e algoritmos que analisam distribuições, densidades, possíveis pontos centrais, e um conjunto de dados. Os dois métodos mais conhecidos são o hierárquico e particionais.

O método de agrupamento hierárquico se subdivide em aglomerativo e divisivo. Os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os elementos. São populares na área de bioinformática e funcionam muito bem, apesar de não terem justificativa teórica baseada em estatística ou teoria da informação, constituindo uma técnica *ad-hoc* de alta efetividade. Para mais detalhes veja (JOHNSON et al., 1992).

Nos métodos 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 extremamente lento para grandes massas de dados. Assim, elevando o custo computacional.

Os métodos divisivos, por outro lado, possui abordagem contrária. Inicia-se a partir de um único grupo o qual então vai se dividindo iterativamente em grupos menores até cada elemento tornar-se seu próprio grupo. Cada um desses métodos possuem suas técnicas as quais podem ser vistas em (KAUFMAN e ROUSSEEUW, 2009), (JAIN e DUBES, 1988).

Os métodos particionais são organizados em dois paradigmas: rígido *(hard)* e difuso *(fuzzy)*. O método rígido tem como princípio particionar os dados em grupos mutuamente exclusivos. E um dos primeiros algoritmos com essa técnica é o algoritmo K-médias (MACQUEEN et al., 1967). O objetivo do agrupamento rígido é alocar cada elemento do conjunto de dados a um e somente um grupo. Em contraste, o agrupamento difuso 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 difuso o qual difere do conceito rígido de limites bem definidos.

Nesta dissertação, trabalha-se com algoritmos de particionamento para dados de formas, em especial o algoritmo K-médias e uma generalização deste, o algoritmo *Kernel* K-médias.

2.2.1 K-Médias

O algoritmo K-médias 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 (representante) do grupo é mínima. O primeiro uso do termo *K-means* foi em 1967 por James MacQueen em seu artigo intitulado *"Some Methods for Classification and Analysis of Multivariate Observation"*. Em 1957, Sturat Llody propôs o algoritmo *"Standard Algorithm"* como uma técnica de modulação de pulso que não tinha sido publicado fora dos laboratórios Bell até 1982. O algoritmo é conhecido como Lloyd Forgy pois em 1965 E. W. Fordy publicou o mesmo algoritmo. No período entre 1975 e 1979, uma versão mais eficiente foi proposta e publicada em Fortran por Hastigan e Wong.

Seja um conjunto de n objetos ou indivíduos a ser agrupados em um conjunto de K grupos, $C = (C_r, r = 1, \ldots, K)$. O algoritmo K-médias encontra uma partição minimizando um critério que mede distância entre pré-formas de grupos e formas média (**forma média Procrustes completa**).

$$J_1(C_r) = \sum_{i \in C_r} \mathrm{Dist}^2(\mathbf{z}_i, \boldsymbol{\mu}_r)$$

onde a função Dist^2 é uma medida de distância geral como as definidas anteriormente pelas Equações (2.12), (2.13) ou (2.14).

O objetivo do K-médias é minimizar a soma do erro quadrático sobre todo cluster K,

$$J_1 = \sum_{r=1}^{K} \sum_{i \in C_r} \mathrm{Dist}^2(\mathbf{z}_i, \boldsymbol{\mu}_r) \tag{2.16}$$

O Algoritmo 3 resume o passo a passo iterativo para obter o agrupamento pelo método K-médias em análise de formas.

Algoritmo 3 Método de Agrupamento K-médias para formas planas

Input: Pré-forma **z** (como definido na Equação (2.6)), número de grupos K e alocação inicial;
Output: Grupos C_r $(1 \leq r \leq K)$;

1: Obtenha a forma média para cada grupo;
2: Atribua cada objeto a forma média do grupo mais próximo, através das Equações (2.12), (2.13) ou (2.14).;
3: Calcule a forma média de cada grupo;
4: Repita o passo 2 e 3 até que a forma média não mude ou um valor ótimo da Equação (2.16) seja encontrada.

Este algoritmo move os objetos entre os agrupamentos até que a função objetivo não se altere ou se altere muito pouco, ou até que o número de iterações máximo pré determinado seja alcançado. O resultado é um conjunto de grupos com indivíduos com característica homogêneas dentro dos grupos e com características heterogêneas entre os grupos.

Apesar do algoritmo convergir rapidamente para uma solução, essa solução encontrada depende da alocação inicial, logo o método pode convergir para um ótimo local. Trata-se de um método prático e computacionalmente eficiente, porém é sensível a ruído, pontos aberrantes, o qual será estudado mais profundamente em trabalhos futuros.

2.2.2 *Kernel* K-Médias

Uma das desvantagens do método K-médias são os resultados pobres quando um conjunto de dados é composto por grupos não linearmente separáveis no espaço de entrada (*input space*). Para superar tal problema, uma solução possível é o método *Kernel* K-médias. Nesse método os pontos são mapeados para um espaço de característica (*feature space*) de alta dimensão usando uma função não linear, e então o *Kernel* K-médias particiona os pontos por separadores lineares no novo espaço.

O método de agrupamento *Kernel* K-médias é uma extensão não linear do método de agrupamento K-médias convencional. Esse é um método iterativo. Mapeia-se os pontos dos dados do espaço de entrada para uma alta dimensão no espaço característico através de uma transformação não linear $\phi(\cdot)$.

O *Kernel* baseado em métodos de aprendizagem usam um mapeamento implícito dos de entrada em um espaço característico de alta dimensão definido por função *Kernel*, isto é, uma função retorna o produto interno $\langle \phi(\mathbf{z}_1), \phi(\mathbf{z}_2) \rangle$ entre a imagem de dois pontos $\mathbf{z}_1$ e $\mathbf{z}_2$ no espaço característico. Isso geralmente é referenciado como o "truque *Kernel*". Dessa forma, o produto $\langle \phi(\mathbf{z}_1), \phi(\mathbf{z}_2) \rangle$ pode ser representado por uma função *Kernel* K_{er}

$$K_{er}(\mathbf{z}_i, \mathbf{z}_j) = \langle \phi(\mathbf{z}_i), \phi(\mathbf{z}_j) \rangle$$

Os conceitos teóricos de tais implicações podem ser estendidas para Manifolds. Em resumo, os pontos em manifold $\mathcal{M}$ são mapeados para elementos no espaço de Hilbert $\mathcal{H}$ de alta dimensão, a realização de Cauchy do espaço medido por valores reais de funções em $\mathcal{M}$, (JAYASUMANA et al., 2013). Uma função *Kernel* $K_{er} : (\mathcal{M} \times \mathcal{M}) \to \mathbb{R}$ é usada para definir o produto interno em $\mathcal{H}$, assim reproduz-se o espaço *Kernel* de Hilbert.

A dificuldade técnica em utilizar o espaço *Kernel* de Hilbert se dá precisamente nas satisfações das condições do teorema de Mercer, (MERCER, 1909), a função *Kernel* deve ser positiva definida para definir a validade do espaço. Uma função *Kernel* positiva definida em análise de forma é o *Procrustes Gaussian kernel*, em que apenas a distância Procrustes completa satisfaz as condições de Mercer. A sua demonstração está feita no artigo de Jayasumana et al. (2013). A função *Kernel* Gaussiana Procrustes é

$$K_{er}(\mathbf{z}_i, \mathbf{z}_j) = \exp\left(-\frac{d_F^2(\mathbf{z}_i, \mathbf{z}_j)}{2h}\right) = \exp\left(-\frac{1 - |\mathbf{z}_i^* \mathbf{z}_j|^2}{2h}\right) \tag{2.17}$$

em que d_F é um *Kernel* positivo definido para todo $h \in I\!R^+$ e h é um parâmetro de alisamento chamado largura de banda. A largura de banda do kernel é um parâmetro livre que possui forte influência sobre o resultado do processo de estimação.

O elemento (i, j) da matriz de similaridade das pré-formas é

$$\begin{aligned} D(i,j) = Dist^2(\phi(\mathbf{z}_i), \phi(\mathbf{z}_j)) &= d_F^2(\phi(\mathbf{z}_i), \phi(\mathbf{z}_j)) \\ &= 1 - |\phi(\mathbf{z}_i^*)\phi(\mathbf{z}_j)|^2 \\ &= 1 - |K_{er}(\mathbf{z}_i, \mathbf{z}_j)|^2 \quad i, j = 1, \ldots, n \end{aligned} \tag{2.18}$$

A função objetiva J_2 aplicada no espaço característico correspondente a função ϕ em análise de formas é definida como segue

$$J_2 = \sum_{r=1}^{K} \sum_{i \in C_r} \mathrm{Dist}^2(\phi(\mathbf{z}_i), \boldsymbol{\mu}_r^\phi) = \sum_{r=1}^{K} \left[\sum_{i \in C_r} 1 - |\phi(\mathbf{z}_i^*) \boldsymbol{\mu}_r^\phi|^2 \right] \tag{2.19}$$

onde, o valor do protótipo (r-ésimo) é expresso por

$$\boldsymbol{\mu}_r^\phi = \sum_{\mathbf{z} \in C_r} \frac{\phi(\mathbf{z})}{tam(C_r)} \tag{2.20}$$

em que $tam(C_r)$ indica o número de elementos dentro do grupo r. A expressão $\phi(\cdot)$ não está definida, e portanto, não é possível determinar explicitamente os protótipos dos grupos através da Equação (2.20). Neste caso, o modulo quadrático de um objeto $\phi(\mathbf{z}_i)$ e o centro do grupo $\boldsymbol{\mu}_r^{\phi}$ no espaço característico pode ser obtida da seguinte maneira

$$\begin{aligned}
\left|\phi(\mathbf{z}_i^*)\boldsymbol{\mu}_r^{\phi}\right|^2 &= \left|\phi(\mathbf{z}_i^*)\sum_{\mathbf{z}\in C_r}\frac{\phi(\mathbf{z})}{tam(C_r)}\right|^2 \\
&= \left|\sum_{\mathbf{z}\in C_r}\phi(\mathbf{z}_i^*)\frac{\phi(\mathbf{z})}{tam(C_r)}\right|^2 \\
&= \left|\sum_{\mathbf{z}\in C_r}\frac{K_{er}(\mathbf{z}_i,\mathbf{z})}{tam(C_r)}\right|^2 \\
&\overset{K_{er}\in\mathbb{R}^+}{=} \left(\sum_{\mathbf{z}\in C_r}\frac{K_{er}(\mathbf{z}_i,\mathbf{z})}{tam(C_r)}\right)^2
\end{aligned} \tag{2.21}$$

Dessa forma, concluímos que a função objetiva J_2 se resume a

$$J_2 = \sum_{r=1}^{K}\left[\sum_{i\in C_r}1 - \left(\sum_{\mathbf{z}\in C_r}\frac{K_{er}(\mathbf{z}_i,\mathbf{z})}{tam(C_r)}\right)^2\right] \tag{2.22}$$

em que $tam(C_r)$ representa o número de elementos no grupo C_r.

O algoritmo K-médias no espaço de características para dados de formas planas não possui propriamente uma etapa de atualização dos protótipos, pois o protótipo não pode ser explicitamente definido. No entanto, mesmo sem a definição explícita do protótipo é possível realizar uma etapa de alocação através do mapeamento implícito via função *Kernel*, apresentada na Equação (2.21). O algoritmo define uma matriz de *Kernel* $K_{er} = \{K_{er}(\mathbf{z}_i,\mathbf{z}_j)\}_{n\times n}$ e um conjunto de partições iniciais, e executa iterativamente reorganizando os indivíduos mais próximos em grupos até que a função objetivo J_2 alcance um valor estacionário, representando um mínimo local. O algoritmo 4 resume os passos iterativos para obter a partição do método de agrupamento *Kernel* K-médias em formas planas.

Algoritmo 4 Método de agrupamento *Kernel* K-médias para formas planas

Input: Pré-forma **z** como definido na Equação (2.6), número de grupos K e alocação inicial;
Output: Grupos C_r $(1 \leq r \leq K)$;

1: Calcule a matriz K_{er} utilizando a função Kernel Gaussiana Procrustes dada pela Equação (2.17) para $i = 1,\ldots,n$ e $j = 1,\ldots,n$.
2: Defina o cluster K cujo o indivíduo i está mais próximo, de acordo com $arg\ min_{r=1,\ldots,K}\left[\sum_{i\in C_r}1 - \left|\sum_{\mathbf{z}\in C_r}\frac{K_{er}(\mathbf{z}_i,\mathbf{z})}{tam(C_r)}\right|^2\right]$
3: Repita os passos 2 até os grupos não mudarem mais ou até valor ótimo da Equação (2.19) for encontrado

CAPÍTULO 3

Contribuições Algorítmicas e Teóricas

Agrupamento é uma das técnicas mais conhecidas de análise multivariada. Avaliar a qualidade dos resultados e determinar o número de grupos nos dados são questões importantes. O algoritmo K-médias, uma das técnicas de agrupamento de partição, é um dos mais utilizados. As vantagens do K-médias incluem eficiência computacional, implementação rápida e fácil conhecimento matemático. No entanto, o algoritmo possui limitações. Elas incluem uma escolha aleatória dos centroides (protótipos) no início do processo, o que pode ocasionar em uma solução de mínimo local; apresentam resultados pobres quando os dados são compostos por grupos não linearmente separáveis; e um número desconhecido de grupos K. A respeito da primeira limitação, vários ensaios podem ser uma solução. A segunda limitação pode ser resolvida com o algoritmo *Kernel* K-médias. Já a terceira questão está relacionada ao número de grupos e portanto a validade do agrupamento.

Agrupamento é por definição uma tarefa subjetiva e isso é o que torna complicado (JAIN; MURTY e FLYNN, 1999). Alguns desafios no agrupamento incluem i) o número de grupos presentes nos dados e ii) a qualidade do agrupamento, (MAULIK e BANDYOPADHYAY, 2002) . Elementos de respostas a essas duas questões são encontradas no campo de validação de cluster. Outros desafios, como condições iniciais e conjuntos de dados de alta dimensão são de importância em clustering. O objetivo das técnicas de validação de agrupamentos é o de avaliar os resultados de agrupamento. Esta avaliação pode ser usada para determinar o número de aglomerados dentro de cada conjunto de dados. Em algumas situações, pesquisadores fazem uso da função objetivo para encontrar o número correto de grupos. Porém de acordo com Jain e Dubes (1988) essa aproximação pode causar alguns problemas tais como número de amostras, características e grupos e a disseminação dos dados. A função objetivo, naturalmente, diminui à medida que o número de clusters aumentam e aumenta conforme o número de amostras e características aumentam. Exemplos desse problemas são discutidos por em Jain e Dubes (1988) e Nasse; Hébert e Hamad (2007).

Dessa forma, é preciso de ferramentas mais robustas para identificar o número correto de grupos para um conjunto de dados. A literatura atual contém vários exemplos de índice de validade, ou índice interno. E vários trabalhos tem sido feitos para avaliá-los tais como Jain e Dubes (1988), Kaufman e Rousseeuw (2009), Kim e Ramakrishna (2005).

Na literatura atual, encontra-se o trabalho Claude (2008) que trabalha com índices internos para coordenadas tangentes. E, não se percebe pesquisas e estudos detalhados sobre a validação do agrupamento ou técnicas que consigam encontrar o número ideal de grupos para um determinado conjunto de dados em análise de formas. Pelo motivo das coordenadas tangentes estarem no espaço euclidiano, os índices internos já encontrados na literatura podem ser utilizados, sem perda de generalidade. Porém no caso de pré-formas, que se encontram no espaço não Euclidiano devem ser utilizadas técnicas apropriadas para a natureza não Euclidiana. Assim, espera-se que índices internos próprios para análise de formas apresentem resultados mais convincentes.

As principais preocupações nessa dissertação são:

- desenvolver índices internos para avaliar o agrupamento em análise de formas;
- encontrar o número ideal de grupos para conjunto de dados.

E como atividade secundária, encontrar o melhor valor para a largura de banda h para o algoritmo *Kernel* K-médias.

3.1 Índices Interno existentes

Nesta seção, dois índices de validação adequados para análise de agrupamento de partição rígido são descritos. Bem como suas versões kernelizadas. Esses índices são medidas para avaliar os resultados das partições do conjunto de dados.

3.1.1 Índice Silhoueta

O índice silhoueta, (ROUSSEEUW, 1987), é uma técnica bem conhecida para estimar o número de grupos em um conjunto de dados. Considere $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ ser um conjunto de dados e seja $C = (C_r : r = 1, \ldots, K)$ ser o agrupamento em K grupos. A técnica de silhoueta atribui para cada i-ésimo indivíduo uma medida de qualidade $s(i), i = 1, \ldots, n$, conhecida como a largura da silhoueta. Esse valor é um indicador de confiança de associação do i-ésimo indivíduo no grupo C_r e é definido como

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{3.1}$$

em que $a(i)$ é a distância média entre o i-ésimo indivíduo e todos outros inclusos no grupo C_r, ao qual esse indivíduo pertence. $b(i)$ é a distância média mínima entre o i-ésimo indivíduo e todos os outros indivíduos dos grupos C_s em que $(s = 1, \ldots, K; s \neq r)$. Assim,

$$a(i) = \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2$$

$$b(i) = \min_{s=1,\ldots,K, s \neq r} \left[\frac{1}{tam(C_s)} \sum_{j \in C_s} ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 \right] \quad i \notin C_s \ (s = 1, 2, \ldots, K)$$

em que $|| \cdot ||$ indica a distância Euclidiana e $tam(C_r)$ representa o número de elementos no grupo r.

O índice Silhoueta (S) global é a média de todas silhoueta dos indivíduos do conjunto de dados, sendo assim

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

Uma versão kernelizada do índice Silhoueta para dados no espaço Euclidiano pode ser vista com maiores detalhes por Camps-Valls (2006, p.79,80). Para introduzir o mapeamento característico $\phi(\cdot)$, a(i) e b(i) podem ser expressos como

$$a^{\phi}(i) = \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2$$

$$b^{\phi}(i) = \min_{s=1,\ldots,K, s \neq r} \left[\frac{1}{tam(C_s)} \sum_{j \in C_s} ||\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)||^2 \right] \quad i \notin C_s \ (s = 1, 2, \ldots, K)$$

em que $||\cdot||$ indica a distância Euclidiana e $tam(C_r)$ representa o número de elementos no grupo r. Consequentemente, o índice silhoueta pode ser calculado no espaço característico como

$$s^{\phi}(i) = \frac{b^{\phi}(i) - a^{\phi}(i)}{\max[a^{\phi}(i), b^{\phi}(i)]} \tag{3.2}$$

3.1.2 Índice Davies-Bouldin

Seja $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ ser um conjunto de dados e seja $C = (C_1, \ldots, C_K)$ ser o agrupamento em K grupos. Considere $d(\mathbf{x}_i, \mathbf{x}_j)$ a distância Euclidiana entre $\mathbf{x}_i$ e $\mathbf{x}_j$. Então o índice Davies–Bouldin, (DAVIES e BOULDIN, 1979), pode ser calculado como segue

$$DB = \frac{1}{K} \sum_{r=1}^{K} \max_{s=1,\ldots,K, s \neq r} (R_{rs})$$

em que

$$R_{rs} = \frac{S_r + S_s}{||\mathbf{c}_r - \mathbf{c}_s||^2} \quad \text{e} \quad S_r = \frac{1}{tam(C_r)} \sum_{i \in C_r} ||\mathbf{x}_i - \mathbf{c}_r||^2$$

S_r é a distância dentro do grupo e $||\mathbf{c}_r - \mathbf{c}_s||^2$ é a distância euclidiana entre os centróides dos

grupos C_r e C_s, $tam(C_r)$ representa o número de elementos no grupo r. Desde que algoritmos produzam agrupamentos com distâncias baixas dentro do grupo (alta similaridade dentro do grupo) e alta distância entre grupos (baixa similaridade entre os grupos) terá baixos valores do índice Davies-Bouldin. Este é considerado o melhor algoritmo baseado nesse critério.

Segundo Kim e Ramakrishna (2005) o índice Davis-Bouldin possuem problemas. Dessa forma, redefiniu-se o índice DB como DB^*.

$$DB^* = \frac{1}{K}\sum_{r=1}^{K}\left(\frac{\max_{s=1,\ldots,K,s\neq r}(R_{rs})}{\min_{s=1,\ldots,K,s\neq r}||\mathbf{c}_r - \mathbf{c}_s||^2}\right)$$

Pode-se aumentar DB^* da seguinte forma

$$DB^{**} = \frac{1}{K}\sum_{r=1}^{K}\left(\frac{\max_{s=1,\ldots,K,s\neq r}(R_{rs}) + \mathrm{maxDiff}_r(K)}{\min_{s=1,\ldots,K,s\neq r}||\mathbf{c}_r - \mathbf{c}_s||^2}\right)$$

em que

$$\mathrm{maxDiff}_r(K) = \max_{K_{max},\ldots,K}\mathrm{diff}_r(K) \qquad \text{e}$$

$$\mathrm{diff}_r(K) = \max_{s=1,\ldots,K,s\neq r}(S_r(K) + S_s(K)) - \max_{s=1,\ldots,K+1,s\neq r}(S_r(K+1) + S_s(K+1))$$

Uma versão do índice Davies-Bouldin para o espaço característico *(feature space)* pode ser encontrada com maior detalhes em (NASSER; HÉBERT e HAMAD, 2007). Depois de transformar o conjunto $\mathbf{X}$ no espaço característico pela função ϕ, o índice DB^ϕ é fácil calcular. A versão kernelizada do índice Davies-Bouldin se resume a

$$DB^\phi = \frac{1}{K}\sum_{r=1}^{K}\max_{s=1,\ldots,K,s\neq r}(R_{rs}^\phi)$$

em que

$$R_{rs}^\phi = \frac{S_r^\phi + S_s^\phi}{||\mathbf{c}_r^\phi - \mathbf{c}_s^\phi||^2} \qquad \text{e} \qquad S_r^\phi = \frac{1}{tam(C_r)}\sum_{j\in C_r}||\phi(\mathbf{x}_i) - \mathbf{c}_r^\phi||^2$$

S_r^ϕ é a dispersão dentro do grupo kernelizada, $||\mathbf{c}_r^\phi - \mathbf{c}_s^\phi||^2$ é a distância entre os dois centróides kernelizados e $\mathbf{c}_r^\phi = \sum_{i\in C_r}\frac{\phi(\mathbf{x}_i)}{tam(C_r)}$ e $tam(C_r)$ é o número de elementos no grupo r.

Como já visto anteriormente, podemos aplicar as coordenadas tangentes a esses índices. Iremos fazer uso dessas técnicas para comparar com os índices propostos e modificados para pré-formas. Nas próximas Seções estamos dispostos a demonstrar ferramentas mais adequadas na descrição dos índices internos para análise de formas, fazendo uso da literatura que envolve essa técnica de pesquisa no espaço não Euclidiano. Logo, as próximas seções, estão descritos um novo método proposto fazendo uso dos resíduos Procrustes e métodos modificados que são extensões dos índices interno Silhoueta e Davies-Bouldin já descritos anteriormente para tratar dados de análise de formas, bem como suas correspondentes versões kernelizadas.

3.2 Índice Residual Procrustes (PR)

O **Índice Residual Procrustes** construído abaixo é útil quando as proximidades estão em uma escala de razão e quando se está à procura de grupos compactos e claramente separados. A metodologia proposta neste trabalho tem como objetivo encontrar o número ideal de grupos, e também avaliar a qualidade do ajuste em um conjunto de dados em *shape analysis*. Essa metodologia baseia-se no cálculo dos resíduos Procrustes definidos na expressão (2.11). A metodologia proposta consiste em duas etapas. Obter a alocação dos indivíduos do conjunto de dados nos grupos por alguma técnica tal como o algoritmo K-médias. Consequentemente, calcule a norma quadrática dos resíduos de cada indivíduo dentro do seu grupo (r_{in}) e fora do seu grupo (r_{out}).

Para cada indivíduo i, seja $r_{in}(i)$ a norma quadrática do resíduo Procrustes do indivíduo i para os indivíduos dentro do seu grupo. Pode-se interpretar $r_{in}(i)$ como tão bem i está atribuído ao seu grupo (valores pequenos, indicam que ele foi mais bem atribuído, ou seja, há grande similaridade do indivíduo i com o seu grupo). Seja $r_{out}(i)$ a menor norma quadrática do resíduo Procrustes do indivíduo i para cada outro grupo o qual i não é um membro. O grupo com a menor norma é conhecida como ser o "grupo vizinho" de i uma vez que esse está mais próximo do melhor grupo ajustado para o indivíduo i. Espera-se que $r_{out}(i)$ seja alto, logo as característica da instância i parecem ser mais similares ao seu verdadeiro grupo, em relação aos demais.

Seja Z^P as coordenadas Procrustes completa de um conjunto de dados, considere algum objeto i, e denote por C_r o grupo ao qual ele foi atribuído. Então, obtenha

$$r_{in}(i) = \left[Z_i^P - \left(\sum_{Z \in C_r}^{n} \frac{Z^P}{tam(C_r)} \right) \right]^* \left[Z_i^P - \left(\sum_{Z \in C_r}^{n} \frac{Z^P}{tam(C_r)} \right) \right].$$

Agora, considere um cluster C_s tal que $i \notin C_s (s = 1, \ldots, K)$, logo

$$r_{out}(i) = \min_{s=1,\ldots,K, s \neq r} \left\{ \left[Z_i^P - \left(\sum_{Z \in C_s}^{n} \frac{Z^P}{tam(C_s)} \right) \right]^* \left[Z_i^P - \left(\sum_{Z \in C_s}^{n} \frac{Z^P}{tam(C_s)} \right) \right] \right\}.$$

Com essas informações podemos definir o índice residual Procrustes $pr(i)$ para cada indivíduo do conjunto de dados como

$$pr(i) = \frac{r_{out}(i) - r_{in}(i)}{\max(r_{out}(i), r_{in}(i))}$$

Da definição acima, é fácil ver que

$$-1 \leqslant pr(i) \leqslant 1.$$

Valores próximos de 1 indicam que o indivíduo i possui dissimilaridade menor dentro do grupo comparando com outro grupo, logo o indivíduo está agrupado apropriadamente. Valores negativos ou próximos de -1 indicam que o indivíduo i pode ter sido alocado no grupo errado

uma vez que o coeficiente indica que este está mais próximo do grupo C_s do que C_r, portanto, parece ser mais apropriado, se este estivesse agrupado no "grupo vizinho". Quando o índice está próximo de 0 não há certeza da proximidade do indivíduo i ao grupo C_r ou C_s.

O Índice residual Procrustes (PR) global é o coeficiente médio de todas as instâncias i, isto é,

$$PR = \frac{1}{n}\sum_{i=1}^{n} pri(i)$$

e fornece uma avaliação da validade do cluster. Assim, a média de $pr(i)$ de todos os indivíduos do conjunto de dados é uma medida de como apropriadamente os dados estão agrupados. Se existem muitos ou poucos grupos, como pode ocorrer quando uma escolha pobre de K usado no algoritmo K-means, alguns dos grupos pode apresentar menores resíduos Procrustes que o resto. Dessa forma, o índice global pode ser usado para determinar o número natural de grupos dentro de um conjunto de dados.

3.3 Métodos modificados para Análises de Formas Planas

Nesta Seção, discuti-se extensões dos índices internos apropriados para dados de formas planas, em que devido a sua natureza, pertence a um espaço não Euclidiano. Logo, é útil detalhar e descrever ferramentas para esse tipo de dados, e consequentemente avaliar os resultados dos métodos de agrupamento e definir o número de grupos correto para tal conjunto de dados.

O índice Silhoueta, (ROUSSEEUW, 1987), é uma abordagem de validação de *cluster* interno para técnicas de partição. Essa silhoueta mostra quais objetos estão bem posicionados dentro de seu grupo, e quais deles estão meramente afastados dentro e entre os grupos. Altos valores do índice para cada objeto ou indivíduo indicam que este possui características similares com indivíduos do seu próprio grupo comparado com outros grupos. A média do coeficiente silhoueta fornece uma avaliação da validade do cluster, e deve ser usado para selecionar um número de grupos apropriados para determinado conjunto de dados.

3.3.1 Índice Silhoutte Modificado (Sm)

O índice silhoueta modificado para análise de formas envolve o espaço métrico não Euclidiano como já discutido na Seção 2.1.3. Dessa forma, para o cálculo do índice silhoueta modificado é necessário o uso das distâncias em análise de formas, estas já definidas pelas expressões (2.12), (2.13) e (2.14).

Suponha i como um objeto que pertence ao grupo C_r. Chame $a(i)$, também conhecida como dissimilaridade dentro do grupo, a distância média de i para todos os outros objetos de C_r. Considere $\mathbf{z}$ uma pré-forma da configuração X, logo $a(i)$ pode ser

$$
\begin{aligned}
a(i) &= \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} (1 - |\mathbf{z}_i^* \mathbf{z}_j|^2) \\
a(i) &= \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} \arccos(|\mathbf{z}_i^* \mathbf{z}_j|) \\
a(i) &= \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} 2(1 - \cos\rho) \quad i, j \in C_r.
\end{aligned} \tag{3.3}
$$

em que $tam(C_r)$ representa o número de elementos no grupo r.

Suponha C_s um grupo diferente de C_r. Defina $b(i)$, também conhecida como dissimilaridade entre grupos, como o mínimo sobre todos os grupos C_s diferentes de C_r da distância média de i a todos os objetos do grupo C_s.

$$
\begin{aligned}
b(i) &= \min_{s=1,\ldots,K, s \neq r} \left\{ \frac{1}{tam(C_s)} \sum_{j \in C_s} (1 - |\mathbf{z}_i^* \mathbf{z}_j|^2) \right\} \\
b(i) &= \min_{s=1,\ldots,K, s \neq r} \left\{ \frac{1}{tam(C_s)} \sum_{j \in C_s} \arccos(|\mathbf{z}_i^* \mathbf{z}_j|) \right\} \\
b(i) &= \min_{s=1,\ldots,K, s \neq r} \left\{ \frac{1}{tam(C_s)} \sum_{j \in C_s} 2(1 - \cos\rho) \right\} \quad i \notin C_s \ (s = 1, 2, \ldots, K).
\end{aligned} \tag{3.4}
$$

em que $tam(C_s)$ representa o número de elementos no grupo s.

Após identificar qual dissimilaridade dentro e entre os grupos definidos pelas expressões (3.3) e (3.4), respectivamente, para pré-formas. A largura da silhoueta modificada $sm(i)$ do objeto i é definida como segue

$$
sm(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{3.5}
$$

A silhoueta modificada global Sm é a média de todos $sm(i)$ do conjunto de dados. Isto é,

$$
Sm = \frac{1}{n} \sum_{i=1}^{n} sm(i)
$$

.

3.3.2 Índice Silhoueta Modificado kernelizado (Sm^{ϕ})

Após utilizar-se o método de agrupamento Kernelizado, ou melhor, o algortimo *Kernel* K-médias não é adequado utilizar o índice Silhoueta Modificado para validar os grupos. Dessa forma, é necessário definir um novo índice que leve em consideração a natureza do espaço característico dos dados.

Antes de definir o índice silhoueta modificado kernelizado, devemos relembrar que apenas

a distância Procrustes completa satisfaz as condições *Kernel* de Mercer, como demonstrado por Jayasumana et al. (2013). Sendo assim, as dissimilaridades dentro e entre os grupos so podem ser definidas fazendo uso da distância Procrustes completa.

Seja $\mathbf{z}$ a pré-forma de uma configuração X e seja $C = (C_r : r = 1, \ldots, K)$ os grupos de um conjunto de dados. A dissimilaridade dentro dos grupos $a^{\phi}(i)$ é definida como

$$\begin{aligned} a^{\phi}(i) &= \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} \left[1 - |\phi(\mathbf{z}_i^*)\phi(\mathbf{z}_j)|^2\right] \\ &= \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} \left[1 - |\phi(\mathbf{z}_i^*)\phi(\mathbf{z}_j)|^2\right] \\ &= \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} \left[1 - |K_{er}(\mathbf{z}_i, \mathbf{z}_j)|^2\right] \\ &\overset{K_{er} \in \mathbb{R}^+}{=} \frac{1}{(tam(C_r)-1)} \sum_{j \in C_r} \left[1 - K_{er}(\mathbf{z}_i, \mathbf{z}_j)^2\right] \quad i, j \in C_r \end{aligned} \tag{3.6}$$

em que $tam(C_r)$ representa o número de elementos no grupo r.

A dissimilaridade entre os grupos $b^{\phi}(i)$ será definida como

$$\begin{aligned} b^{\phi}(i) &= \min_{s=1,\ldots,K, s \neq r} \left\{ \frac{1}{tam(C_s)} \sum_{j \in C_s} \left[1 - |\phi(\mathbf{z}_i^*)\phi(\mathbf{z}_j)|^2\right] \right\} \\ &= \min_{s=1,\ldots,K, s \neq r} \left\{ \frac{1}{tam(C_s)} \sum_{j \in C_s} \left[1 - |K_{er}(\mathbf{z}_i, \mathbf{z}_j)|^2\right] \right\} \\ &\overset{K_{er} \in \mathbb{R}^+}{=} \min_{s=1,\ldots,K, s \neq r} \left\{ \frac{1}{tam(C_s)} \sum_{j \in C_s} \left[1 - K_{er}(\mathbf{z}_i, \mathbf{z}_j)^2\right] \right\} \quad i \notin C_s \ (s = 1, 2, \ldots, K) \end{aligned} \tag{3.7}$$

em que $tam(C_s)$ representa o número de elementos no grupo s.

Consequentemente, a largura da silhoueta modificada kernelizada para cada indivíduo i do conjunto de dados pode ser computado no espaço característico como

$$sm^{\phi}(i) = \frac{b^{\phi}(i) - a^{\phi}(i)}{\max[a^{\phi}(i), b^{\phi}(i)]} \tag{3.8}$$

A silhoueta modificada kernelizada global Sm^{ϕ} é a média de todos $sm^{\phi}(i)$ do conjunto de dados. Ou seja,

$$Sm^{\phi} = \frac{1}{n} \sum_{i=1}^{n} sm^{\phi}(i)$$

.

Segue abaixo propriedades do índice silhoueta modificada e kernelizada.

1. $sm(i)$ e $sm^{\phi}(i)$ assumem valores entre $[-1, 1]$;
2. valores de $sm(i)$ e $sm^{\phi}(i)$ próximo de 1 indica que a dissimilaridade dentro do grupo é

muito menor em relação a dissimilaridade entre outros grupos, ou seja, o objeto está bem alocado;

3. valores de $sm(i)$ e $sm^{\phi}(i)$ próximo de -1 indica que a dissimilaridade dentro do grupo é maior em relação a dissimilaridade entre outros grupos, ou seja, o objeto foi mal alocado, uma vez que que está mais próximo do outro grupo ao qual não foi alocado;

4. valores de $sm(i) \simeq 0$ e $sm^{\phi}(i) \simeq 0$ tem-se uma indecisão em considerar qual o grupo mais adequado para o indivíduo.

3.3.3 Índice Davies-Bouldin Modificado (DBm)

O índice Davies-Bouldin, (DAVIES e BOULDIN, 1979), é uma das técnicas de validação interna em análise de agrupamento que avalia a qualidade de um resultado com base em propriedades estatísticas, como a combinação das distâncias entre os indivíduos e a forma média dentro e entre os grupos em relação a distância da forma média entre os grupos. Valores pequenos desse índice indicam que o resultado do agrupamento é válido. Pelo motivo da medida não depender do número de grupos analisados e do método (algoritmo) de partição do conjunto de dados, pode-se utilizá-lo para encontrar o número (K) ideal de grupos.

O índice Davies-Bouldin modificado para análise de formas envolve o espaço métrico não Euclidiano como já discutido na Seção 2.1.3. Dessa forma, para o cálculo do índice é necessário o uso das distâncias em análise de formas, estas já definidas pelas expressões (2.12), (2.13) e (2.14).

Suponha i ser um objeto que pertence ao grupo C_r. Considere $\mathbf{z}$ uma pré-forma da configuração X. O índice Davies-Bouldin é calculado pela média dos pares dos grupos como

$$
\begin{aligned}
DBm &= \frac{1}{K}\sum_{r=1}^{K} \max_{s=1,\ldots,K, s\neq r}(R_{rs}) \quad \text{com} \\
R_{rs} &= \frac{D_r + D_s}{\text{Dist}^2(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)} \quad \text{e} \\
D_r &= \frac{1}{tam(C_r)} \sum_{i \in C_r} \text{Dist}^2(\mathbf{z}_i, \boldsymbol{\mu}_r)
\end{aligned}
$$

em que K denota o número de grupos, r, s são os níveis dos grupos, $tam(C_r)$ representa o número de elementos no grupo r, então D_r e D_s são as distâncias médias de todos indivíduos em cada grupo para sua respectiva forma média no grupo $\boldsymbol{\mu}_r$ e $\boldsymbol{\mu}_s$; $\text{Dist}^2(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$ é a distância entre as formas médias (definida no Resultado 1), em formas planas, a distância $Dist^2(\cdot, \cdot)$ é umas das já definidas em (2.12), (2.13) e (2.14). Assim, obtém-se o índice Davies-Bouldin Modificado.

Como descrito por Kim e Ramakrishna (2005) o índice Davies-Bouldin sofre alguns problemas, e o índice modificado para análises planas também. Sendo assim, propõem dois novos

índices ajustados. O índice DBm é a média do máximo de R_{rs} de cada grupo, e tem o máximo nas seguintes situações:

1. quando $d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$ domina;
2. quando $(D_r + D_s)$ domina;
3. pela combinação de $d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$ e $(D_r + D_s)$;

A expressão 'Dominar' indica que é um fator decisivo em determinar o max(R_{rs}).

No caso de 1., $d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$ tem um valor relativamente pequeno comparado com $(D_r + D_s)$ e é usualmente min($d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$). Isso indica uma situação onde dois grupos estão localizados bem próximos um do outro e eles precisam ser fundidos.

No caso de 2. $(D_r + D_s)$ tem valores relativamente muito altos, usualmente quando $(D_r + D_s) = \max(D_r + D_s)$. Isso revela que uma junção desnecessária foi tomada.

No caso 3. R_{rs} tem o valor máximo quando nem min($d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$) nem max$(D_r + D_s)$ ocorre, isto é, por algumas combinações de $d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$ e $(D_r + D_s)$.

Resumindo, observa-se que DBm pode efetivamente ter um valor ótimo, isso é, o valor mínimo quando $d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$ domina $K > K_{otimo}$ e $(D_r + D_s)$ domina $K < K_{otimo}$. A partir da suposição que em uma situação $1/\min d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)$ e max$(D_r + D_s)$ tem valores relativamente altos quando $K > K_{otimo}$ e $K < K_{otimo}$, respectivamente. Redefine-se o índice DBm como DBm^*.

$$DBm^* = \frac{1}{K} \sum_{r=1}^{K} \left(\frac{\max_{s=1,\ldots,K, s \neq r}(R_{rs})}{\min_{s=1,\ldots,K, s \neq r} d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)} \right)$$

Outra suposição oculta sobre esse índice é abordado. Refere-se à condição de que o padrão indicado da distância intra-classe (da) e a distância inter-classe (de) em torno do $K = K_{otimo}$ ocorre apenas uma vez. Entretanto, isso não é garantido em aplicações do mundo real.

Por exemplo, se um amontoado de grupos com similares distâncias intra-classe são separados por distâncias similares, o padrão pode ocorrer várias vezes em $K < K_{otimo}$, e além disso, eles podem ter alterações nas configurações. Isto é, mesmo no $K < K_{otimo}$ pode haver situações em que distância intra-classe em K seja menor que a distância inter-classe em K, isto é o índice de validade para K será menor que o índice de validade para K_{otimo} quando estabelecemos $da(K)$ e $de(K)$ como $da(K)/da(K_{otimo})$ e $de(K)/de(K_{otimo})$, respectivamente.

Antes de apresentar alternativas para aliviar este problemas, fazemos algumas observações. Como para $K < K_{otimo}$, um aumento no valor de da leva a um aumento no valor do índice de validade. Isso é desejável em ordem para o índice de validade para ter o mínimo em $K = K_{otimo}$. Vamos definir dois termos como segue-se: $\mathrm{diff}_r(K) = \max_{s=1,\ldots,K, s \neq r}(D_r(K) + D_s(K)) - \max_{s=1,\ldots,K+1, s \neq r}(D_r(K+1) + D_s(K+1))$ e $\mathrm{maxDiff}_r(K) = \max_{K_{max},\ldots,K} \mathrm{diff}_r(K)$. $\mathrm{maxDiff}_r(K)$ tem as seguintes propriedades:

(i) Valores relativamente pequenos de $K \geq K_{otimo}$,

(ii) Valores grandes somente em $K < K_{otimo}$.

Essa são propriedades desejáveis da distância intra-classe. Além disso, $\mathrm{maxDiff}(K)$ pode ser aumentado da distância intra-classe sem qualquer efeito colateral devido a essas propriedades. Assim, propõe-se novo índice de validade, DBm^{**}, aumentado por $\mathrm{maxDiff}(K)$ de DBm^{*} da seguinte maneira

$$DBm^{**} = \frac{1}{K}\sum_{r=1}^{K}\left(\frac{\max_{s=1,\ldots,K,s\neq r}(R_{rs}) + \mathrm{maxDiff}_r(K)}{\min_{s=1,\ldots,K,s\neq r} d(\boldsymbol{\mu}_r, \boldsymbol{\mu}_s)}\right)$$

3.3.4 Índice Davies-Bouldin Modificado kernelizado (DBm^{ϕ})

Uma modificação desse índice para amostras que estão mapeadas no espaço de entrada para uma dimensão maior no espaço característico através de uma transformação não linear pode ser chamado como índice Davies-Bouldin modificado kernelizado (DBm^{ϕ}) dado por

$$\begin{aligned} DBm^{\phi} &= \frac{1}{K}\sum_{r=1}^{K}\max_{s=1,\ldots,K,r\neq s}(R_{rs}) \quad \text{com} \\ R_{rs} &= \frac{D_r^{\phi} + D_s^{\phi}}{\mathrm{Dist}^2(\boldsymbol{\mu}_r^{\phi}, \boldsymbol{\mu}_s^{\phi})} \quad \text{e} \\ D_r^{\phi} &= \frac{1}{tam(C_r)}\sum_{i\in C_r}\mathrm{Dist}^2(\phi(\mathbf{z}_i), \boldsymbol{\mu}_r^{\phi}) \end{aligned}$$

em que K denota o número de grupos, C_r representa o número de elementos no grupo r, r, s são os níveis dos grupos, então D_r^{ϕ} e D_s^{ϕ} são as distâncias médias de todos indivíduos em cada grupo r e s para sua respectiva forma média no espaço característico em seu grupo $\boldsymbol{\mu}_r^{\phi}$ e $\boldsymbol{\mu}_s^{\phi}$, seja $\boldsymbol{\mu}^{\phi} = \frac{1}{|C|}\sum_{\mathbf{z}\in C}\phi(\mathbf{z}^{*})$, então

$$
\begin{aligned}
D_r^{\phi} &= \frac{1}{tam(C_r)} \sum_{i \in C_r} \mathrm{Dist}^2(\phi(\mathbf{z}_i), \boldsymbol{\mu}_r^{\phi}) \\
&= \frac{1}{tam(C_r)} \sum_{i \in C_r} (1 - |\phi(\mathbf{z}_i^*) \boldsymbol{\mu}_r^{\phi}|^2) \\
&= \frac{1}{tam(C_r)} \sum_{i \in C_r} \left[1 - \left| \phi(\mathbf{z}_i) \sum_{j \in C_r} \frac{\phi(\mathbf{z}_j)}{|C_r|} \right|^2 \right] \\
&= \frac{1}{tam(C_r)} \sum_{i \in C_r} \left[1 - \left| \sum_{j \in C_r} \frac{\phi(\mathbf{z}_i^*)\phi(\mathbf{z}_j)}{|C_r|} \right|^2 \right] \\
&= \frac{1}{tam(C_r)} \sum_{i \in C_r} \left[1 - \left| \sum_{j \in C_r} \frac{K_{er}(\mathbf{z}_i, \mathbf{z}_j)}{|C_r|} \right|^2 \right] \\
&\overset{K_{er} \in \mathbb{R}^+}{=} \frac{1}{tam(C_r)} \sum_{i \in C_r} \left[1 - \left(\sum_{j \in C_r} \frac{K_{er}(\mathbf{z}_i, \mathbf{z}_j)}{|C_r|} \right)^2 \right]
\end{aligned}
$$

$\mathrm{Dist}^2(\boldsymbol{\mu}_r^{\phi}, \boldsymbol{\mu}_s^{\phi})$ é a distância entre as formas médias no espaço característico. Fazendo uso da equação (2.18) obtém-se

$$
\begin{aligned}
\mathrm{Dist}^2(\boldsymbol{\mu}_r^{\phi}, \boldsymbol{\mu}_s^{\phi}) &= (1 - |\boldsymbol{\mu}_r^{\phi *} \boldsymbol{\mu}_s^{\phi}|^2) \\
&= \left(1 - \left| \sum_{i \in C_r} \frac{\phi(\mathbf{z}_i^*)}{tam(C_r)} \sum_{j \in C_s} \frac{\phi(\mathbf{z}_j)}{tam(C_s)} \right|^2 \right) \\
&= \left(1 - \left| \sum_{i \in C_r} \sum_{j \in C_s} \frac{\phi(\mathbf{z}_i^*)\phi(\mathbf{z}_j)}{tam(C_r)tam(C_s)} \right|^2 \right) \\
&= \left(1 - \left| \sum_{i \in C_r} \sum_{j \in C_s} \frac{K_{er}(\mathbf{z}_i, \mathbf{z}_j)}{tam(C_r)tam(C_s)} \right|^2 \right) \\
&\overset{K_{er} \in \mathbb{R}^+}{=} \left[1 - \left(\sum_{i \in C_r} \sum_{j \in C_s} \frac{K_{er}(\mathbf{z}_i, \mathbf{z}_j)}{tam(C_r)tam(C_s)} \right)^2 \right]
\end{aligned}
\tag{3.9}
$$

É desejável que os grupos tenham o mínimo possível de similaridade entre eles. Por isso buscamos por grupos que minimizem DBm e DBm^{ϕ}. Os índices DBm e DBm^{ϕ} não apresentam tendência com respeito ao número de grupos e assim busca-se o valor mínimos desses índices em gráficos contra o número de grupos. Valores pequenos de DBm e DBm^{ϕ} correspondem a grupos mais compactos, e cujos centros (forma média) estão distantes um do outro. Consequentemente, o número de grupos que minimiza esses índices são tomados como o número ótimo de grupos.

CAPÍTULO 4

Avaliação Numérica - Apresentação e Resultados

A proposta dessa Seção é avaliar os índices existentes dispostos na Seção 3.1 para lidar com as coordenadas tangentes e os índices propostos nas Seções 3.2 e 3.3 próprios para pré-formas como ferramentas de estimação para os parâmetros dos algoritmos K-médias e *Kernel* K-médias em dados de formas planas. Para o segundo algoritmo usamos a distância Procrustes completa aplicada à função do *Kernel* Gaussiano, uma vez que esta é a única distância que satisfaz as condições de Mercer, como provado em Jayasumana et al. (2013). Essa função possui capacidade de fazer grupos separáveis que são inicialmente compactos e contínuos, mas não linearmente separáveis.

O algoritmo K-médias precisa de um parâmetro, o número de grupos K, já o *Kernel* K-médias Gaussiano necessita, além deste, a largura de banda (h). Realizou-se uma estimação a posteriori deles conjuntamente, por considerar que a melhor escolha dos parâmetros representa a melhor qualidade dos agrupamentos, isto é, altos valores dos índices Silhoueta e Procrustes Resíduos e baixos valores do índice Davies-Bouldin.

Com o objetivo de validar os métodos propostos para os dados de morfometria, foram realizados experimentos para conjuntos de dados sintéticos e conjunto de dados reais. Os conjuntos de dados sintéticos foram gerados com diferentes graus de dificuldade no agrupamento com grupos de formas e tamanhos diferentes, dados mais heterogêneos (com características diferentes) e outros mais homogêneos (com características iguais) . Para cada conjunto de dados simulados o índice interno é estimado em 100 réplicas de Monte Carlo. A finalidade da aplicação do método Monte Carlo é propiciar uma melhor avaliação quantitativa do desempenho dos métodos considerando situações com diferentes graus de dificuldades de agrupamento.

Para ambos os parâmetros K e h, que indicam o número de grupos e a largura de banda para o algoritmo *Kernel* K-médias, respectivamente, alguns valores podem ser escolhidos razoavelmente a priori. Nos dados simulados e reais, consideramos $K = 2, 3, \ldots, 8$ e para $h = [h_{min}; h_{max}]$ onde h_{min} e h_{max} detonam, respectivamente, o mínimo e o máximo definido

para o parâmetro com valores espaçados igualmente.

4.1 Avaliação Numérica - Simulações

Para avaliar a performance dos índices internos propostos e modificados para análises de formas planas, dois experimentos foram realizados. O primeiro compreende simulações de marcos no espaço dos marcos e com diferentes variabilidades. E o segundo, envolve simulação no espaço das pré-formas, que corresponde a uma esfera no espaço complexo através da distribuição Bingham complexa, são considerados diferentes níveis de concentração. As Tabelas com os resultados das análises estão nos Apêndices A e B. Essas Tabelas apresentam os índices propostos para pré-formas e os índices existentes na literatura para coordenadas tangentes, o número de grupos, a largura de banda escolhido para os índices Kernelizados e a taxa de acerto (Tx.) a qual é calculada como a quantidade de vezes que o índice encontrou o número correto de grupos para as 100 réplicas Monte Carlo.

4.1.1 Simulações no espaço dos marcos

Dessa forma, faz-se simulações dos *landmarks* no espaço dos *landmarks*. São considerados seis marcos, os quais são descritos por pontos no $\mathbb{R}^2$ seguindo distribuição normal bivariada de componentes independentes, Equação (4.1). Cada conjunto de dados possui n indivíduos divididos em dois grupos. O objetivo é gerar sobreposição entre as configurações para valores distintos da matriz de covariância.

$$(x_l^\top, y_l^\top) \sim N_{12}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \quad l = 1 \text{ e } 2 \tag{4.1}$$

onde, $\boldsymbol{\mu}_l = (\boldsymbol{\mu}_{xl}^\top, \boldsymbol{\mu}_{yl}^\top)^\top$ e $\boldsymbol{\Sigma}_l = \sigma_l \mathbf{I}_{12}$, são o vetor média e a matriz de covariância, respectivamente, denotado como

$$\boldsymbol{\Sigma}_l = \begin{pmatrix} \sigma^2\mathbf{I}_6 & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_6 \end{pmatrix}, \quad \boldsymbol{\mu}_l = \begin{pmatrix} \boldsymbol{\mu}_{xl} \\ \boldsymbol{\mu}_{yl} \end{pmatrix}$$

Assim esse modelo é isotrópico. Isotropia significa que cada Marco tem, aproximadamente, a mesma variabilidade. Nessa simulação é considerado que os marcos são independentes.

O mesmo vetor média é considerado para os cenários, sendo assim,

$$\begin{aligned} \boldsymbol{\mu}_1 &= (\boldsymbol{\mu}_{x1}^\top, \boldsymbol{\mu}_{y1}^\top)^\top = (1,1,2,0,0,0,0,0,0,0,0,0)^\top, \\ \boldsymbol{\mu}_2 &= (\boldsymbol{\mu}_{x2}^\top, \boldsymbol{\mu}_{y2}^\top)^\top = (1,1,2,5,5,5,-1,-1,-1,2,2,2)^\top \end{aligned}$$

No Apêndice A encontram-se os resultados dos índices de validação existentes e os propostos para dados de formas planas. A Tabela A.1 refere-se aos dados de alta concentração com $\sigma = 0,1$. Pode-se observar que todos os índices tanto para coordenadas tangentes quanto para

pré-forma apresentaram alta taxa de acerto, referente ao número correto de grupos no conjunto de dados.

A segunda configuração é dada por $\sigma = 0,3$, na Tabela A.2, pode-se observar que os índices Davies-Bouldin e Davies-Bouldin Kernelizados para tamanho de amostra $n = 30$ esses índices existentes para coordenadas tangentes não conseguem estimar o verdadeiro número de grupos do conjunto de dados simulados. No caso de $n = 50$, neste cenário, é possível identificar o número correto de grupos.

Na geração de dados com baixa concentração, dado pela configuração $\sigma = 0,5$, na Tabela A.3, para tamanho de amostra $n = 30$ os índices DB_m e DB_m^* e DB_m^ϕ não são bons estimadores no caso das pré-formas; considerando as coordenadas tangentes, os índices Davies-Bouldin não satisfazem nossa necessidade de estimar o número de grupos. Quando simulamos as amostras com tamanho $n = 50$, neste mesmo cenário, apenas o índice DB_m para pré-formas erra na identificação do número de grupos, enquanto que para coordenadas tangentes todos os índices Davies-Bouldin erram na estimação. A razão é que com a baixa concentração medidas de distância no espaço tangente não representam bem a real magnitude. Isso ocorre porque o espaço tangente é uma aproximação linear da esfera estudada.

4.1.2 Simulação Distribuição Bingham Complexa

Para gerar dois grupos com distribuição Bingham complexa, com densidade definida em (2.15), utilizou-se $k = 3$ e diferentes valores para a concentração $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ para diferentes tamanhos de amostras. Um dos grupos das amostras foi rotacionado para obter dois grupos com médias diferentes. E para tanto, o grau de rotação para as análises foi de 0.05π. Para garantir que as amostras geradas possuem diferenças estatísticas significativas, foram considerados os testes apresentados por Amaral; Dryden e Wood (2007) para diferenças de médias. Ou seja, iremos testar se as médias dos dois grupos são diferentes, se o *p-valor* do teste for menor que o nível de significância ($\alpha = 5\%$, adotado) rejeita-se a hipótese nula e conclui-se que os grupos possuem médias diferentes.

Diferentes cenários foram considerados: dois com alta concentração e dois com baixa concentração. Os resultados estão inseridos no Apêndice B. A Tabela B.1 representa os resultados de amostras de tamanho $n = 30$ e $n = 50$ para a distribuição Bingham complexa, em que as duas amostras possuem alta concentração, ou seja, $\boldsymbol{\lambda} = (200, 200, 0)$ e pode-se observar que a taxa de acerto do número de grupos é alta para todos os índices, tanto para coordenadas tangentes que pertence ao espaço Euclidiano e para as pré-formas que não estão no espaço Euclidiano.

No segundo cenário, diminuímos a concentração para a geração da distribuição, nesse caso $\boldsymbol{\lambda} = (100, 100, 0)$. E, segundo a Tabela B.2, pode-se notar que a taxa de acerto dos índices diminuem, mas os índices não deixam de estimar o número correto de grupos no conjunto de dados.

A Tabela B.3 retrata os resultados dos dados simulados para $\boldsymbol{\lambda} = (2; 1,5; 0)$ indicando baixa concentração na geração dos dados da distribuição Bingham. Os índices para coordenadas tangentes não conseguem estimar o verdadeiro número de grupos do conjunto de dados neste

cenário. Já para os índices modificados e proposto para pré-forma, apenas os índices Silhoueta S_m, Silhoueta Kernelizado S_m^{ϕ} e Davies-Bouldin DB_m e sua versão Kernelizada DB_m^{ϕ} não são bons estimadores para determinar o parâmetro K.

No último cenário, este contém a menor concentração com $\boldsymbol{\lambda} = (1; 1; 0)$, pode-se notar que apenas os índices Procrustes Residual e o índice Davies-Bouldin modificado aumentado DB_m^{**} para pré-forma conseguem estimar o verdadeiro valor do número de grupos. Isso realmente diferencia os índices propostos em relação aqueles usais para o espaço tangente.

4.2 Avaliação Numérica - Dados Reais

O problema de inicialização dos algoritmos K-médias para cada parâmetro K e *Kernel* K-médias para cada par de parâmetros (K, h) é resolvido por computar os algoritmos 50 vezes com diferentes alocações iniciais. Para a largura de banda, nos dados reais, usamos 20 valores entre $h = [h_{min}; h_{max}] = [0{,}1; 2{,}0]$ igualmente espaçados. Somente os valores máximos dos índices Silhoueta Modificado e sua versão Kernelizada, do índice Procrustes Residual e os valores mínimos dos índices Davies-Bouldin Modificado e sua versão Kernelizada são apresentados nas tabelas e nos gráficos.

Finalmente, nós apresentamos para cada exemplo o agrupamento obtido com o número conhecido a priori K de grupos, que deve corresponder aos melhores valores dos índices. Isso, permite verificar a precisão dos agrupamentos.

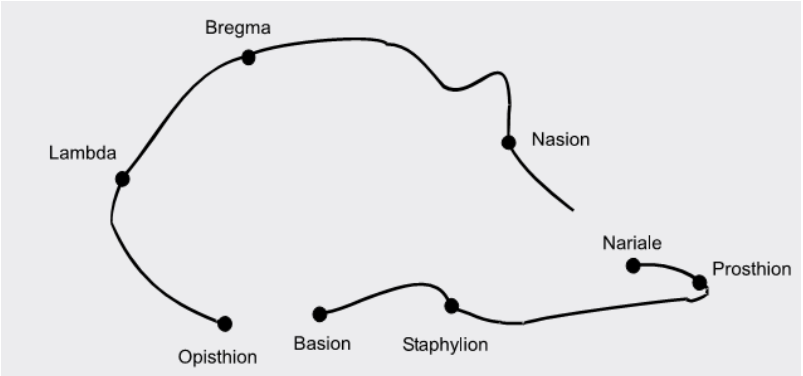
4.2.1 Crânios de Gorilas

Em um estudo de avaliação de diferenças do crânio entre os sexos de macacos, 29 gorilas machos e 30 gorilas fêmeas adultos foram tomados. Esse conjunto de dados foram produzidos por O'Higgins (1989) e analisado por O'Higgins e Dryden (1993). Oito marcos foram escolhidos na linha média do plano de cada crânio como descrito na Figura 4.1 desenvolvida por Hammer e Harper (2008). Os marcos são marcos anatômicos e foram alocados por um biólogo experiente.

Os dados brutos como mostrados na Figura 4.3(a) dos marcos dos gorilas machos e fêmeas são difíceis de analisar diretamente, porque as espécies são de tamanhos diferentes e têm sido posicionado e orientado diferentemente pelo biólogo experiente. Dessa forma, para analisar o agrupamento do conjunto de dados será usado as pré-formas. As pré-formas não possuem representação gráfica, dessa forma, a Figura 4.3(b) tem-se as pré-formas centradas, esta já possui, como definido na Definição 3. A Figura 4.3(c) representa as coordenadas dos crânios de cada gorila macho e fêmea após o ajuste dos Procrustes completa, já a Figura 4.3(d) indica as formas médias de acordo com o sexo dos animais.

Na Tabela 4.1, tem-se os valores máximo dos índices silhoueta modificados e do Procrustes resíduos e o mínimo dos índices Davies-Bouldin versus o número de grupos. Ainda, para o *Kernel* K-médias, deve-se encontrar o valor ideal para a largura de banda, e através das Figuras 4.4(a) e 4.4(b), pode-se notar que serão utilizados os valores $(h = 0,4)$ e $(h = 1,1)$, para

Figura 4.1: Oito marcos anatômicos no plano médio de um crânio do gorila.



Fonte: (HAMMER e HARPER, 2008).

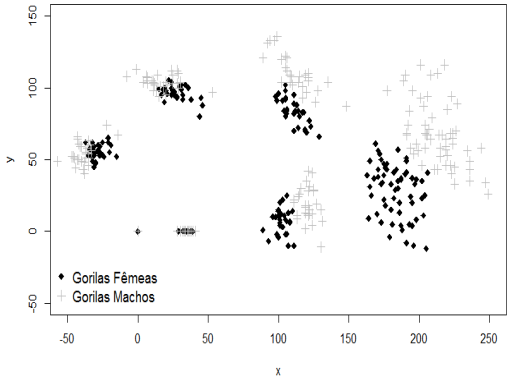
os índices Silhoueta Kernelizado S_m^{ϕ} e Davies-Bouldin Kernelizado DB_m^{ϕ}, uma vez que estes remetem os valores ótimos dos respectivos índices.

Nota-se que os índices existentes e os modificados para formas planas detecta bem o verdadeiro número de grupos, $K = 2$, para o conjunto de dados dos gorilas que se dividem em grupo de fêmeas e machos. Outra questão a ser analisada é, os índices internos kernelizados apresentam valores melhores, ou seja, o índice Sm^{ϕ} é maior que Sm e o índice DBm^{ϕ} é menor do que DBm. Isso indica que ao fazer análise de agrupamento com esses dados, aconselha-se o uso do algoritmo *Kernel* K-médias. Devido ao conhecimento a priori dos indivíduos, podemos calcular a taxa de acerto do agrupamento para cada algoritmo. O valor médio, das 50 inicializações, da taxa de acerto para o algoritmo *Kernel* K-médias é 0.9491 enquanto que para o algoritmo K-médias é 0.9152, assim, verificamos que o índice interno nos auxilia na escolha do melhor método para agrupamento dos dados.

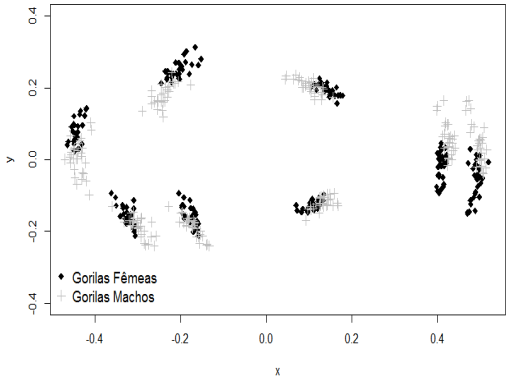
Tabela 4.1: Validação do agrupamento do conjunto de dados dos gorilas.

		Número de grupos						
	Índices	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	PR	**0,2797**	0,2193	0,2260	0,2156	0,1831	0,2314	0,1692
	S_m	**0,2485**	0,2176	0,1891	0,1709	0,1708	0,1576	0,1276
	DB_m	**1,4918**	1,7450	1,7200	1,7132	1,7491	1,6960	1,6785
Pré-Forma	DB_m^*	**1,5281**	1,8758	1,8866	1,9248	1,9722	1,9603	1,9545
	DB_m^{**}	**1,3918**	1,5191	1,6311	1,6244	1,5995	1,6195	—
	S_m^{ϕ} $(h = 0, 4)$	**0,4219**	0,3651	0,3392	0,3133	0,2990	0,3073	0,2890
	$DB_m^{\phi}(h = 1, 1)$	**1,0587**	1,1453	1,1242	1,1467	1,1584	1,1466	1,1356
	S	**0,2468**	0,2178	0,1893	0,1710	0,1711	0,1578	0,1277
	DB	**1,4866**	1,5657	1,5515	1,5335	1,4900	1,5299	1,5469
Coordenadas	DB^*	**1,2754**	1,4883	1,5712	1,5193	1,4473	1,7002	1,7497
Tangentes	DB^{**}	**1,1263**	1,2783	1,4911	1,5054	1,4310	1,5016	—
	S^{ϕ} $(h = 0, 4)$	**0,4120**	0,3451	0,3241	0,2977	0,2565	0,2645	0,2680
	$DB^{\phi}(h = 1, 1)$	**1,1183**	1,5164	1,4134	1,5583	1,5979	1,5140	1,5520

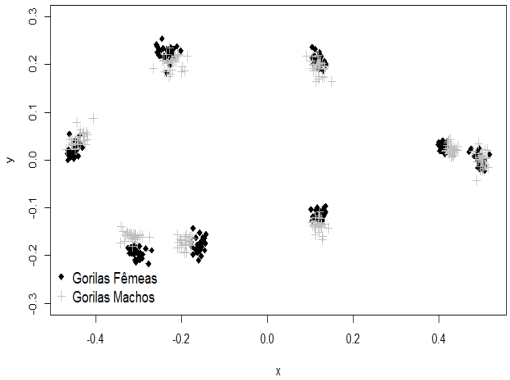
Figura 4.2: Conjunto de dados gorilas machos e fêmeas.



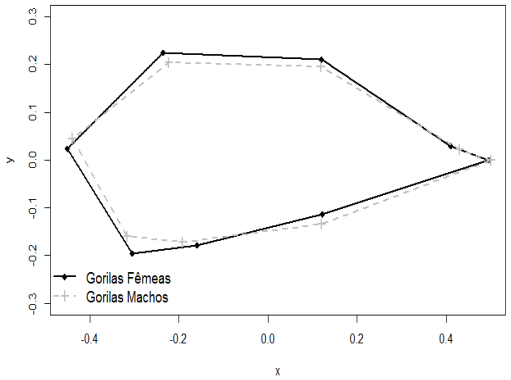
(a) Marcos anatômicos dos gorilas



(b) Pré-formas dos gorilas



(c) Ajuste Procrustes completo



(d) Forma média para cada grupo dos gorilas

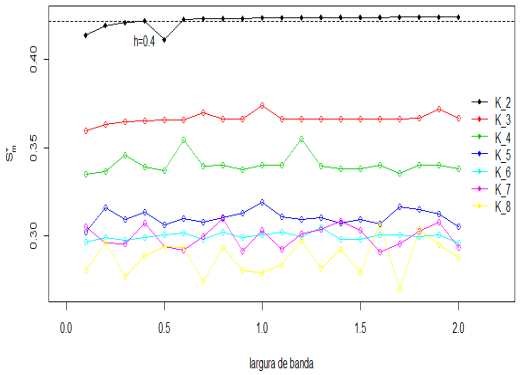
Fonte: Elaborada pelo autor.

4.2.2 Vértebra de ratos

Em um experimento para avaliar os efeitos da seleção para peso corporal na forma das vértebras dos ratos, três grupos de ratos foram obtidos: Controle, Grande e Pequeno. O grupo Controle contém ratos não selecionados, o grupo Grande contém ratos selecionados de cada geração de acordo ao grande peso corporal e o grupo Pequeno são selecionados pelo pequeno peso corporal. Dryden e Mardia (1998) consideraram a segunda vértebra torácica $T2$, em que cada vértebra dos grupos de ratos foram colocadas sob um microscópico e digitalizado usando uma câmera de vídeo para dar um nível de cinza na imagem, veja Figura 4.4. Nesse conjunto de dados de vértebras existem 30 Controles, 23 Grandes e 23 Pequenos. O objetivo é avaliar se existe uma diferença entre tamanho e forma entre esses três grupos e providenciar descrições de algumas diferenças.

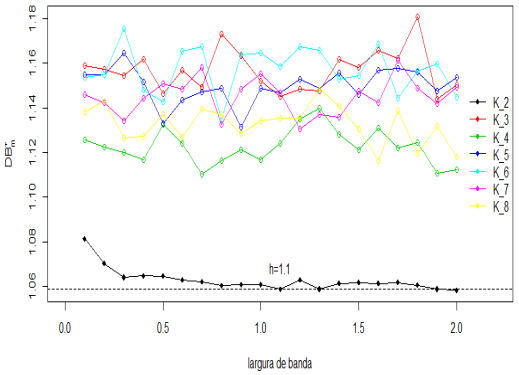
Os dados possuem 6 marcos os quais são obtidos usando um método semi automático em pontos de alta curvatura, esse procedimento é descrito em Mardia (1989) e Dryden (1989, Cap. 5). A Figura 4.6(a) representa as coordenadas Procrustes completa e a Figura 4.6(b) representa

Figura 4.3: Índice interno para validação do agrupamento nos dados dos gorilas



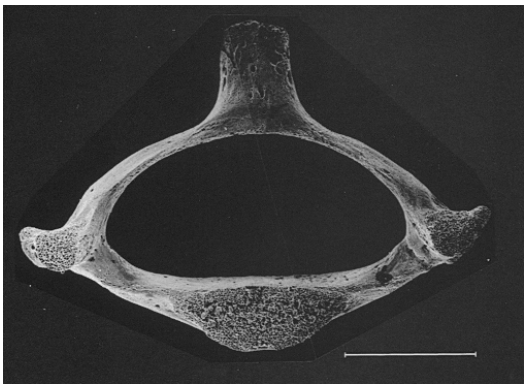
(a) Índice Silhoueta Modificado Kernelizado para os dados dos gorilas.

Fonte: Elaborada pelo autor.



(b) Índice Davies-Bouldin Modificado Kernelizado para os dados dos gorilas.

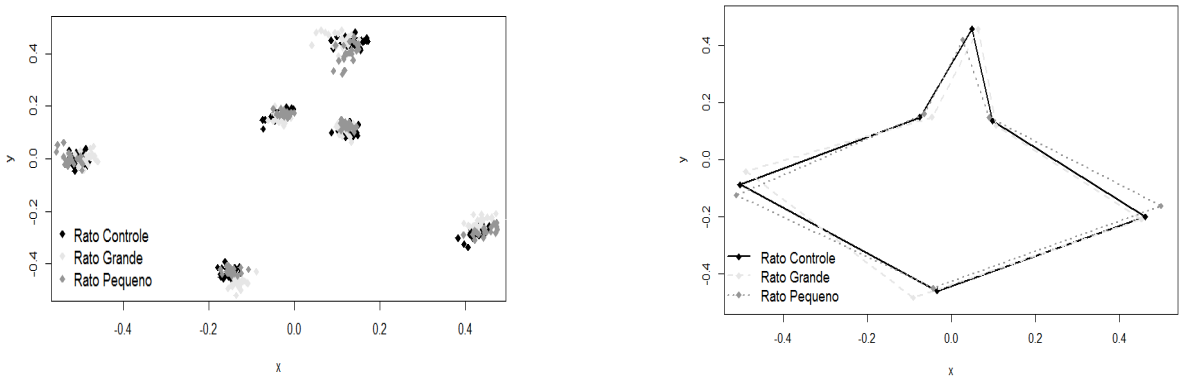
Figura 4.4: Digitalização da segunda vértebra torácica $T2$ de um rato.



Fonte: (DRYDEN e MARDIA, 1998).

a forma média Procrustes para cada grupo dos dados das vértebras $T2$ dos ratos.

Figura 4.5: Conjunto de dados vértebras de ratos $T2$ Controle, Grande e Pequeno.



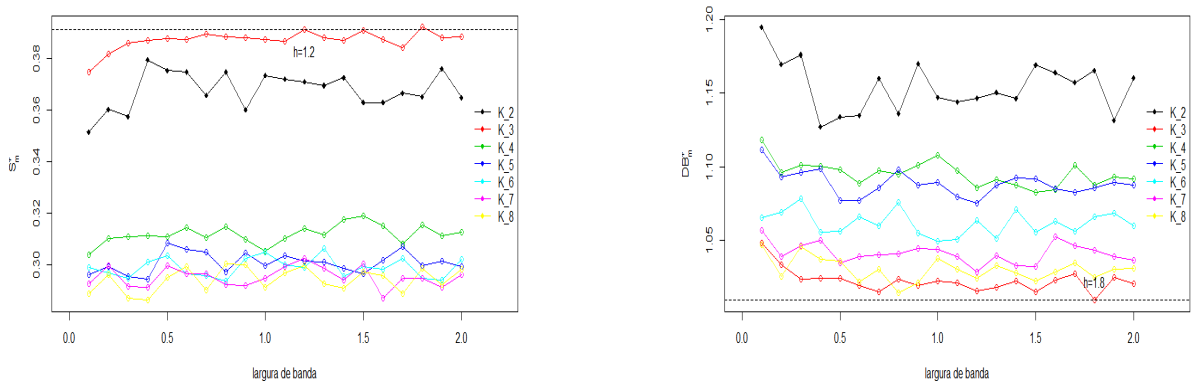
(a) Ajuste Procrustes completo

(b) Forma média para cada grupo das ratos

Fonte: Elaborada pelo autor.

Os dados das vértebras $(T2)$ dos ratos possuem alta concentração, dessa forma é de se esperar que os índices para coordenadas tangentes e os índices modificados para pré-formas funcionem bem. E, realmente, através da Tabela 4.2 podemos concluir que os índices em estudo conseguem estimar o verdadeiro valor do número de grupos no conjunto de dados. As Figuras 4.7(a) e 4.7(b) auxilia a encontrar o melhor valor para a largura de banda para os índices Silhoueta $(h = 1, 2)$ e Davies-Bouldin Kernelizados $(h = 1, 8)$, os quais remetem os valores ótimos dos índices, respectivamente.

Figura 4.6: Índice interno para validação do agrupamento nos dados das vértebras $(T2)$ dos ratos.



(a) Índice Silhoueta Modificado Kernelizado para os dados das vértebras $(T2)$ dos ratos.

(b) Índice Davies-Bouldin Modificado Kernelizado para os dados das vértebras $(T2)$ dos ratos.

Fonte: Elaborada pelo autor.

Tabela 4.2: Validação do agrupamento do conjunto de dados da vértebra torácica $(T2)$ dos ratos.

		Número de grupos						
	Índices	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	PR	0,0566	**0,0608**	0,0487	0,0412	0,0280	0,0204	0,0156
	S_m	0,2142	**0,2323**	0,1838	0,1688	0,1636	0,1556	0,1471
	DB_m	1,7270	**1,4507**	1,6164	1,6366	1,5665	1,5350	1,5322
Pré-Forma	DB_m^*	1,7866	**1,5048**	1,7465	1,7916	1,7491	1,7321	1,7526
	DB_m^{**}	1,5791	**1,4649**	1,7030	1,7476	1,7157	1,7066	—
	$S_m^\phi \quad (h=1,2)$	0,3711	**0,3912**	0,3140	0,3013	0,2988	0,3022	0,2996
	$DB_m^\phi (h=1,8)$	1,1653	**1,0092**	1,0873	1,0858	1,0658	1,0434	1,0148
	S	0,2147	**0,2329**	0,1833	0,1696	0,1642	0,1601	0,1471
	DB	1,7291	**1,4515**	1,6206	1,6397	1,5784	1,5405	1,5384
Coordenadas	DB^*	2,1626	**1,3949**	1,9303	2,0183	1,9344	1,8663	1,9095
Tangentes	DB^{**}	1,5364	**1,3127**	1,8308	1,9064	1,8302	1,7938	—
	$S^\phi \quad (h=1,2)$	0,3617	**0,3843**	0,2981	0,2714	0,2638	0,2518	0,2362
	$DB^\phi (h=1,8)$	1,4189	**1,0831**	1,3260	1,3896	1,3028	1,2735	1,2440

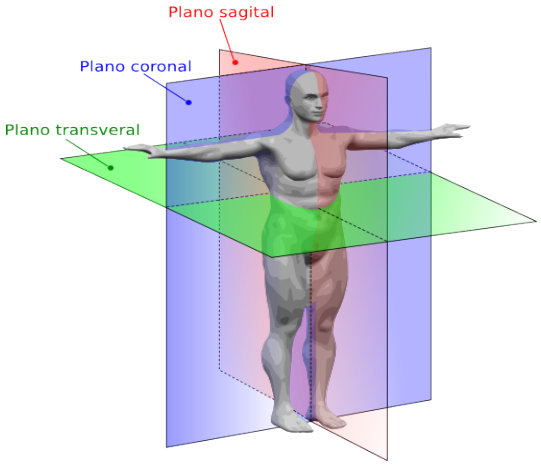
4.2.3 Medicina: digitalização da ressonância magnética em pacientes esquizofrênicos

Bookstein (1996) considera 13 marcos tomados próximo ao corte no plano sagital mediano (as descrições anatômicas, tanto do corpo humano quantos dos órgãos, são baseadas em 3 principais planos de secção que passam através do corpo na posição anatômica, sagital, transversal e coronal, observe a Figura 4.8(a)) bidimensional do escaneamento do cérebro de Ressonância Magnética (RM) de 14 pacientes esquizofrênicos e 14 pacientes normais. É de interesse estudar alguma diferença de forma nos cérebros entre os dois grupos, se em forma média ou em variabilidade da forma. Se diferenças morfométricas entre os dois grupos podem ser demonstrados, então isso deveria permitir pesquisadores a ganhar um aumento no entendimento sobre a condição. Na Figura 4.8(b) podemos ver a disposição dos marcos o qual possuem um contexto neuropsiquiátrico para serem escolhidos, veja (DEQUARDO et al., 1996).

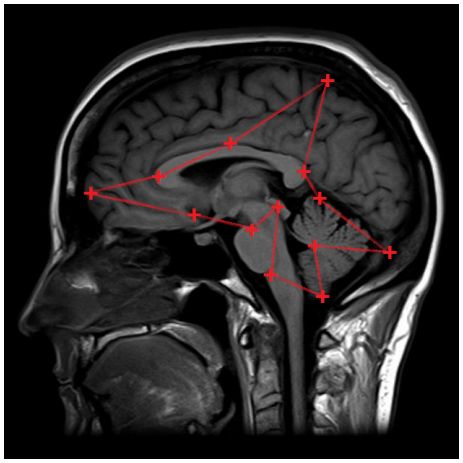
A Figura 4.9(a) representa as coordenadas Procrustes completa e a Figura 4.9(b) representa a forma média Procrustes para cada grupo dos pacientes controle e dos pacientes com esquizofrenia.

Os dados dos pacientes normais e com esquizofrenia possuem alta concentração, sendo assim, é de se esperar que os índices para coordenadas tangentes e os índices modificados para pré-formas funcionem bem. Através da Tabela 4.3 podemos concluir que os índices Silhoueta e sua versão kernelizada para coordenadas tangentes e o índice proposto Procrustes Residual para pré-formas identificam o verdadeiro número de grupos. Porém os índices Davies-Bouldin e suas versões kernelizadas e aumentada não conseguem estimar o verdadeiro parâmetro de interesse. Uma suposição para esse feito se dá pelo motivo da amostra ser pequena, apenas 28 indivíduos, ou seja, a partir do momento que aumentamos o número de grupos, os grupos vão se diluindo e o numerador dos índices vai se aproximando do denominador. Isto é, a distância entre o indivíduo e a forma média (ou centróide); e a distância entre as formas médias (ou centróides) serão próximas, uma vez que cada grupo terá poucos indivíduos. As Figuras 4.10(a) e 4.10(b) auxilia a encontrar o melhor valor para a largura de banda para os índices Silhoueta $(h = 1, 6)$

Figura 4.7: Disposição dos 3 tipos de planos na posição anatômica e os 13 marcos dispostos próximo ao plano sagital mediano do cérebro humano, segundo DeQuardo et al. (1996).



(a) Posições anatômicas: sagital, transversal e coronal

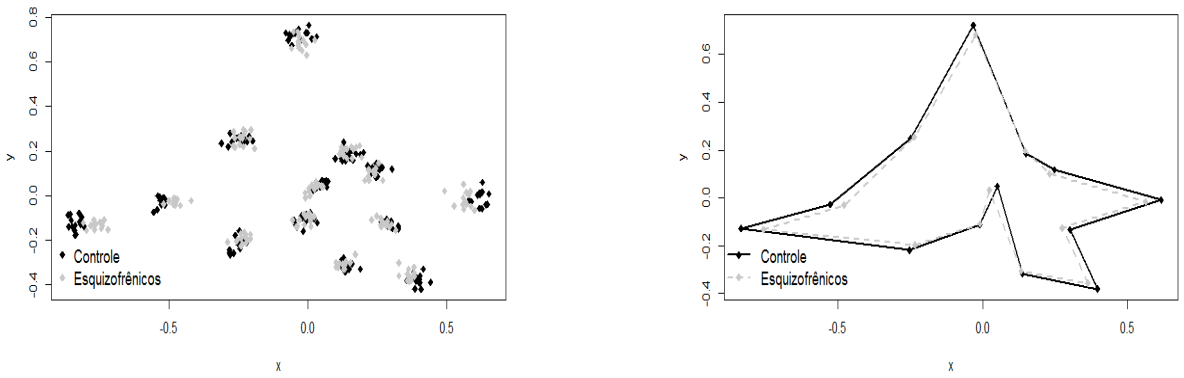


(b) Plano sagital mediano do cérebro humano

Fonte: Primeira Figura retirada da internet; segunda Figura elaborada pelo autor.

e Davies-Bouldin Kernelizados $(h = 0,6)$, os quais remetem os valores ótimos desses índices respectivamente.

Figura 4.8: Conjunto de dados vértebras de ratos $T2$ Controle, Grande e Pequeno.



(a) Ajuste Procrustes completo

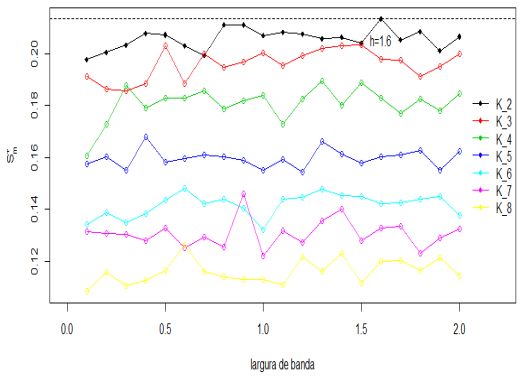
(b) Forma média para cada grupo das ratos

Fonte: Elaborada pelo autor.

Tabela 4.3: Validação do agrupamento do conjunto de dados dos pacientes normais e pacientes com esquizofrenia.

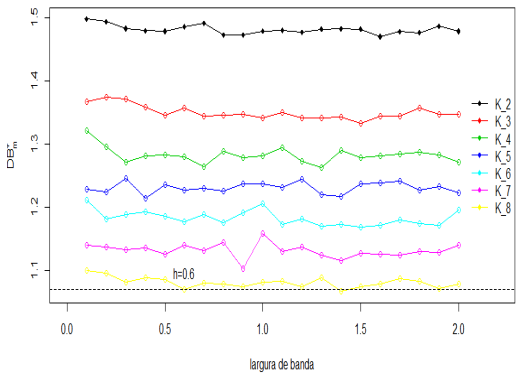
		Número de grupos						
	Índices	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	PR	**0,1448**	0,1030	0,0617	0,0563	0,0284	-0,0213	-0,0569
	S_m	**0,1214**	0,0971	0,0755	0,0690	0,0576	0,0422	0,0167
	DB_m	2,3098	2,0877	1,9240	1,8176	1,7250	1,6139	**1,5561**
Pré-Forma	DB_m^*	2,3971	2,1500	2,0510	1,9351	1,8882	1,7975	**1,7688**
	DB_m^{**}	2,2614	2,0728	1,9595	1,8964	1,8484	**1,7339**	—
	$S_m^\phi \quad (h=1,6)$	**0,2133**	0,1979	0,1828	0,1604	0,1421	0,1327	0,1200
	$DB_m^\phi (h=0,6)$	1,4856	1,3568	1,2790	1,2272	1,1765	1,1397	**1,0700**
	S	**0,1217**	0,0973	0,0786	0,0705	0,0607	0,0592	0,0918
	DB	2,3150	2,0926	1,9270	1,8344	1,7427	1,6487	**1,4551**
Coordenadas	DB^*	3,1481	2,4850	2,3690	2,2007	2,0701	1,8220	**1,3623**
Tangentes	DB^{**}	2,7828	2,3655	2,1433	2,0391	1,9136	**1,6752**	—
	$S^\phi \quad (h=1,6)$	**0,1973**	0,1671	0,1251	0,0756	0,0657	0,0577	0,0254
	$DB^\phi (h=0,6)$	2,9498	2,2957	1,9335	1,8030	1,5827	1,4305	**1,2845**

Figura 4.9: Índice interno para validação do agrupamento nos dados de pacientes esquizofrênicos.



(a) Índice Silhoueta Modificado Kernelizado para os dados de pacientes esquizofrênicos.

Fonte: Elaborada pelo autor.



(b) Índice Davies-Bouldin Modificado Kernelizado para os dados de pacientes esquizofrênicos.

CAPÍTULO 5

Considerações finais

Após o agrupamento do conjunto de dados, tem-se a finalidade de validar os grupos e identificar o número correto de grupos. Para tanto, introduziu-se ferramentas para formas planas para estimar o número ideal de grupos em um determinado conjunto de dados de acordo com a natureza e característica desses dados, os quais não pertence ao espaço Euclidiano. As estruturas planas podem ser representadas pelas coordenadas tangentes as quais podem-se utilizar os conhecimentos padrões referentes as medidas de validação de Análise Multivariada. Ou também, podem ser representadas pelas coordenadas de Kendall, estas, como já definido anteriormente não apresentam literatura. Desta forma, esta dissertação tem como proposta abordar medidas de validação interna modificados para lidar com tal natureza não Euclidiana.

Simulações são consideradas, para avaliar a qualidade dos índices existentes na literatura para as coordenadas tangentes e comparar com os índices modificados para coordenadas de Kendall. Diferentes cenários são supostos, baseados na alta concentração e baixa concentração dos dados. Diante os resultados, quando simulados as amostras no espaço dos marcos, para $\sigma = 0,1$, (remetem a alta concentração) tanto os índices para coordenadas tangentes como para coordenadas de Kendall mostram resultados pertinentes quanto ao número correto de grupos. Mas, quando $\sigma = 0,5$, baixa concentração, os índices Davies-Bouldin para coordenadas tangentes não estimam o verdadeiro número de grupos; já os índices modificados propostos são mais eficientes.

Simulações da distribuição Bingham que possui suporte na esfera unitária e é bastante utilizada em análise de formas planas também foi estudada. Nos casos em que foi considerado alta concentração, ou seja, $\boldsymbol{\lambda} = (200; 200; 0)$ e $\boldsymbol{\lambda} = (100; 100; 0)$ todos os índices para coordenadas tangentes e os índices modificados apresentaram bons resultados e estimaram o verdadeiro número de grupos. Já no caso de baixa concentração, isto é, $\boldsymbol{\lambda} = (2; 1,5; 0)$ e $\boldsymbol{\lambda} = (1; 1; 0)$ nenhum índice em estudo para coordenadas tangentes mostrou-se eficiente na identificação do número correto de grupos, e os índices modificados Davies-Bouldin ajustado pela segunda vez e

o índice Procruste Residual conseguem estimar o verdadeiro número de grupos para as amostras de duas populações da distribuição Bingham.

Os dados reais apresentados nas análises apresentam alta concentração, então é de esperar que os índices modificados funcione tão bem quanto os índices para coordenadas tangentes. Nos dados de esquizofrenia o índice Davies-Bouldin não estimar o verdadeiro número de grupos, uma suposição se dá pelo motivo da amostra ser pequena. Através dessas análises recomenda-se o uso dos índices propostos na Dissertação para identificar o verdadeiro número de grupos de um determinado conjunto de dados. Através da Tabela 5.1 para dados simulados e Tabela 5.2 para dados reais, pode-se notar que o índice Procrustes Residual obteve 100% de êxito na identificação do número correto de grupos. É interessante perceber que este último faz uso de técnicas próprias de pré-formas. Em segundo lugar, o índice Davies-Bouldin aumentado modificado (DB_m^{**}) para pré-formas apresentou-se melhor resultado.

Tabela 5.1: Número estimado de grupos para sete dados simulados com tamanho de amostra variando entre 30 e 50, sendo três no espaço dos *landmarks*, dois da distribuição Bingham com alta *(high)* concentração, e dois da distribuição Bingham com baixa *(low)* concentração; e três conjunto de dados reais e sete índices para pré-forma e seis para coordenadas tangentes.

	Pré-Forma							**Coordenadas Tangentes**					
	PR	S_m	DB_m	DB_m^*	DB_m^{**}	S_m^ϕ	DB_m^ϕ	S	DB	DB^*	DB^{**}	S^ϕ	DB^ϕ
$Land_{0,1}(n=30)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
$Land_{0,1}(n=50)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
$Land_{0,3}(n=30)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	8(×)	2(○)	2(○)	2(○)	8(×)
$Land_{0,3}(n=50)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
$Land_{0,5}(n=30)$	2(○)	2(○)	8(×)	8(×)	2(○)	2(○)	8(×)	2(○)	8(×)	8(×)	6(×)	2(○)	8(×)
$Land_{0,5}(n=50)$	2(○)	2(○)	8(×)	2(○)	2(○)	2(○)	2(○)	2(○)	8(×)	8(×)	7(×)	2(○)	8(×)
$Bing_{high1}(n=30)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
$Bing_{high1}(n=50)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
$Bing_{high2}(n=30)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
$Bing_{high2}(n=50)$	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
$Bing_{low3}(n=30)$	2(○)	8(×)	8(×)	2(○)	2(○)	5(×)	8(×)	6(×)	8(×)	8(×)	6(×)	4(×)	8(×)
$Bing_{low3}(n=50)$	2(○)	8(×)	8(×)	3(×)	2(○)	7(×)	8(×)	6(×)	8(×)	8(×)	6(×)	7(×)	8(×)
$Bing_{low4}(n=30)$	2(○)	7(×)	3(×)	3(×)	2(×)	5(×)	8(×)	5(×)	8(×)	8(×)	7(×)	5(×)	8(×)
$Bing_{low4}(n=50)$	2(○)	7(×)	3(×)	3(×)	2(○)	7(×)	8(×)	7(×)	8(×)	8(×)	7(×)	6(×)	8(×)
Total	14	10	8	10	13	10	9	10	7	8	8	10	7

Notação indica quando o número correto de grupos tem sido encontrado (○) ou não (×).

Tabela 5.2: Número estimado de grupos para três conjunto de dados reais e sete índices para pré-forma e seis para coordenadas tangentes.

	Pré-Forma							**Coordenadas Tangentes**					
	PR	S_m	DB_m	DB_m^*	DB_m^{**}	S_m^ϕ	DB_m^ϕ	S	DB	DB^*	DB^{**}	S^ϕ	DB^ϕ
Gorilas	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)	2(○)
Vértebra $T2$ Ratos	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)	3(○)
Schiz	2(○)	2(○)	8(×)	8(×)	7(×)	2(○)	8(×)	2(○)	8(×)	8(×)	7(×)	2(○)	8(×)
Total	3	3	2	2	2	3	2	3	2	2	2	3	2

Notação indica quando o número correto de grupos tem sido encontrado (○) ou não (×).

Sugestões para trabalhos futuros: desenvolver outros índices modificados para lidar com formas planas, tais como índice de Dunn, C_index ; Calinski Harabasz; SD Scat; S Dbw; Tau; Trace W; Xie Beni, entre outros e desenvolver versões kernelizadas. Em análises de formas, não há um estudo detalhado sobre métodos de agrupamento particionado difuso, e também sua versão kernelizada; e após desenvolver essas técnicas pode-se avançar os estudos sobre índices internos próprios para os algoritmos difusos. Além disso, pode-se desenvolver algoritmos particionais os quais são robustos a ruídos e outlier, e também que convergem para um ótimo global, já que os algoritmos K-médias e *Kernel* K-médias convergem para o ótimo local. Encontrar diferentes funções de *Kernel* para formas planas, nessa dissertação foi utilizada apenas o *Kernel* gaussiano.

Referências

AMARAL, G. J. et al. 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. et al. Bootstrap confidence regions for the planar mean shape. *Journal of Statistical Planning and Inference*, Elsevier, v. 140, n. 11, p. 3026–3034, 2010.

BOOKSTEIN, F. L. A statistical method for biological shape comparisons. *Journal of Theoretical Biology*, Elsevier, v. 107, n. 3, p. 475–520, 1984.

BOOKSTEIN, F. L. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, JSTOR, p. 181–222, 1986.

BOOKSTEIN, F. L. Biometrics, biomathematics and the morphometric synthesis. *Bulletin of mathematical biology*, Springer, v. 58, n. 2, p. 313–365, 1996.

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

CAMPS-VALLS, G. *Kernel methods in bioengineering, signal and image processing*. [S.l.]: Igi Global, 2006.

CLAUDE, J. *Morphometrics with R*. [S.l.]: Springer Science & Business Media, 2008.

DEQUARDO, J. R. et al. Spatial relationships of neuroanatomic landmarks in schizophrenia. *Psychiatry Research: Neuroimaging*, Elsevier, v. 67, n. 1, p. 81–95, 1996.

DRYDEN, I. L. *The Statistical Analysis of Shape analysis*. Tese (Doutorado) — University of Leeds, 1989.

DRYDEN, I. L.; MARDIA, K. V. *Statistical shape analysis*. [S.l.]: J. Wiley Chichester, 1998.

GIROLAMI, M. Mercer kernel-based clustering in feature space. *Neural Networks, IEEE Transactions on*, IEEE, v. 13, n. 3, p. 780–784, 2002.

GOODALL, C. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 285–339, 1991.

GOODALL, C.; MARDIA, K. V. The noncentral bartlett decompositions and shape densities. *Journal of Multivariate Analysis*, Elsevier, v. 40, n. 1, p. 94–108, 1992.

GOODALL, C. R.; MARDIA, K. V. Multivariate aspects of shape theory. *The Annals of Statistics*, JSTOR, p. 848–866, 1993.

GREENBERG, H. J. *A Simplified Introduction to LATEX*. Denver, Colorado, 2000. Disponível em: <http://www.tex.ac.uk/ctan/info/simplified-latex/simplified-intro.pdf>.

HAMMER, Ã. yvind; HARPER, D. A. *Paleontological data analysis*. [S.l.]: John Wiley & Sons, 2008.

JAIN, A. K. Data clustering: 50 years bey k-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651–666, 2010.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. [S.l.]: Prentice-Hall, Inc., 1988.

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

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

KENDALL, D. G. The diffusion of shape. *Advances in applied probability*, JSTOR, 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*, Oxford University Press, v. 16, n. 2, p. 81–121, 1984.

KENDALL, D. G. et al. *Shape and shape theory*. [S.l.]: John Wiley & Sons, 2009.

KENT, J. T. The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 285–299, 1994.

KIM, M.; RAMAKRISHNA, R. New indices for cluster validity assessment. *Pattern Recognition Letters*, Elsevier, v. 26, n. 15, p. 2353–2363, 2005.

LAMPORT, L. A document preparation system. In: . [S.l.: s.n.], 1994.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MARDIA, K. Discussion of"a survey of the statistical theory of shape,"by dg kendall. *Statistical Science*, v. 4, p. 108–111, 1989.

MERCER, J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, JSTOR, v. 209, p. 415–446, 1909.

O'HIGGINS, P. *A morphometric study of cranial shape in the Hominoidea*. Tese (Doutorado) — University of Leeds, 1989.

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

R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2009. Disponível em: <http://www.R-project.org>.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.

SMALL, C. G. *The statistical theory of shape*. [S.l.]: Springer Science & Business Media, 2012.

APÊNDICE A

Tabelas para dados simulados no espaço dos marcos

Tabela A.1: Validação do agrupamento do conjunto de dados simulados para $N = 30$ e $N = 50$ para $\sigma = 0, 1$.

		Número de grupos							
N=30	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,3470**	0,1564	0,1086	0,0864	0,0769	0,0712	0,0624	0,98
		(0,0495)	(0,0398)	(0,0393)	(0,0355)	(0,0322)	(0,0331)	(0,0324)	
	S_m	**0,3993**	0,1635	0,1023	0,0840	0,0749	0,0729	0,0732	0,99
		(0,0402)	(0,0732)	(0,0352)	(0,0205)	(0,0205)	(0,0196)	(0,0210)	
	DB_m	**1,0903**	1,8912	1,9919	1,9465	1,8748	1,7822	1,7276	0,99
		(0,2462)	(0,2768)	(0,2582)	(0,1921)	(0,1802)	(0,1266)	(0,1112)	
Pré-Forma	DB_m^*	**1,3105**	2,8799	3,2266	3,2189	3,2512	3,1567	3,1369	0,99
		(0,2322)	(0,5502)	(0,3960)	(0,3368)	(0,3294)	(0,2582)	(0,2594)	
	DB_m^{**}	**1,1091**	2,8019	3,2015	3,2375	3,3431	3,3487	—-	0,99
		(0,2473)	(0,5738)	(0,4272)	(0,3604)	(0,3951)	(0,3517)	—-	
	$S_m^{\phi \cdot (h=1,75)}$	**0,5322**	0,2397	0,1493	0,1229	0,0992	0,0927	0,0953	0,99
		(0,0551)	(0,0885)	(0,0487)	(0,0395)	(0,0470)	(0,0452)	(0,0391)	
	$DB_m^{\phi \cdot (h=1,75)}$	**0,9024**	1,3047	1,3535	1,3300	1,3033	1,2769	1,2355	0,99
		(0,1093)	(0,1024)	(0,1030)	(0,0969)	(0,0968)	(0,0916)	(0,0753)	
	S	**0,4099**	0,1659	0,1031	0,0845	0,0753	0,0732	0,0741	0,99
		(0,0417)	(0,0760)	(0,0367)	(0,0215)	(0,0216)	(0,0209)	(0,0221)	
	DB	**1,1402**	1,9273	2,0418	2,0056	1,9262	1,8318	1,7744	0,99
		(0,2698)	(0,2720)	(0,2784)	(0,2117)	(0,1966)	(0,1381)	(0,1187)	
Coordenadas	DB^*	**1,1967**	6,5184	8,1489	8,1763	8,3796	7,9602	7,9267	0,99
		(0,7470)	(2,6567)	(2,2378)	(2,0089)	(1,9924)	(1,5461)	(1,5169)	
Tangentes	DB^{**}	**0,9576**	6,4914	8,2232	8,4339	8,9390	9,0629	—-	0,99
		(0,7200)	(2,8230)	(2,3431)	(2,1182)	(2,2822)	(1,9663)	——	
	$S^{\phi \cdot (h=1,75)}$	**0,5449**	0,2426	0,1481	0,1201	0,0959	0.0909	0.0931	0,99
		(0,0574)	(0,0923)	(0,0529)	(0,0424)	(0,0505)	(0,0477)	(0,0412)	
	$DB^{\phi \cdot (h=1,75)}$	0,8436	2,2136	2,2502	2,0424	1,9281	1,8362	1,6719	0,99
		(0,8326)	(0,6870)	(0,5200)	(0,4092)	(0,4018)	(0,3491)	(0,2436)	

		Número de grupos							
N=50	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,3272**	0,1546	0,1195	0,0979	0,0799	0,0745	0,0701	0,99
		(0,0412)	(0,0287)	(0,0312)	(0,0282)	(0,0232)	(0,0235)	(0,0219)	
	S_m	**0,3996**	0,1653	0,1047	0,0841	0,0761	0,0697	0,0694	0,99
		(0,0378)	(0,0679)	(0,0276)	(0,0182)	(0,0155)	(0,0138)	(0,0142)	
	DB_m	**1,1011**	1,9940	2,1340	2,1146	2,0891	2,0144	1,9511	0,99
		(0,2540)	(0,2039)	(0,1977)	(0,1881)	(0,1479)	(0,1187)	(0,1071)	
Pré-Forma	DB_m^*	**1,3272**	3,0511	3,4121	3,4631	3,4177	3,4164	3,3516	0,99
		(0,2352)	(0,5458)	(0,3945)	(0,3147)	(0,2931)	(0,2839)	(0,2569)	
	DB_m^{**}	**1,1240**	2,9595	3,4078	3,5137	3,4967	3,582	—-	0,99
		(0,2446)	(0,5568)	(0,4482)	(0,3835)	(0,3820)	(0,3698)	—-	
	$S_m^{\phi \cdot (h=2,5)}$	**0,5407**	0,2567	0,1604	0,1262	0,1066	0,1021	0,0949	1,00
		(0,0199)	(0,0850)	(0,0478)	(0,0291)	(0,0293)	(0,0308)	(0,0323)	
	$DB_m^{\phi \cdot (h=2,5)}$	**0,8943**	1,3447	1,4201	1,4282	1,4162	1,3773	1,3645	1,00
		(0,0458)	(0,0594)	(0,1022)	(0,0802)	(0,0683)	(0,0599)	(0,0551)	
	S	**0,4100**	0,1672	0,1054	0,0845	0,0769	0,0702	0,0703	0,99
		(0,0389)	(0,0701)	(0,0295)	(0,0196)	(0,0166)	(0,0147)	(0,0149)	
	DB	**1,1526**	2,0381	2,1913	2,1797	2,1567	2,0754	2,0082	0,99
		(0,2753)	(0,2006)	(0,2239)	(0,2122)	(0,1603)	(0,1266)	(0,1133)	
Coordenadas	DB^*	**1,2305**	7,3853	9,1339	9,3752	9,0878	9,1259	8,8449	0,99
		(0,7523)	(2,9121)	(2,4755)	(1,9314)	(1,7723)	(1,7025)	(1,6345)	
Tangentes	DB^{**}	**0,9726**	7,2549	9,2516	9,7533	9,6927	10,1982	—-	0,99
		(0,6589)	(2,9472)	(2,6918)	(2,1997)	(2,2472)	(2,0657)	——	
	$S^{\phi \cdot (h=2,5)}$	**0,5508**	0,2584	0,1595	0,1247	0,1046	0,1006	0,0929	1,00
		(0,0230)	(0,0873)	(0,0518)	(0,0321)	(0,0316)	(0,0341)	(0,0351)	
	$DB^{\phi \cdot (h=2,5)}$	**0,7686**	2,3787	2,6186	2,5115	2,4192	2,2169	2,1655	1,00
		(0,0814)	(0,5419)	(0,5455)	(0,3927)	(0,3566)	(0,2953)	(0,2858)	

Tabela A.2: Validação do agrupamento do conjunto de dados simulados para $N = 30$ e $N = 50$ para $\sigma = 0,3$.

		Número de grupos							
N=30	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,2215**	0,1260	0,0963	0,0847	0,0785	0,0651	0,0561	0,77
		(0,0786)	(0,0347)	(0,0306)	(0,0359)	(0,0328)	(0,0314)	(0,0350)	
	S_m	**0,2205**	0,0906	0,0637	0,0576	0,0533	0,0494	0,0434	0,81
		(0,0959)	(0,0463)	(0,0182)	(0,0193)	(0,0208)	(0,0204)	(0,0212)	
	DB_m	**1.5299**	2,2067	2,1185	1,9933	1,9131	1,8218	1,7849	0,71
		(0,6002)	(0,2085)	(0,1626)	(0,1376)	(0,1220)	(1,1096)	(1,1067)	
Pré-Forma	DB_m^*	**1,9768**	2,7749	2,7488	2,6577	2,5994	2,5142	2,5121	0,73
		(0,5335)	(0,3304)	(0,2083)	(0,2097)	(0,1957)	(0,1981)	(0,2038)	
	DB_m^{**}	**1,8169**	2,6873	2,7012	2,6296	2,6054	2,5703	—	0,76
		(0,5872)	(0,3202)	(0,2149)	(0,2219)	(0,2504)	(0,2538)		
	$S_m^\phi \cdot_{(h=2,5)}$	**0,3311**	0,1372	0,1004	0,0721	0,0540	0,0478	0,0387	0,80
		(0,1233)	(0,0710)	(0,0339)	(0,0378)	(0,0421)	(0,0439)	(0,0426)	
	$DB_m^\phi \cdot_{(h=1,0)}$	**1,2908**	1,5150	1,5196	1,5055	1,4594	1,4295	1,3910	0,73
		(0,1044)	(0,0506)	(0,0443)	(0,0392)	(0,0350)	(0,0346)	(0,0502)	
	S	**0,2310**	0,0890	0,0618	0,0542	0,0498	0,0469	0,0398	0,83
		(0,0930)	(0,0482)	(0,0196)	(0,0224)	(0,0241)	(0,0225)	(0,0232)	
	DB	2,1151	2,3326	2,2732	2,1175	2,0183	1,9260	**1,8843**	0,46
		(0,6186)	(0,2083)	(0,1825)	(0,1438)	(0,1390)	(0,1116)	(0,1135)	
Coordenadas	DB^*	**3,3825**	5,6794	5,4571	5,0728	4,9131	4,5944	4,6515	0,73
		(1,7748)	(1,3209)	(1,0050)	(0,9679)	(0,9075)	(0,9072)	(0,9168)	
Tangentes	DB^{**}	**3,0260**	5,4679	5,3384	4,9966	4,9857	4,8941	—	0,75
		(1,7877)	(1,3049)	(1,0266)	(0,9803)	(1,0755)	(1,2210)	—	
	$S^\phi \cdot_{(h=1,00)}$	**0,3137**	0,1116	0,0735	0,0651	0,0491	0,0284	0,0190	0,88
		(0,1216)	(0,0583)	(0,0426)	(0,0439)	(0,0363)	(0,0473)	(0,0455)	
	$DB^\phi \cdot_{(h=1,00)}$	2,7666	3,0904	2,9089	2,6179	2,3505	2,1850	**2,0507**	0,54
		(1,8201)	(0,5478)	(0,5401)	(0,3880)	(0,3488)	(0,3089)	(0,2627)	

		Número de grupos							
N=50	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,2368**	0,1220	0,0900	0,0809	0,0776	0,0660	0,0595	0,96
		(0,0429)	(0,0326)	(0,0268)	(0,0261)	(0,0257)	(0,0252)	(0,0274)	
	S_m	**0,2580**	0,0951	0,0721	0,0643	0,0573	0,0534	0,0513	0,93
		(0,0570)	(0,0551)	(0,0252)	(0,0190)	(0,0147)	(0,0133)	(0,0137)	
	DB_m	**1,6216**	2,3117	2,2628	2,1440	2,0589	1,9790	1,9243	0,93
		(0,3704)	(0,1996)	(0,1758)	(0,1279)	(0,1039)	(0,0963)	(0,0954)	
Pré-Forma	DB_m^*	**1,8130**	2,8926	2,8628	2,7574	2,7060	2,6313	2,6117	0,93
		(0,3314)	(0,3674)	(0,2558)	(0,2022)	(0,1709)	(0,1650)	(0,1409)	
	DB_m^{**}	**1,6302**	2,8106	2,8307	2,7564	2,7276	2,7070	—	0,93
		(0,3628)	(0,3600)	(0,2619)	(0,2346)	(0,2005)	(0,2023)		
	$S_m^\phi \cdot_{(h=2,5)}$	**0,3634**	0,1404	0,1034	0,0980	0,0797	0,0735	0,0674	0,92
		(0,0866)	(0,0724)	(0,0348)	(0,0409)	(0,0250)	(0,0277)	(0,0250)	
	$DB_m^\phi \cdot_{(h=1,75)}$	**1,3276**	1,5460	1,5689	1,5435	1,5119	1,4842	1,4508	0,92
		(0,1376)	(0,0459)	(0,0532)	(0,0440)	(0,0397)	(0,0401)	(0,0374)	
	S	**0,2670**	0,0944	0,0710	0,0630	0,0549	0,0523	0,0491	0,95
		(0,0554)	(0,0568)	(0,0279)	(0,0210)	(0,0173)	(0,0149)	(0,0150)	
	DB	**1,9414**	2,4758	2,4113	2,2871	2,1883	2,0917	2,0331	0,79
		(0,3880)	(0,1745)	(0,1953)	(0,1360)	(0,1144)	(0,1049)	(0,0972)	
Coordenadas	DB^*	**2,9027**	6,2322	5,9942	5,5554	5,4100	5,0376	4,9845	0,91
		(1,1688)	(1,4017)	(1,259)	(0,9224)	(0,8799)	(0,8051)	(0,6600)	
Tangentes	DB^{**}	**2,5214**	6,0326	5,8877	5,5621	5,4848	5,4005	—	0,92
		(1,1565)	(1,4070)	(1,2563)	(1,0135)	(0,9652)	(1,0007)	—	
	$S^\phi \cdot_{(h=1,00)}$	**0,3488**	0,1163	0,0885	0,0649	0,0623	0,0560	0,0536	0,94
		(0,0844)	(0,0540)	(0,0432)	(0,0284)	(0,0328)	(0,0309)	(0,0307)	
	$DB^\phi \cdot_{(h=2,5)}$	**2,1989**	3,2830	3,1118	2,7859	2,5657	2,3603	2,2722	0,74
		(1,0625)	(0,4804)	(0,4925)	(0,3122)	(0,2754)	(0,2245)	(0,2264)	

Tabela A.3: Validação do agrupamento do conjunto de dados simulados para $N = 30$ e $N = 50$ para $\sigma = 0,5$.

		Número de grupos							
N=30	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,1623**	0,1048	0,0843	0,0603	0,0483	0,0398	0,0235	0,75
		(0,0702)	(0,0344)	(0,0353)	(0,0392)	(0,0351)	(0,0404)	(0,0322)	
	S_m	**0,1181**	0,0664	0,0549	0,0437	0,0414	0,0371	0,0304	0,67
		(0,0784)	(0,0285)	(0,0210)	(0,0169)	(0,0172)	(0,0210)	(0,0182)	
	DB_m	2,2492	2,2222	2,0871	2,0076	1,9090	1,8540	**1,7979**	0,26
		(0,5348)	(0,1815)	(0,1399)	(0,1421)	(0,1154)	(0,1053)	(0,1201)	
Pré-Forma	DB_m^*	2,3406	2,6001	2,4978	2,4467	2,3610	2,3305	**2,2891**	0,36
		(0,4822)	(0,2145)	(0,1937)	(0,1835)	(0,1421)	(0,1390)	(0,1434)	
	DB_m^{**}	**2,0828**	2,5235	2,4478	2,4077	2,3276	2,3338	—	0,38
		(0,3959)	(0,2091)	(0,1908)	(0,1838)	(0,1572)	(0,1864)	—	
	S_m^ϕ $\cdot(h=3,25)$	**0,1825**	0,0988	0,0746	0,0626	0,0501	0,0446	0,0306	0,70
		(0,1159)	(0,0372)	(0,0364)	(0,0365)	(0,0347)	(0,0383)	(0,0355)	
	$DB_m^\phi \cdot(h=0,5)$	1,6615	1,6665	1,6219	1,6089	1,5738	1,5416	**1,5175**	0,32
		(0,1758)	(0,0711)	(0,0543)	(0,0523)	(0,0370)	(0,0453)	(0,0409)	
	S	**0,1376**	0,0725	0,0597	0,0465	0,0440	0,0386	0,0330	0,78
		(0,0753)	(0,0300)	(0,0243)	(0,0195)	(0,0186)	(0,0229)	(0,0213)	
	DB	2,6519	2,4681	2,2830	2,1769	2,0473	1,9737	**1,8965**	0,11
		(0,5246)	(0,2076)	(0,1628)	(0,1527)	(0,1281)	(0,1158)	(0,1126)	
Coordenadas	DB^*	4,7245	5,0655	4,4896	4,2771	3,9174	3,8271	**3,6237**	0,25
		(1,5938)	(0,8543)	(0,7970)	(0,7449)	(0,5395)	(0,5468)	(0,5575)	
Tangentes	DB^{**}	4,4276	4,8138	4,3195	4,1431	**3,8141**	3,8367	—	0,33
		(1,6350)	(0,8327)	(0,7589)	(0,7208)	(0,6077)	(0,7139)		
	S^ϕ $\cdot(h=0,5)$	**0,1842**	0,0839	0,0666	0,0485	0,0318	0,0239	0,0168	0,85
		(0,1002)	(0,0286)	(0,0316)	(0,0344)	(0,0384)	(0,0403)	(0,0382)	
	$DB^\phi \cdot(h=0,5)$	4,8436	4,2035	3,3710	3,1730	2,8359	2,6112	**2,4433**	0,18
		(2,0917)	(0,8691)	(0,4927)	(0,5141)	(0,3235)	(0,3074)	(0,2835)	

		Número de grupos							
N=50	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,1741**	0,1077	0,0836	0,0679	0,0571	0,0428	0,0407	0,81
		(0,0597)	(0,0247)	(0,0257)	(0,0280)	(0,0278)	(0,0239)	(0,0233)	
	S_m	**0,1393**	0,0744	0,0625	0,0548	0,0524	0,0453	0,0465	0,59
		(0,0816)	(0,0259)	(0,0143)	(0,0135)	(0,0135)	(0,0118)	(0,0139)	
	DB_m	2,1712	2,2932	2,1562	2,0751	2,0120	1,9616	**1,8874**	0,43
		(0,5511)	(0,1672)	(0,1131)	(0,1085)	(0,1000)	(0,0928)	(0,0824)	
Pré-Forma	DB_m^*	**2,0908**	2,6934	2,5667	2,5130	2,4583	2,4137	2,3597	0,49
		(0,3735)	(0,2394)	(0,1449)	(0,1561)	(0,1328)	(0,1284)	(0,1111)	
	DB_m^{**}	**2,0327**	2,6215	2,5227	2,4787	2,4403	2,4254	—	0,51
		(0,4213)	(0,2314)	(0,1416)	(0,1542)	(0,1509)	(0,1464)		
	S_m^ϕ $\cdot(h=2,5)$	**0,2242**	0,1114	0,0986	0,0837	0,0738	0,0628	0,0571	0,68
		(0,1166)	(0,0332)	(0,0320)	(0,0258)	(0,0216)	(0,0223)	(0,0279)	
	$DB_m^\phi \cdot(h=0,5)$	**1,5263**	1,6565	1,6457	1,6398	1,6250	1,6051	1,5822	0,54
		(0,0931)	(0,0414)	(0,0477)	(0,0382)	(0,0360)	(0,0322)	(0,0304)	
	S	**0,1580**	0,0798	0,0674	0,0599	0,0567	0,0498	0,0504	0,72
		(0,0769)	(0,0263)	(0,0154)	(0,0167)	(0,0156)	(0,0145)	(0,0163)	
	DB	2,6089	2,5312	2,3492	2,2519	2,1693	2,0938	**2,0131**	0,12
		(0,5091)	(0,1570)	(0,1187)	(0,1205)	(0,1060)	(0,0970)	(0,0927)	
Coordenadas	DB^*	4,6244	5,4239	4,7726	4,5287	4,3108	4,1058	**3,9327**	0,38
		(1,5350)	(0,8830)	(0,6249)	(0,6580)	(0,5475)	(0,5074)	(0,4728)	
Tangentes	DB^{**}	4,2953	5,1937	4,6194	4,4101	4,2638	**4,1712**	—	0,45
		(1,5861)	(0,8619)	(0,5995)	(0,6178)	(0,5785)	(0,5577)	—	
	S^ϕ $\cdot(h=1,00)$	**0,2262**	0,0997	0,0819	0,0679	0,0615	0,0552	0,0488	0,91
		(0,0984)	(0,0365)	(0,0343)	(0,0311)	(0,0303)	(0,0298)	(0,0313)	
	$DB^\phi \cdot(h=0,5)$	4,7812	4,2322	3,6538	3,4057	3,1965	3,0216	**2,8562**	0,25
		(2,3092)	(0,5619)	(0,5063)	(0,4034)	(0,3574)	(0,3076)	(0,2572)	

APÊNDICE B

Tabelas para dados simulados da Distribuição Bingham Complexa

Tabela B.1: Validação do agrupamento do conjunto de dados simulados para $N = 30$ e $N = 50$ para distribuição Bingham com alta concentração $\boldsymbol{\lambda} = (200; 200; 0)$.

		Número de grupos							
N=30	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,4020**	0,2219	0,1961	0,1673	0,1525	0,1418	0,1355	0,98
		(0,0614)	(0,0385)	(0,0331)	(0,0380)	(0,0350)	(0,0330)	(0,0299)	
	S_m	**0,4592**	0,2688	0,2331	0,2021	0,1798	0,1715	0,1725	0,98
		(0,0577)	(0,0331)	(0,0342)	(0,0320)	(0,0310)	(0,0289)	(0,0323)	
	DB_m	**0,9084**	1,3753	1,4247	1,4439	1,4627	1,4202	1,3593	0,98
		(0,2683)	(0,1522)	(0,1652)	(0,1403)	(0,1562)	(0,1207)	(0,1141)	
Pré-Forma	DB_m^*	**0,9261**	1,4538	1,5717	1,6367	1,6509	1,6552	1,6178	0,98
		(0,2698)	(0,1479)	(0,1663)	(0,1476)	(0,1620)	(0,1600)	(0,1516)	
	DB_m^{**}	**0,8628**	1,3833	1,5113	1,5724	1,5861	1,5835		0,98
		(0,1610)	(0,1317)	(0,1594)	(0,1476)	(0,1562)	(0,1456)		
	$S_m^\phi \cdot_{(h=0,25)}$	**0,6563**	0,4107	0,3503	0,3005	0,2638	0,2553	0,2404	1,00
		(0,0349)	(0,0525)	(0,0586)	(0,0584)	(0,0514)	(0,0523)	(0,0515)	
	$DB_m^\phi \cdot_{(h=0,25)}$	**0,6158**	1,0016	1,0319	1,0538	1,0634	1,0101	0,9837	1,00
		(0,0588)	(0,0940)	(0,1140)	(0,1089)	(0,0967)	(0,0866)	(0,0802)	
	S	**0,4623**	0,02698	0,2339	0,2027	0,1803	0,1719	0,1730	0,98
		(0,0580)	(0,0334)	(0,0344)	(0,0323)	(0,0312)	(0,0291)	(0,0324)	
	DB	**0,9098**	1,3823	1,4305	1,4489	1,4674	1,4243	1,3626	0,98
		(0,2736)	(0,1540)	(0,1668)	(0,1412)	(0,1573)	(0,1223)	(0,1144)	
Coordenadas	DB^*	**0,5640**	1,3442	1,6025	1,7261	1,7791	1,8107	1,7309	0,98
		(0,5792)	(0,3214)	(0,3765)	(0,3474)	(0,3841)	(0,4339)	(0,3532)	
Tangentes	DB^{**}	**0,4724**	1,2229	1,4887	1,6084	1,6493	1,6657		0,98
		(0,2659)	(0,2712)	(0,3433)	(0,3241)	(0,3461)	(0,3473)		
	$S^\phi \cdot_{(h=0,25)}$	**0,6725**	0,4190	0,3566	0,3046	0,2669	0,2578	0,2424	1,00
		(0,0347)	(0,0540)	(0,0604)	(0,0599)	(0,0528)	(0,0534)	(0,0525)	
	$DB^\phi \cdot_{(h=0,25)}$	**0,4135**	1,0229	1,1142	1,1649	1,1956	1,0685	1,0116	1,00
		(0,0581)	(0,2285)	(0,2964)	(0,2846)	(0,2778)	(0,2003)	(0,1683)	

		Número de grupos							
N=50	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,3916**	0,2205	0,1897	0,1661	0,1535	0,1493	0,1438	1,00
		(0,0410)	(0,0315)	(0,0324)	(0,0273)	(0,0266)	(0,0301)	(0,0260)	
	S_m	**0,4657**	0,2701	0,2310	0,1969	0,1826	0,1734	0,1721	1,00
		(0,0241)	(0,0258)	(0,0285)	(0,0279)	(0,0256)	(0,0210)	(0,0192)	
	DB_m	**0,8809**	1,4238	1,5185	1,5821	1,5665	1,5302	1,4800	1,00
		(0,0518)	(0,1106)	(0,1382)	(0,1585)	(0,1195)	(0,0972)	(0,0924)	
Pré-Forma	DB_m^*	**0,8938**	1,4847	1,6258	1,7144	1,7148	1,7026	1,6682	1,00
		(0,0538)	(0,1194)	(0,1503)	(0,1540)	(0,1238)	(0,1331)	(0,1228)	
	DB_m^{**}	**0,8533**	1,4186	1,5606	1,6558	1,6661	1,6432		1,00
		(0,0531)	(0,1134)	(0,1446)	(0,1415)	(0,1197)	(0,1171)		
	$S_m^\phi \cdot_{(h=0,25)}$	**0,6567**	0,4129	0,3479	0,3088	0,2904	0,2709	0,2635	1,00
		(0,0276)	(0,0340)	(0,0444)	(0,0410)	(0,0389)	(0,0324)	(0,0332)	
	$DB_m^\phi \cdot_{(h=0,25)}$	**0,6242**	1,0433	1,1221	1,1336	1,1160	1,0952	1,0724	1,00
		(0,0493)	(0,0634)	(0,0847)	(0,0836)	(0,0753)	(0,0607)	(0,0631)	
	S	**0,4688**	0,2711	0,2319	0,1976	0,1832	0,1741	0,1727	1,00
		(0,0242)	(0,0261)	(0,0287)	(0,0280)	(0,0258)	(0,0211)	(0,0193)	
	DB	**0,8818**	1,4316	1,5251	1,5875	1,5710	1,5342	1,4829	1,00
		(0,0526)	(0,1119)	(0,1398)	(0,1593)	(0,1206)	(0,0977)	(0,0930)	
Coordenadas	DB^*	**0,4875**	1,3942	1,6926	1,8828	1,8925	1,8912	1,7935	1,00
		(0,0586)	(0,2572)	(0,3349)	(0,3547)	(0,3166)	(0,3423)	(0,2964)	
Tangentes	DB^{**}	**0,4466**	1,2788	1,5755	1,7664	1,7904	1,7459		1,00
		(0,0578)	(0,2368)	(0,3150)	(0,3152)	(0,2903)	(0,2757)		
	$S^\phi \cdot_{(h=0,25)}$	**0,6729**	0,4214	0,3543	0,3140	0,2945	0,2745	0,2666	1,00
		(0,0274)	(0,0351)	(0,0457)	(0,0422)	(0,0398)	(0,0335)	(0,0338)	
	$DB^\phi \cdot_{(h=0,25)}$	**0,4211**	1,0986	1,3112	1,3383	1,2945	1,2390	1,1891	1,00
		(0,0493)	(0,1668)	(0,2443)	(0,2434)	(0,2029)	(0,1647)	(0,1598)	

Tabela B.2: Validação do agrupamento do conjunto de dados simulados para $N = 30$ e $N = 50$ para distribuição Bingham com alta concentração $\boldsymbol{\lambda} = (100; 100; 0)$.

		Número de grupos							
N=30	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,2994**	0,1933	0,1767	0,1621	0,1511	0,1422	0,1429	0,89
		(0,0791)	(0,0355)	(0,0361)	(0,0383)	(0,0338)	(0,0331)	(0,0350)	
	S_m	**0,3190**	0,2040	0,1849	0,1731	0,1596	0,1539	0,1577	0,89
		(0,0756)	(0,0268)	(0,0296)	(0,0288)	(0,0258)	(0,0262)	(0,0272)	
	DB_m	**1,3274**	1,6312	1,5720	1,5173	1,5084	1,4610	1,4095	0,82
		(0,4493)	(0,1472)	(0,1775)	(0,1505)	(0,1453)	(0,1218)	(0,1126)	
Pré-Forma	DB_m^*	**1,3531**	1,7266	1,7047	1,6798	1,6953	1,6784	1,6543	0,89
		(0,4608)	(0,1491)	(0,1939)	(0,1771)	(0,1664)	(0,1585)	(0,1488)	
	DB_m^{**}	**1,2336**	1,6441	1,6347	1,6166	1,6321	1,6158		0,89
		(0,2920)	(0,1421)	(0,1754)	(0,1591)	(0,1544)	(0,1444)		
	$S_m^\phi \cdot_{(h=2,5)}$	**0,4963**	0,3298	0,2886	0,2737	0,2546	0,2390	0,2325	0,90
		(0,1069)	(0,0484)	(0,0467)	(0,0417)	(0,0470)	(0,0435)	(0,0545)	
	$DB_m^\phi \cdot_{(h=2,5)}$	**0,9110**	1,1438	1,1166	1,0833	1,0538	1,0419	1,0119	0,83
		(0,2229)	(0,0835)	(0,0835)	(0,0832)	(0,0846)	(0,0746)	(0,0791)	
	S	**0,3227**	0,2057	0,1862	0,1742	0,1606	0,1547	0,1586	0,89
		(0,0765)	(0,0271)	(0,0300)	(0,0292)	(0,0261)	(0,0265)	(0,0273)	
	DB	**1,3407**	1,6470	1,5843	1,5274	1,5174	1,4688	1,4159	0,82
		(0,4636)	(0,1503)	(0,1804)	(0,1538)	(0,1478)	(0,1230)	(0,1135)	
Coordenadas	DB^*	**1,2648**	1,9206	1,8795	1,8248	1,8695	1,8524	1,8142	0,89
		(1,1140)	(0,3852)	(0,4839)	(0,4516)	(0,4529)	(0,4226)	(0,3919)	
Tangentes	DB^{**}	**1,0069**	1,7472	1,7356	1,6982	1,7365	1,7286		0,89
		(0,5760)	(0,3439)	(0,4216)	(0,3874)	(0,3960)	(0,3882)		
	$S^\phi \cdot_{(h=2,5)}$	**0,5030**	0,3333	0,2910	0,2760	0,2566	0,2404	0,2339	0,9
		(0,1081)	(0,0492)	(0,0473)	(0,0421)	(0,0475)	(0,0443)	(0,0551)	
	$DB^\phi \cdot_{(h=2,5)}$	**0,9649**	1,3893	1,2986	1,2091	1,1532	1,1206	1,0690	0,85
		(0,7782)	(0,2738)	(0,2364)	(0,2167)	(0,2249)	(0,1810)	(0,1858)	

		Número de grupos							
N=50	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,2807**	0,1804	0,1630	0,1598	0,1489	0,1452	0,1418	0,92
		(0,0626)	(0,0320)	(0,0331)	(0,0287)	(0,0307)	(0,0301)	(0,0297)	
	S_m	**0,3245**	0,2090	0,1915	0,1819	0,1737	0,1695	0,1645	0,92
		(0,0647)	(0,0236)	(0,0235)	(0,0212)	(0,0208)	(0,0201)	(0,0204)	
	DB_m	**1,3060**	1,6588	1,5927	1,5543	1,5222	1,4960	1,4723	0,92
		(0,4013)	(0,1220)	(0,1365)	(0,1351)	(0,1070)	(0,0997)	(0,0939)	
Pré-Forma	DB_m^*	**1,3256**	1,7256	1,7001	1,6774	1,6618	1,6640	1,6520	0,92
		(0,4071)	(0,1217)	(0,1454)	(0,1487)	(0,1264)	(0,1251)	(0,1300)	
	DB_m^{**}	**1,2232**	1,6456	1,6291	1,6150	1,6113	1,6066		0,92
		(0,2654)	(0,1148)	(0,1364)	(0,1384)	(0,1201)	(0,1202)		
	$S_m^\phi \cdot_{(h=1,0)}$	**0,5043**	0,3327	0,3036	0,2915	0,2777	0,2660	0,2608	0,95
		(0,0768)	(0,0361)	(0,0410)	(0,0359)	(0,0324)	(0,0332)	(0,0352)	
	$DB_m^\phi \cdot_{(h=4,0)}$	**0,8930**	1,1525	1,1146	1,0999	1,0817	1,0723	1,0452	0,94
		(0,1674)	(0,0610)	(0,0563)	(0,0744)	(0,0667)	(0,0639)	(0,0543)	
	S	**0,3283**	0,2108	0,1928	0,1833	0,1750	0,1708	0,1656	0,92
		(0,0655)	(0,0240)	(0,0237)	(0,0213)	(0,0210)	(0,0204)	(0,0206)	
	DB	**1,3190**	1,6752	1,6065	1,5646	1,5306	1,5032	1,4785	0,92
		(0,4149)	(0,1250)	(0,1390)	(0,1368)	(0,1085)	(0,1016)	(0,0950)	
Coordenadas	DB^*	**1,2065**	1,9060	1,8615	1,8099	1,7907	1,7927	1,7811	0,92
		(1,0374)	(0,3212)	(0,3599)	(0,3555)	(0,3167)	(0,3088)	(0,3328)	
Tangentes	DB^{**}	**0,9838**	1,7469	1,7192	1,6803	1,6848	1,6776		0,92
		(0,5459)	(0,2863)	(0,3255)	(0,3125)	(0,2972)	(0,2801)		
	$S^\phi \cdot_{(h=1,00)}$	**0,5138**	0,3380	0,3079	0,2952	0,2811	0,2691	0,2636	0,95
		(0,0781)	(0,0370)	(0,0420)	(0,0367)	(0,0329)	(0,0338)	(0,0356)	
	$DB^\phi \cdot_{(h=4,0)}$	**0,8751**	1,3783	1,2706	1,2408	1,1986	1,1755	1,1120	0,94
		(0,6388)	(0,1854)	(0,1601)	(0,2062)	(0,1811)	(0,1608)	(0,1254)	

Tabela B.3: Validação do agrupamento do conjunto de dados simulados para $N = 30$ e $N = 50$ para distribuição Bingham com baixa concentração $\boldsymbol{\lambda} = (2; 1, 5; 0)$.

		Número de grupos							
N=30	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,1726**	0,1264	0,0912	0,0735	0,0538	0,0510	0,0394	0,58
		(0,0862)	(0,0516)	(0,0604)	(0,0581)	(0,0545)	(0,0463)	(0,0570)	
	S_m	0,1036	0,1023	0,1071	0,1075	0,1082	0,1098	**0,1118**	0,11
		(0,0180)	(0,0151)	(0,0166)	(0,0184)	(0,0212)	(0,0240)	(0,0253)	
	DB_m	1,4225	1,4386	1,4164	1,4141	1,4023	1,3896	**1,3591**	0,13
		(0,0860)	(0,1008)	(0,0926)	(0,0953)	(0,1006)	(0,1012)	(0,1083)	
Pré-Forma	DB_m^*	**1,4408**	1,4710	1,4711	1,5003	1,4985	1,5084	1,5035	0,34
		(0,0881)	(0,1041)	(0,0974)	(0,1005)	(0,1142)	(0,1125)	(0,1074)	
	DB_m^{**}	**1,3218**	1,3721	1,3929	1,4306	1,4363	1,4473		0,50
		(0,0718)	(0,0919)	(0,0839)	(0,0849)	(0,0941)	(0,0991)		
	$S_m^\phi \cdot_{(h=3,25)}$	0,1607	0,1589	0,1641	**0,1660**	0,1649	0,1618	0,1582	0,09
		(0,0258)	(0,0222)	(0,0243)	(0,0287)	(0,0334)	(0,0335)	(0,0341)	
	$DB_m^\phi \cdot_{(h=0,1)}$	1,9681	1,9484	1,9270	1,9042	1,8799	1,8543	**1,8286**	0,00
		(0,0038)	(0,0060)	(0,0074)	(0,0103)	(0,0121)	(0,0130)	(0,0170)	
	S	0,1662	0,1612	0,1642	0,1630	**0,1669**	0,1665	0,1662	0,18
		(0,0258)	(0,0179)	(0,0168)	(0,0195)	(0,0214)	(0,0226)	(0,0258)	
	DB	2,1081	1,8320	1,6604	1,5667	1,4845	1,4441	**1,3851**	0,00
		(0,1951)	(0,1207)	(0,1137)	(0,1228)	(0,1060)	(0,1088)	(0,1065)	
Coordenadas	DB^*	2,5803	2,0223	1,7346	1,6370	1,5077	1,5052	**1,4688**	0,01
		(0,4679)	(0,2857)	(0,2667)	(0,2738)	(0,2467)	(0,2512)	(0,2183)	
Tangentes	DB^{**}	2,1924	1,7632	1,5395	1,4628	**1,3889**	1,3963		0,02
		(0,3573)	(0,2372)	(0,2103)	(0,2014)	(0,1962)	(0,2250)		
	$S^\phi \cdot_{(h=2,5)}$	0,2432	0,2456	**0,2541**	0,2476	0,2443	0,2452	0,2423	0,16
		(0,0314)	(0,0270)	(0,0277)	(0,0343)	(0,0332)	(0,0409)	(0,0429)	
	$DB^\phi \cdot_{(h=0,5)}$	3,8378	2,8164	2,2965	1,9957	1,8127	1,6912	**1,5457**	0,00
		(0,4531)	(0,3105)	(0,2948)	(0,2297)	(0,2314)	(0,2553)	(0,1919)	

		Número de grupos							
N=50	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,1959**	0,1560	0,1181	0,0890	0,0682	0,0515	0,0502	0,58
		(0,0809)	(0,0429)	(0,0407)	(0,0443)	(0,0451)	(0,0482)	(0,0461)	
	S_m	0,1025	0,1066	0,1093	0,1118	0,1137	0,1113	**0,1148**	0,05
		(0,0140)	(0,0141)	(0,0146)	(0,0133)	(0,0162)	(0,0157)	(0,0186)	
	DB_m	1,4486	1,4269	1,4310	1,4095	1,4104	1,4167	**1,4071**	0,06
		(0,0749)	(0,0792)	(0,0789)	(0,0773)	(0,0794)	(0,0869)	(0,1008)	
Pré-Forma	DB_m^*	1,4621	**1,4517**	1,4703	1,4620	1,4815	1,4990	1,5006	0,16
		(0,0765)	(0,0802)	(0,0839)	(0,0809)	(0,0800)	(0,0891)	(0,1090)	
	DB_m^{**}	**1,3390**	1,3631	1,3897	1,4020	1,4248	1,4477		0,4
		(0,0647)	(0,0722)	(0,0688)	(0,0698)	(0,0760)	(0,0876)		
	$S_m^\phi \cdot_{(h=2,5)}$	0,1573	0,1623	0,1717	0,1733	0,1727	**0,1748**	0,1748	0,06
		(0,0215)	(0,0185)	(0,0182)	(0,0224)	(0,0235)	(0,0248)	(0,0278)	
	$DB_m^\phi \cdot_{(h=0,25)}$	1,8734	1,8170	1,7719	1,7265	1,6952	1,6618	**1,6338**	0,00
		(0,0133)	(0,0178)	(0,0219)	(0,0223)	(0,0265)	(0,0294)	(0,0258)	
	S	0,1654	0,1673	0,1697	0,1720	**0,1753**	0,1710	0,1747	0,09
		(0,0197)	(0,0144)	(0,0150)	(0,0126)	(0,0144)	(0,0134)	(0,0168)	
	DB	2,1606	1,8288	1,6666	1,5500	1,4949	1,4666	**1,4251**	0,00
		(0,1602)	(0,0833)	(0,0903)	(0,0931)	(0,0781)	(0,0716)	(0,0779)	
Coordenadas	DB^*	2,6787	1,9799	1,7045	1,5142	1,4412	1,4506	**1,4178**	0,00
		(0,3904)	(0,1828)	(0,1922)	(0,1964)	(0,1629)	(0,1654)	(0,1767)	
Tangentes	DB^{**}	2,2910	1,7344	1,5048	1,3743	**1,3339**	1,3446		0,00
		(0,3155)	(0,1425)	(0,1447)	(0,1605)	(0,1386)	(0,1532)		
	$S^\phi \cdot_{(h=2,5)}$	0,2507	0,2533	0,2643	0,2657	0,2682	**0,2722**	0,2665	0,07
		(0,0289)	(0,0190)	(0,0179)	(0,0237)	(0,0231)	(0,0223)	(0,0272)	
	$DB^\phi \cdot_{(h=1,0)}$	3,1451	2,2524	1,8258	1,6296	1,4764	1,3854	**1,3151**	0,00
		(0,3888)	(0,2076)	(0,1844)	(0,1873)	(0,1617)	(0,1466)	(0,1584)	

Tabela B.4: Validação do agrupamento do conjunto de dados simulados para $N = 30$ e $N = 50$ para distribuição Bingham com baixa concentração $\boldsymbol{\lambda} = (1; 1; 0)$.

		Número de grupos							
N=30	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,1391**	0,1080	0,0751	0,0467	0,0363	0,0204	0,0068	0,49
		(0,0860)	(0,0612)	(0,0583)	(0,0605)	(0,0504)	(0,0567)	(0,0545)	
	S_m	0,0963	0,0994	0,1011	0,1025	0,1018	**0,1041**	0,0990	0,05
		(0,0143)	(0,0152)	(0,0167)	(0,0168)	(0,0182)	(0,0195)	(0,0218)	
	DB_m	1,4031	**1,3832**	1,3962	1,4095	1,4044	1,3969	1,3843	0,04
		(0,0539)	(0,0808)	(0,0819)	(0,0966)	(0,0937)	(0,1011)	(0,1115)	
Pré-Forma	DB_m^*	1,4203	**1,4132**	1,4436	1,4816	1,4977	1,5122	1,5111	0,19
		(0,0550)	(0,0870)	(0,0881)	(0,1051)	(0,1057)	(0,1070)	(0,1224)	
	DB_m^{**}	**1,2940**	1,3194	1,3644	1,4151	1,4342	1,4409		0,48
		(0,,0434)	(0,0712)	(0,0701)	(0,0901)	(0,0935)	(0,0911)		
	$S_m^\phi \cdot_{(h=1,75)}$	0,1403	0,1427	0,1482	**0,1516**	0,1478	0,1495	0,1509	0,7
		(0,0201)	(0,0213)	(0,0205)	(0,0244)	(0,0310)	(0,0349)	(0,0364)	
	$DB_m^\phi \cdot_{(h=0,1)}$	1,9710	1,9537	1,9330	1,9112	1,8874	1,8647	**1,8430**	0,00
		(0,0028)	(0,0048)	(0,0071)	(0,0102)	(0,0106)	(0,0147)	(0,0155)	
	S	0,1629	0,1631	0,1660	**0,1668**	0,1648	0,1656	0,1607	0,16
		(0,0208)	(0,0153)	(0,0190)	(0,0172)	(0,0226)	(0,0203)	(0,0240)	
	DB	2,1313	1,8162	1,6514	1,5530	1,4811	1,4344	**1,3994**	0,00
		(0,1703)	(0,1010)	(0,1131)	(0,1330)	(0,1006)	(0,1045)	(0,1013)	
Coordenadas	DB^*	2,6042	1,9599	1,6900	1,5656	1,4979	1,4824	**1,4618**	0,00
		(0,4175)	(0,2254)	(0,2555)	(0,3188)	(0,2661)	(0,2272)	(0,2597)	
Tangentes	DB^{**}	2,2253	1,6912	1,4877	1,4155	1,3692	**1,3551**		0,00
		(0,3198)	(0,1682)	(0,1933)	(0,2618)	(0,2436)	(0,2262)		
	$S^\phi \cdot_{(h=1,75)}$	0,2352	0,2359	0,2426	**0,2465**	0,2379	0,2360	0,2395	0,1
		(0,0292)	(0,0237)	(0,0255)	(0,0273)	(0,0369)	(0,0414)	(0,375)	
	$DB^\phi \cdot_{(h=1,0)}$	3,0832	2,2642	1,8670	1,6493	1,4658	1,3981	**1,2662**	0,00
		(0,4381)	(0,2464)	(0,2524)	(0,2446)	(0,2096)	(0,2320)	(0,1825)	

		Número de grupos							
N=50	**Índices**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Tx.**
	PR	**0,1391**	0,1160	0,0792	0,0622	0,0496	0,0362	0,0252	0,45
		(0,0851)	(0,0490)	(0,0470)	(0,0548)	(0,0478)	(0,0475)	(0,0503)	
	S_m	0,0946	0,1029	0,1063	0,1058	0,1072	**0,1073**	0,1066	0,06
		(0,0122)	(0,0120)	(0,0124)	(0,0149)	(0,0151)	(0,0144)	(0,0158)	
	DB_m	1,4161	**1,3764**	1,3983	1,4171	1,4120	1,4119	1,4247	0,05
		(0,0486)	(0,0736)	(0,0581)	(0,0817)	(0,0761)	(0,0837)	(0,0859)	
Pré-Forma	DB_m^*	1,4285	**1,3950**	1,4272	1,4705	1,4747	1,4926	1,5162	0,16
		(0,0530)	(0,0773)	(0,0603)	(0,0928)	(0,0805)	(0,0939)	(0,0917)	
	DB_m^{**}	**1,2988**	1,3090	1,3577	1,4034	1,4240	1,4427		0,45
		(0,0390)	(0,0609)	(0,0528)	(0,0786)	(0,0720)	(0,0833)		
	$S_m^\phi \cdot_{(h=4,0)}$	0,1474	0,1603	0,1634	0,1688	0,1714	**0,1718**	0,1668	0,03
		(0,0182)	(0,0185)	(0,0175)	(0,0211)	(0,0208)	(0,0250)	(0,0297)	
	$DB_m^\phi \cdot_{(h=0,1)}$	1,7230	1,6294	1,5535	1,4961	1,4550	1,4238	**1,3840**	0,00
		(0,0244)	(0,0344)	(0,0393)	(0,0440)	(0,0481)	(0,0480)	(0,0491)	
	S	0,1625	0,1675	0,1734	0,1719	0,1752	**0,1759**	0,1758	0,11
		(0,0171)	(0,0116)	(0,0137)	(0,0146)	(0,0165)	(0,0155)	(0,0164)	
	DB	2,1885	1,8243	1,6527	1,5577	1,4925	1,4484	**1,4200**	0,00
		(0,1345)	(0,0800)	(0,0823)	(0,0998)	(0,0821)	(0,0793)	(0,0740)	
Coordenadas	DB^*	2,7162	1,9414	1,6275	1,5190	1,4227	1,3797	**1,3664**	0,00
		(0,3240)	(0,1798)	(0,1723)	(0,2251)	(0,1975)	(0,1691)	(0,1649)	
Tangentes	DB^{**}	2,3103	1,6847	1,4565	1,3678	1,3092	**1,2774**		0,00
		(0,2627)	(0,1377)	(0,1442)	(0,1821)	(0,1696)	(0,1471)		
	$S^\phi \cdot_{(h=2,5)}$	0,2439	0,2501	0,2638	0,2654	**0,2755**	0,2731	0,2689	0,04
		(0,0257)	(0,0168)	(0,0204)	(0,0215)	(0,0223)	(0,0241)	(0,0280)	
	$DB^\phi \cdot_{(h=1,0)}$	3,2422	2,2890	1,8700	1,6525	1,4691	1,4019	**1,2836**	0,00
		(0,2957)	(0,2322)	(0,1904)	(0,2048)	(0,1665)	(0,1465)	(0,1283)	