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

Agrupamento Espectral para Dados de Formas

Diogo Vasconcelos Cândido

Junho/2017

Diogo Vasconcelos Cândido

Agrupamento Espectral para Dados de Formas

Monografia apresentada ao Curso de Bacharelado em Estatística da Universidade Federal da Paraíba como requisito parcial para obtenção do Grau de Bacharel. Área de Concentração: Estatística Aplicada.

João Pessoa
Junho de 2017

Catalogação na publicação
Biblioteca Setorial do CCEN/UFPB
Josélia M.O. Silva – CRB-15/113

C217a Cândido, Diogo Vasconcelos.
Agrupamento espectral para dados de formas / Diogo Vasconcelos Cândido. – João Pessoa, 2017.
55 p. : il. color.

Monografia (Bacharelado em Estatística) – Universidade Federal da Paraíba.
Orientador(a): Profº. Drº. Marcelo Rodrigo Portela Ferreira.

1. Análise estatística. 2. Agrupamento espectral. 3. Formas. 4. Algoritimo de Ng. I. Título.

UFPB/BS-CCEN CDU 519.23(043.2)

Ata da Sessão Pública de Defesa de Monografia de **Diogo Vasconcelos Cândido**, realizada em 02 de junho de 2017 no Departamento de Estatística da Universidade Federal da Paraíba - UFPB.

Aos dois dias do mês de junho de dois mil e dezessete, às quatorze horas, na Sala 20 do Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para avaliar o Trabalho de Conclusão Curso intitulado "AGRUPAMENTO ESPECTRAL PARA DADOS DE FORMAS" de autoria de DIOGO VASCONCELOS CÂNDIDO. A Banca Examinadora foi composta pelos professores: Prof. Dr. Marcelo Rodrigo Portela Ferreira (DE-UFPB, orientador), Prof. Dr. Eufrásio de Andrade Lima Neto (DE-UFPB, examinador) e Prof. Dr. João Agnaldo do Nascimento (DE-UFPB, examinador). Dando início aos trabalhos, o presidente da banca cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou à palavra ao discente para que se fizesse, oralmente, a exposição de seu trabalho de monografia. Concluída a apresentação, o discente foi arguido pela Banca Examinadora que sugeriu algumas alterações até o dia 12 de junho de 2017, de acordo com a Resolução No. 02/2014 do Colegiado do Curso de Bacharelado em Estatística da UFPB. Uma vez entregue a versão final do Trabalho de Conclusão de Curso à Coordenação do Bacharelado em Estatística, com as alterações solicitadas pela Banca Examinadora dentro do prazo estabelecido, o discente Diogo Vasconcelos Cândido será aprovado com nota ([redacted]), que é a média aritmética das notas atribuídas pelos membros da Banca Examinadora.

João Pessoa, 02 de junho de 2017.

Prof. Marcelo Rodrigo Portela Ferreira (Orientador)

Prof. Eufrásio de Andrade Lima Neto (Examinador)

Prof. João Agnaldo do Nascimento (Examinador)

Diogo Vasconcelos Cândido (Discente)

Este trabalho é dedicado à minha família e amigos, que, certamente, em muito me agregaram ao longo deste curso.

AGRADECIMENTOS

Agradeço ao Senhor Deus, pelo dom da vida e por me permitir dar mais um passo, ao concluir esta graduação. Sem Ele, nada poderia ter sido feito.

Aos meus pais, Ginaldo e Sandra, obrigado por serem meus maiores incentivadores e patrocinadores. Queridos pais, essa conquista é vossa. Minha eterna gratidão por todo o esforço empenhando em meu favor.

À você, minha segunda professora de todas as disciplinas, um "obrigado" do tamanho do meu amor por ti. Foram 8 períodos de contínuo suporte. Das mais simples até as mais complexas disciplinas, dos numerosos trabalhos em equipe e das infinitas listas de exercícios. Certamente eu ainda estaria no meio do caminho se você não estivesse nele. Obrigado, Adenice.

Meus amigos de turma, André e Anny, vocês são pra vida toda. Minha gratidão se estende a vocês, que em muito me agregaram ao longo desses anos. Obrigado por todo o bom humor e enorme parceria. Zé e Lukas, vocês são irmãos. Me sinto privilegiado por poder conviver com pessoas tão especiais.

Professor Marcelo, você é fantástico. Minha admiração por ti enquanto pessoa e profissional é grande. Foi um presente ter meu TCC orientado por você, e agora colhemos o fruto de um trabalho tão excelente que executamos em conjunto. Obrigado por ser tão paciente e fazer jus ao título de orientador.

Minha gratidão também não poderia de deixar de existir para com os atenciosos professores do DE, em especial aos que tive o privilégio de cursar alguma disciplina. Professora Ana Flávia, é até difícil expressar em poucas palavras o sentimento que temos a respeito da senhora. A palavra "obrigado", apesar de simples, traz um enorme peso de alegria em poder ter compartilhado vários momentos descontraídos ao seu lado. Professor Hemílio, você foi um paizão. Minha gratidão também se aplica ao senhor, que me deu a oportunidade de ser seu aluno projetista PIBIC. Apesar de toda a dificuldade, as lições extraídas ao longo de dois anos de trabalho valeram todo o esforço.

Agradeço a banca avaliadora deste trabalho pela disposição em contribuir com esta

monografia.

Aos amigos e familiares, sintam-se incluídos nesta seção de agradecimentos. Vocês são responsáveis por partilhar de momentos alegres e tristes, sempre com muito apoio. Deus te abençoe Adelma, Alysson, Andreia, Arielly, Bárbara, Brenda, Carol, Chico, Clarissa, Cris, Douglas, Edgard, Ellen, Flávia, Flor, Gê, Gilvandro, Jéssica, Jhon, João, Joyce, Leoncio, Livinha, Marinalva, Paloma, Phâmella e Samara.

"Não vos amoldeis às estruturas deste mundo,
mas transformai-vos pela renovação da mente,
a fim de distinguir qual é a vontade de Deus:
o que é bom, o que Lhe é agradável, o que é perfeito".
(Bíblia Sagrada, Romanos 12, 2)

RESUMO

A coleta de informações geométricas, a partir do avanço da tecnologia, e o estudo das formas de objetos tem se tornado cada vez mais comum e importante. A análise estatística de formas (AEF) utiliza métodos estatísticos para a análise de estruturas geométricas e suas aplicações podem ser encontradas em diferentes áreas da ciência. No entanto, um dos problemas de interesse na AEF é estender os métodos clássicos de análise estatística para dados de formas de objetos, ou propor novos métodos para esse tipo de dado. Na AEF é comum existir a necessidade de agrupamento em um conjunto de dados de modo a obter grupos com características mais homogêneas. Os métodos de agrupamento são ferramentas úteis para explorar estruturas em conjuntos de dados sendo utilizados, por exemplo, em para reconhecimento não-supervisionado de padrões. O método k-médias figura entre os métodos de agrupamento mais antigos e mais comumente utilizados na prática. Mas, apesar de sua simplicidade e eficiência, o algoritmo k-médias apresenta algumas deficiências. Por causa disso, há a necessidade da proposição de métodos alternativos que possam apresentar bons resultados em situações onde o algoritmo k-médias falha. Os métodos de agrupamento espectral surgem a partir de conceitos da teoria espectral dos grafos onde o problema de agrupamento é configurado como um problema de corte no grafo em que uma função objetivo apropriada deve ser otimizada. Neste trabalho apresentamos uma adaptação do algoritmo de agrupamento espectral de Ng, Jordan & Weiss para dados de formas planas de objetos e o comparamos a uma adapatação do algoritmo k-médias para dados de formas planas. Foram realizadas aplicações com 14 conjuntos de dados reais e verificou-se que o algoritmo espectral adaptado de Ng, Jordan & Weiss, considerando as distâncias de procrustes completa e euclidiana no espaço tangente obteve desempenho superior ao método de agrupamento k-médias, fornecendo evidências de que a adaptação proposta é eficiente para dados dessa natureza.

Palavras-chave: Agrupamento espectral; Formas; Algoritmo de Ng, Jordan & Weiss; k-médias.

ABSTRACT

With the advance of tecnology, the collection of geometrical information from images became usual. Statistical shape analysis uses statistical methods to analyse geometrical structures and can be applied in several areas. One particular problem of interest in statistical shape analysis is the adaptation of classical statistical methods for shape data or the proposition of new methods. In statistical shape analysis it is common the need for clustering shape data to obtain clusters with similar characteristics. Clustering methods are useful tools to explore structures in data and have been used for unsupervised pattern recognition. The k-means algorithm is among the oldest and most widely used clustering methods. Despite its simplicity and efficiency, the k-means algorithm has some problems. Because of this, it is important to propose alternate methods that can be useful where the k-means fails. Spectral clustering methods arise from spectral theory of graphs and the clustering problem can be formulated as a graph cut where an appropriate objective function should be optimized. In this work we propose an adaptation of the Ng, Jordan & Weiss spectral clustering algorithm for plannar shape data. We performed applications on 14 plannar shape data sets and verified that the adapted version of the Ng, Jordan & Weiss algorithm considering the full procrustes distance and the euclidean distance on the shapes tangent space outperforms the version of the$k$-means algorithm for plannar shapes, corroborating that the proposed adaptation is efficient for shape data.
Keywords: Spectral clustering; Shapes; Ng, Jordan & Weiss algorithm; k-means.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

CAPÍTULO 1

INTRODUÇÃO

1.1 Definição e Motivação

Objetos, sejam eles naturais ou artificiais, podem ser encontrados por toda parte. Com o avanço da tecnologia, a coleta de informações geométricas tem se tornado rotina e o estudo da forma de objetos tem se tornado cada vez mais importante (DRYDEN; MARDIA, 1998). A análise estatística de formas (AEF) é uma área de pesquisa relativamente nova, que aplica métodos estatísticos para a análise de estruturas geométricas. Em um sentido amplo, a formulação matemática da AEF é semelhante à análise estatística multivariada. Aplicações da AEF podem ser encontradas em diferentes áreas da ciência. Em visão computacional, as formas das fronteiras em imagens têm um importante papel. Em medicina, a análise da forma de dados anatômicos pode ajudar no diagnóstico de diversas doenças. As formas de caracteres escritos podem ser usadas como meios primários de identificação. Em biometria humana, a identificação de pessoas através de suas digitais ou escaneamento facial são exemplos de análise de formas. A Tabela 1.1 apresenta algumas aplicações onde a análise de formas pode ser imprescindível (COSTA; JR, 2000).

O trabalho seminal "*The diffusion of shape*" em análise de formas foi publicado em 1977, por David Kendall, e apresenta um breve resumo no qual é introduzida uma nova representação de formas de objetos em espaços complexos projetados (KENDALL, 1977). Alguns anos depois, Kendall (1984) definiu forma como toda informação geométrica que resta quando os efeitos de tamanho, rotação e escala são retiradas de um objeto. Kendall também propôs um sistema de coordenadas com finalidade de obter a forma de um objeto através de pontos dispostos em seu contorno chamados de marcos anatômicos. A formulação matemática para o estudo das formas foi introduzida em Bookstein (1984)

Tabela 1.1: Áreas do conhecimento e algumas aplicações da análise formas.

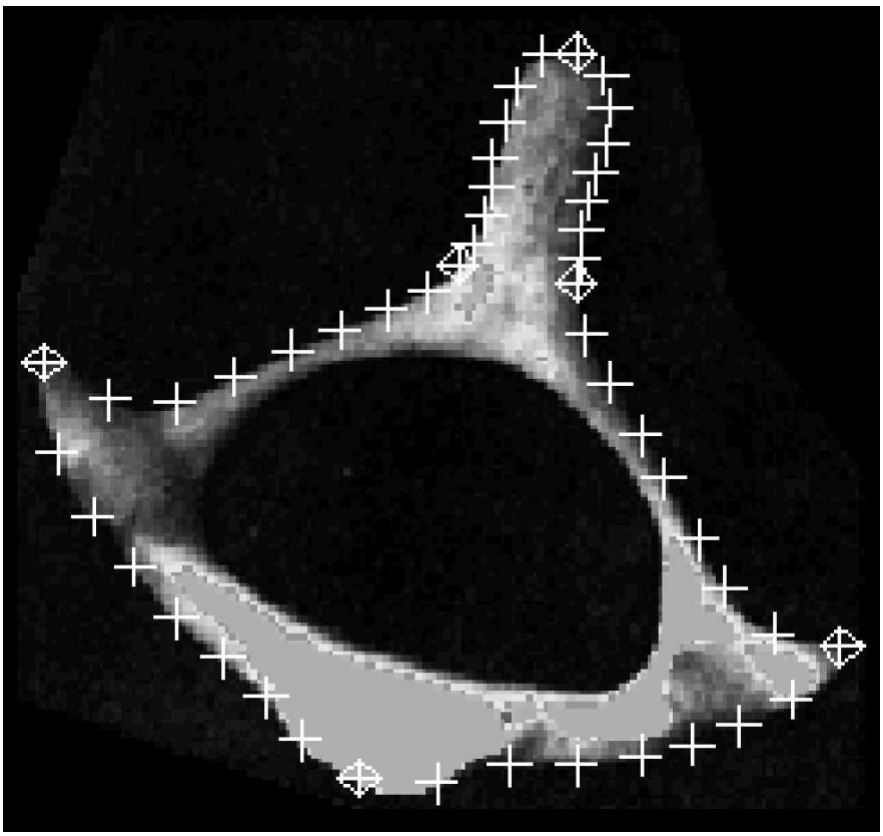
Área do Conhecimento	**Exemplo de Aplicações**
Neurociência	Taxonomia morfológica de neurônios, investigações sobre a relação função e forma das células, comparação entre células de diferentes áreas corticais e de diferentes espécies, modelagem de células biologicamente realistas.
Análise de Documentos	Análise de documentos eletrônicos (textuais ou visuais) como em aplicações CBIR, sistemas OCR (*Optical Character Recognitions*), análise de dados em banco de dados multimídia.
"Artes Visuais"	Restauração de vídeo, efeitos especiais, monitoramento de vídeo, jogos, computação gráfica, síntese de imagens.
Medicina	Identificação de tumor, quantificação de mudança e deformação de estruturas anatômicas, análise numérica de cromossomos, identificação de doenças genéticas.
Biologia	Identificação de espécies, taxonomia, relação entre forma e função, comparativo de anatomias, citologia, identificação e contagem de células (como glóbulos brancos no sangue),caracterização de células e formas nucleares, crescimento e modificação de formas de estruturas, análise da forma de caminhar dos seres vivos.
Física	Aplicações envolvendo microscopia como a análise da trajetória,comportamento e distribuição de partículas em um meio ou material, análise de estruturas como polímeros e cristais,caracterização de grupos de estrelas ou análise de propriedades de corpos celestes em astronomia, análise do movimento de objetos macroscópicos (velocidade, aceleração, etc.).
Engenharia	Controle visual de qualidade de produtos em linha de produção,detecção de perigo, interpretação (pelas máquinas) de desenhos feitos à mão, automação, robótica, sensoriamento remoto.
Segurança	Detecção de impressão digital, face e íris, verificação de assinatura e modo de andar de uma pessoa (biometria).
Agricultura	Controle de colheita, contagem de sementes, análise da maturação de frutos.

e Bookstein (1986).

Apesar de análise de formas poder ser investigada através de métodos matemáticos tradicionais (ROHLF; BOOKSTEIN, 1990), há um interesse em realizar comparação direta da forma dos organismos (MARE; CORSEUIL, 2004). A utilização de marcos anatômicos é um desses métodos de comparação direta, usado para representar as formas de uma maneira compreensiva por meio de dados numéricos médios. Os dados podem ser obtidos através de coordenadas médias de marcos distribuídos perifericamente, ou dentro das estruturas analisadas, ou das distâncias entre os marcos anatômicos escolhidos diretamente para cada espécie.

Um marco é um ponto de correspondência em cada objeto que coincide entre e dentro das populações ou grupos, ou seja, todos os objetos possuem os mesmos marcos. Um marco anatômico, por sua vez, é um ponto atribuído por um especialista, que corresponde entre organismos, de alguma forma biologicamente significativa, por exemplo, o canto de um olho ou a reunião de duas suturas em um crânio. Já marcos matemáticos são pontos localizados num objeto de acordo com alguma propriedade matemática ou geométrica da imagem, por exemplo, pontos de alta curvatura ou pontos extremos. Por fim, pseudo-marcos são pontos construídos em um objeto, localizados ao redor do contorno ou entre marcos anatômicos ou matemáticos. Na figura 1.1 vemos 6 marcos matemáticos em pontos de alta curvatura e 42 pseudo-marcos marcados no contorno de uma segunda vértebra torácica (T2) de rato. Os marcos ainda podem ser definidos em três tipos adicionais, onde os marcos do tipo I ocorrem na junção de tecidos/ossos; os do tipo II são definidos por propriedades locais, tais como curvaturas máximas e os marcos do tipo III ocorrem em pontos extremos ou marcos construídos, tais como diâmetros máximos e centróides.

Figura 1.1: Vértebra T2 de rato com 6 marcos matemáticos (junções de linhas) e 42 pseudo-marcos.



Fonte: Dryden e Mardia (1998).

Um dos problemas de interesse em AEF é estender métodos clássicos de análise estatística para dados de formas de objetos ou propôr novos métodos para este tipo de dado.

É comum existir, em análise de formas, a necessidade de agrupamento em um conjunto de dados, de tal forma que se obtenha grupos com características mais homogêneas. Por exemplo, quando se deseja detectar o número de diferentes espécies, a análise de agrupamento pode ser um excelente recurso para detecção dos grupos (AMARAL et al., 2010). Dessa forma, diversos pesquisadores têm conduzido estudos e têm estudado a performance desses métodos.

Os métodos de agrupamento são ferramentas úteis para explorar estruturas em conjuntos de dados sendo utilizados, por exemplo, em para reconhecimento não-supervisionado de padrões. A tarefa de agrupar significa organizar um conjunto de observações (indivíduos, objetos, etc.) em grupos de tal forma que observações pertencentes a um dado grupo têm um alto grau de similaridade, enquanto que observações pertencentes a grupos diferentes têm um alto grau de dissimilaridade (GORDON, 1999; JAIN; MURTY; FLYNN, 1999; XU; WUNSCH, 2005). Esses métodos vêm sendo largamente aplicados

em diversas áreas da ciência, tais como, taxonomia, processamento de imagens, mineração de dados, recuperação de informação, dentre outras.

As técnicas de agrupamento mais populares podem ser divididas em métodos hierárquicos e métodos particionais. Os métodos hierárquicos produzem uma resposta representada por uma estrutura completa de hierarquia, i.e., uma sequência aninhada de partições do conjunto de observações de entrada; sua saída é uma estrutura hierárquica de grupos conhecida como dendrograma. Por outro lado, nos métodos particionais o objetivo é obter uma partição única do conjunto de observações em um número fixo de grupos, tipicamente através da otimização (geralmente local) de uma função objetivo; o resultado é a criação de hipersuperfícies de separação entre os grupos.

Os métodos de agrupamento particionais foram desenvolvidos sob dois diferentes paradigmas: agrupamento rígido (*hard*) e agrupamento difuso (*fuzzy*). Nos métodos de agrupamento do tipo rígido, os grupos são naturalmente disjuntos e não se sobrepõem. Nesse caso, cada padrão pode pertencer a um, e somente um, grupo. No caso dos métodos de agrupamento do tipo difuso, um padrão pode pertencer a todos os grupos com um certo grau de pertinência. Uma exposição detalhada dos principais métodos de agrupamento difuso pode ser encontrada em Höppner (1999). Em adição, uma boa revisão sobre os vários métodos de agrupamento pode ser encontrada, por exemplo, em Jain (2010) ou em Jain, Murty e Flynn (1999).

O método k-médias (HARTIGAN; WONG, 1979) figura entre os métodos de agrupamento mais antigos e mais utilizados em problemas práticos (JAIN, 2010). Tal método consiste na divisão de um conjunto de observações sobre um espaço métrico em k grupos, de maneira que a soma dos quadrados das distâncias entre cada observação e a média do grupo ao qual ela pertence seja a mínima possível. Esta divisão é, em geral, obtida por meio de algoritmos iterativos. Amaral et al. (2010) propuseram uma adaptação do método k-médias clássico para dados de formas planas (bidimensionais) de objetos. Eles consideraram três tipos de distâncias apropriadas para dados de formas, além de uma versão do método baseado na distância Euclidiana clássica obtida no espaço tangente das formas. Apesar de sua simplicidade e eficiência, o algoritmo k-médias apresenta algumas deficiências. Por essa razão, surge a necessidade da proposição de métodos alternativos, que possam apresentar bons resultados em situações onde o algoritmo k-médias falha.

Os métodos de agrupamento espectral surgem de conceitos da teoria espectral dos grafos e o agrupamento é configurado como um problema de poda no grafo, onde uma função objetivo apropriada deve ser otimizada. A ideia básica é construir um grafo ponderado a partir do conjunto de dados inicial, onde cada nó representa uma observação e cada aresta ponderada simplesmente leva em consideração a similaridade entre duas observa-

ções. A ideia principal desta teoria é a decomposição espectral da matriz laplaciana do grafo ponderado obtido a partir dos dados originais (FILIPPONE et al., 2008). Verma e Meila (2003), Kannan, Vempala e Vetta (2004), Shawe-Taylor e Kandola (2001) apresentam comparações entre diversos métodos de agrupamento espectral e métodos clássicos de agrupamento, enquanto Luxburg (2007) apresenta um tutorial sobre métodos de agrupamento espectral.

Neste trabalho apresentaremos uma adaptação do algoritmo de agrupamento espectral de Ng, Jordan e Weiss (NG; JORDAN; WEISS, 2001) para dados de formas planas de objetos. Partiremos da definição de funções kernel adequadas para dados de formas (JAYASUMANA et al., 2013), introduzindo um algoritmo de agrupamento espectral adequado para dados de formas e um algoritmo de agrupamento espectral no espaço tangente das formas. Aplicações com conjuntos de dados reais irão ilustras a utilidade dos métodos propostos.

1.2 Organização do Trabalho

Além do capítulo de introdução, este trabalho é composto por mais três capítulos. No Capítulo 2 apresentamos uma revisão geral sobre Análise Estatística de Formas, em que introduzindo uma representação matemática de formas seguido de subseções contendo informações acerca da distância de procrustes para caso planar, coordenadas no espaço tangente, além de apresentarmos o algoritmo k-média para dados de formas e uma seção explicando as funções kernel. Ainda neste capítulo, é feita uma revisão sobre agrupamento espectral e o algoritmo espectral de Ng, Jordan & Weiss. Finalizaremos o capítulo com uma seção contendo nossa proposta de trabalho. No Capítulo 3, índices de avaliação serão utilizados em conjuntos de dados reais, objetivando-se mensurar a eficácia dos algoritmos. Finalmente, no Capítulo 4, são apresentadas as considerações finais acerca do trabalho e algumas sugestões para trabalhos futuros.

CAPÍTULO 2

REFERENCIAL TEÓRICO

2.1 Análise Estatística de Formas (AEF)

Nesta seção, apresentaremos uma revisão geral sobre AEF, introduzindo uma representação matemática desse tipo de análise, em que conceitos básicos para compreensão da teoria abordada serão definidos; seguido das explicações para o entendimento da distância de procrustes para o caso planar e estudo de coordenadas no espaço tangente.

2.1.1 Representação Matemática de Formas

A especificação de um sistema de coordenadas é essencial para descrição sobre a forma de um objeto. São vários os sistemas de coordenadas: coordenadas de Bookstein, propostas por Bookstein (1984, 1986); coordenadas polares de Kent, proposta por Kent (1994); coordenadas de forma Goodall-Mardia QR desenvolvida por Goodall e Mardia (1992, 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 $\boldsymbol{X}$ 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.

$$\boldsymbol{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ x_{2,1} & \cdots & x_{2,m} \\ \vdots & \ddots & \vdots \\ x_{k,1} & \cdots & x_{k,m} \end{pmatrix} \tag{2.1}$$

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

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

Algumas transformações devem ser feitas na matriz $\boldsymbol{X}$ 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

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

o qual corresponde as coordenadas complexas para os marcos.

Neste trabalho serão consideradas as coordenadas de Kendall. O primeiro trabalho feito nesta área foi Kendall (1977), mas somente em Kendall (1984) que realmente formalizou-se e definiu-se os conceitos básicos. Um dos pontos mais relevantes do trabalho de Kendall foi a proposta do sistema de coordenadas. Estes sistemas visam a obtenção da forma de um objeto por meio dos pontos dispostos através dos marcos. Primeiramente, deve-se retirar os efeitos de locação. Contudo, para remover a locação da forma deve-se definir a submatriz de Helmert ($\boldsymbol{H}$). A matriz de Helmert ($\boldsymbol{H}^F$) é uma matriz quadrática ortogonal $k \times k$ com a primeira linha de elementos igual a $1/\sqrt{k}$, então a j-ésima linha (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 matriz de Helmert sem a primeira linha. Por exemplo, para $k = 4$ a matriz de Helmert é

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

e a sub-matriz de Helmert será

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

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

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

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

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

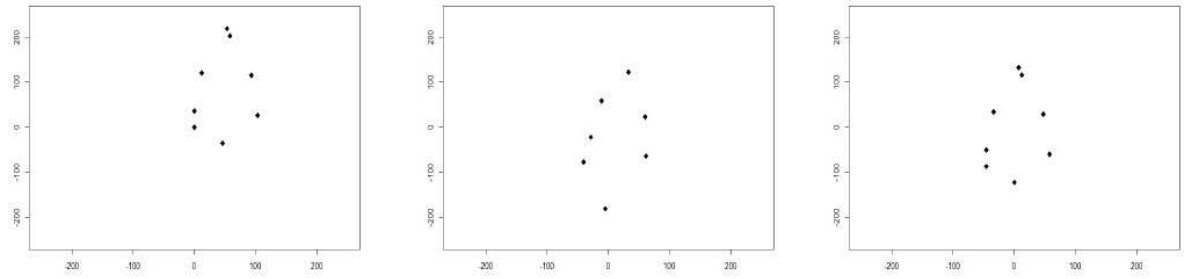
$$\boldsymbol{H}^\top\boldsymbol{H} = \left(\boldsymbol{I}_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^\top\right)$$

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

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

Para exemplificar, tem-se a configuração matemática de um indivíduo obtido dos dados de gorilas machos, (DRYDEN; MARDIA, 1998), em que $\boldsymbol{z}^0 = (53+220i, 46-35i, 0+0i, 0+37i, 12+122i, 58+204i, 93+117i, 103+28i)^\top$. A Figura 2.1 representa as configurações original, Helmertizada e centralizada para marcos de um gorila macho. Vale notar que a configuração Helmertizada, Figura 2.1b, perde a dimensão original dos dados. Este problema é corrigido com a multiplicação por $\boldsymbol{H}^\top$.

Figura 2.1: Configurações de um gorila macho.



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

Fonte: Oliveira (2016).

Para remover o efeito de escala deve-se dividir a configuração Helmertizada, obtida na expressão (2.4), pela sua norma. Sendo assim,

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

onde $\boldsymbol{w}^*$ é o transposto do conjugado de $\boldsymbol{w}$ e $|\cdot|$ denota a norma complexa de $\boldsymbol{w}$. O vetor $\boldsymbol{z}$, de acordo com Kendall (1984), é chamado de pré-forma da configuração complexa $\boldsymbol{z}^0$. 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 relevantes devem ser considerados.

Definição 2.1. *(Pré-forma). As pré-formas de uma matriz de configuração $\boldsymbol{X}$, da Equação* (2.1)*, é dado por*

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

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

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

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

O espaço das pré-formas é o espaço de todas as 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

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

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

Definição 2.2. *(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 espaço projetivo complexo $\mathbb{C}P^{k-2}$, o espaço de linhas complexas que passam pela origem.

Definição 2.3. *(Ícone). Um ícone é um membro particular do conjunto de formas $[\boldsymbol{z}]$ 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. A pré-forma centrada $\boldsymbol{z}_{c_e}$ é uma escolha apropriada de ícone. Dessa forma, iremos usar a pré-forma centrada para ter uma representação da configuração original.

2.1.2 Distância Procrustes para caso planar

O objetivo desta seção é apresentar 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 entre 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 $\boldsymbol{z}_1^0, \ldots, \boldsymbol{z}_n^0$ como uma amostra aleatória de configurações complexas de uma população de n objetos ou indivíduos o qual $\boldsymbol{z}_i^0$ foi definido pela Equação (2.3).

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

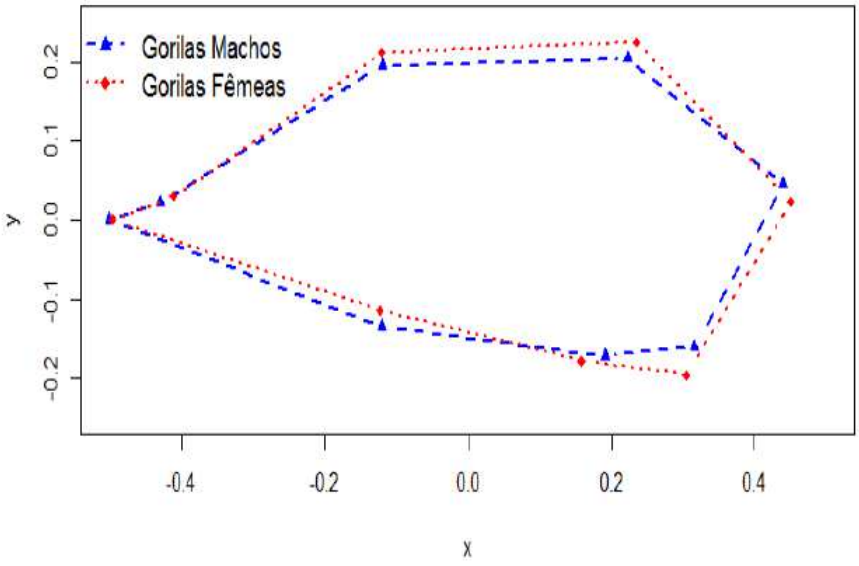
Resultado 2.1. *A* ***forma média Procrustes completa*** $\hat{\boldsymbol{\mu}}$ *pode ser encontrada como o autovetor correspondente ao maior autovalor da* ***soma quadrática complexa*** *e* ***matriz produto***

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

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 $\boldsymbol{S}$. O autovetor é único (até uma rotação - todas rotações de $\hat{\boldsymbol{\mu}}$ são também soluções, mas todos este correspondem a mesma forma), desde que exista um único autovalor maior de $\boldsymbol{S}$. A forma média de dados dos gorilas 29 machos e 30 fêmeas (DRYDEN; MARDIA, 1998) são apresentados na Figura 2.2.

Figura 2.2: A forma média Procrustes completa de gorilas machos e fêmeas.



Fonte: Oliveira (2016).

A configuração possui uma rotação arbitrária (DRYDEN; 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 que o **ajuste Procrustes completo** ou **coordenadas Procrustes completa** de $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ são

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

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

Os resíduos Procrustes são calculados como

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

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

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

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 $\boldsymbol{z}_x$ e $\boldsymbol{z}_y$ é

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

Essa quantidade também denominada como geodésica é definida como o caminho mais curto entre $\boldsymbol{z}_x$ e $\boldsymbol{z}_y$ na hiperesfera da pré-forma e não é afetada pela rotação (ZELDITCH; SWIDERSKI; SHEETS, 2012). Consequentemente, pode-se ver explicitamente que a **distância de Procrustes** ρ é o ângulo entre as pré-formas $\boldsymbol{z}_x$ e $\boldsymbol{z}_y$. Também, desde que o raio da esfera da pré-forma é 1, pode-se considerar ρ a distância ótima no círculo na esfera da pré-forma.

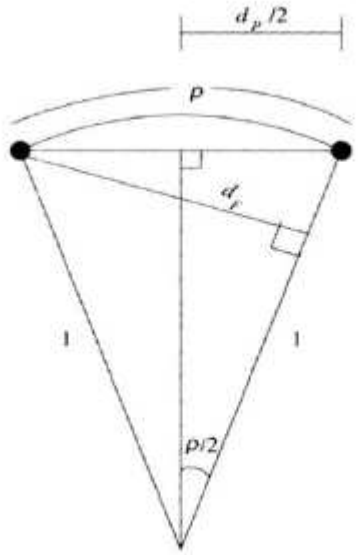
A **distância de Procrustes Parcial** também é invariante quanto a rotação entre $\boldsymbol{z}_x$ e $\boldsymbol{z}_y$ e é dada por

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

A Figura 2.3 representa as distâncias (2.12), (2.13) e (2.14) na hiperesfera da pré-

forma.

Figura 2.3: Ilustração da relação entre as distâncias d_F, ρ e d_P na esfera da pré-forma.



Fonte: Oliveira (2016).

2.1.3 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 podem ser usadas as técnicas padrão de análise multivariada. Existem vários tipos diferentes de coordenadas no espaço tangente. Serão consideradas as coordenadas tangente Procrustes parcial.

Considere $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ uma amostra de configurações. Dessa forma, as coordenadas tangentes serão

$$\boldsymbol{t}_i = \exp^{i\theta}[I_{k-1} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^*]\boldsymbol{z}_i, i = 1, \ldots, n \tag{2.15}$$

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

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

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

Assim, o vetor de pré-formas $\boldsymbol{z}_i \in \mathbb{C}^{k-1}$ é representado nas coordenadas tangentes pelo vetor $\boldsymbol{v}_i \in \mathbb{R}^{2k-2}$. A distância Euclidiana no espaço tangente para o espaço de formas é uma boa aproximação para situações de alta concentração, ou seja, variância pequena ((DRYDEN; MARDIA, 1998), p.76) das distâncias de 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 formas médias das pré-formas propostas por Amaral, Dryden e Wood (2007) e a região de confiança bootstrap para a forma média planar, desenvolvida por Amaral et al. (2010).

2.2 Algoritmo k-médias para formas

De acordo com Amaral et al. (2010), para dados usuais, o algoritmo k-médias objetiva 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 termo *k-means* foi usado primeiramente em 1967 por James MacQueen em seu artigo intitulado "*Some Methods for Classification and Analysis of Multivariate Observation*". No ano de 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 Llody Forgy pois em 1965 E. W. Fordy publicou o mesmo algoritmo. 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**).

$$J(C_r) = \sum_{i \in C_r} d^2(\boldsymbol{z}_i, \boldsymbol{\mu}_r)$$

onde a função d^2 é uma medida de distância geral como as distâncias definidas em (2.12), (2.13) e (2.14).

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

$$J(C) = \sum_{r=1}^{k} \sum_{i \in C_r} d^2(\boldsymbol{z}_i, \boldsymbol{\mu}_r) \tag{2.16}$$

O algoritmo a seguir resume o passo a passo iterativo para obter o agrupamento pelo método k-médias em análise de formas.

Algoritmo 1: Método de Agrupamento k-médias para formas planas

Entrada: Pré-forma $\boldsymbol{z}$ (como definido na Equação (2.6)), número de grupos k e alocação inicial;

Saída: 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.

Esse algoritmo move os objetos entre os agrupamentos até que não haja alteração significativa na função objetivo, ou até que o número de iterações máximo pré determinado seja alcançado. O resultado é um conjunto de grupos com indivíduos com características 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 e pontos aberrantes, por exemplo.

2.3 Funções kernel

Desde o início da última década, muitos pesquisadores têm demonstrado interesse em métodos baseados em kernel (FILIPPONE et al., 2008). A principal ideia por trás desses métodos é o uso de um mapeamento não-linear arbitrário Φ do espaço original das observações para um espaço de mais alta dimensão (possivelmente infinita), chamado espaço de características, $\mathcal{F}$. Nesta seção, apresentamos uma breve revisão acerca da teoria básica sobre funções kernel.

Seja $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ um conjunto não-vazio, onde $\boldsymbol{x}_i \in \mathbb{R}^p$, $\forall i$. Uma função $h : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ é dita um kernel positivo-definido (ou kernel de Mercer) se, e somente se, h é simétrica (isto é, $h(\boldsymbol{x}_i, \boldsymbol{x}_k) = h(\boldsymbol{x}_k, \boldsymbol{x}_i)$) e a seguinte desigualdade é válida (MERCER, 1909):

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_i c_k h(\boldsymbol{x}_i, \boldsymbol{x}_k) \geq 0 \quad \forall n \geq 2, \tag{2.17}$$

onde $c_r \in \mathbb{R}$ $\forall r = 1, \ldots, n$.

Um conjunto de observações não-linearmente separável pode tornar-se separável linearmente através de um mapeamento não-linear arbitrário para um espaço de características de alta dimensão (HAYKIN, 1998). Seja $\Phi : \boldsymbol{X} \to \mathcal{F}$ um mapeamento não-linear arbitrário do espaço original das observações para um espaço de características de alta dimensão $\mathcal{F}$. Aplicando o mapeamento não-linear Φ, o produto interno $\boldsymbol{x}_i^\top \boldsymbol{x}_k$ no espaço original é mapeado para $\Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_k)$ no espaço de características. A essência dos métodos baseados em kernel é que o mapeamento não-linear Φ não precisa ser explicitamente especificado porque todo kernel de Mercer pode ser expresso como

$$h(\boldsymbol{x}_i, \boldsymbol{x}_k) = \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_k), \tag{2.18}$$

que é usualmente referida como *kernel trick*.

Por causa da Equação (2.18), é possível calcular distâncias Euclidianas em $\mathcal{F}$ da seguinte maneira:

$$\begin{aligned} ||\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_k)||^2 &= (\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_k))^\top (\Phi(\boldsymbol{x}_i) - \Phi(\boldsymbol{x}_k)) \\ &= \Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_i) - 2\Phi(\boldsymbol{x}_i)^\top \Phi(\boldsymbol{x}_k) + \Phi(\boldsymbol{x}_k)^\top \Phi(\boldsymbol{x}_k) \\ &= h(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2h(\boldsymbol{x}_i, \boldsymbol{x}_k) + h(\boldsymbol{x}_k, \boldsymbol{x}_k). \end{aligned} \tag{2.19}$$

Exemplos de funções kernel tipicamente utilizadas são:

- Linear: $h(\boldsymbol{x}_i, \boldsymbol{x}_k) = \boldsymbol{x}_i^\top \boldsymbol{x}_k$,
- Polinomial de grau d: $h(\boldsymbol{x}_i, \boldsymbol{x}_k) = (\gamma \boldsymbol{x}_i^\top \boldsymbol{x}_k + \theta)^d$, $\gamma > 0$, $\theta \geq 0$, $d \in \mathbb{N}$,
- Gaussiana: $h(\boldsymbol{x}_i, \boldsymbol{x}_k) = e^{-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_k||^2}{2\sigma^2}}$, $\sigma > 0$,
- Laplaciana: $h(\boldsymbol{x}_i, \boldsymbol{x}_k) = e^{-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_k||}$, $\gamma > 0$,
- Sigmóide: $h(\boldsymbol{x}_i, \boldsymbol{x}_k) = \tanh(\gamma \boldsymbol{x}_i^\top \boldsymbol{x}_k + \theta)$, $\gamma > 0$, $\theta \geq 0$,

onde γ, σ, θ e d são parâmetros do kernel.

2.4 Agrupamento Espectral

Os métodos de agrupamento espectral relacionam-se com teoria dos grafos (SHAWE-TAYLOR; KANDOLA, 2001). Uma comparação de alguns métodos de agrupamento espectral foi apresentada por Verma e Meila (2003). Seja $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ o conjunto de obervações a serem agrupadas. Iniciando a partir de $\boldsymbol{X}$, podemos construir um grafo completo, ponderado não-direcionado $G(V, A)$ contendo um conjunto de nós $V = \{v_1, \ldots, v_n\}$ correspondendo às n observações e arestas definidas através da matriz de adjacência $\boldsymbol{A}$ (também chamada de afinidade), de dimensão $n \times n$. A matriz de adjacência para um grafo ponderado é dada pela matriz cujo elemento a_{ij} representa o peso da aresta que liga os nós i e j. Sendo um gráfico não-direcionado, a propriedade $a_{ij} = a_{ji}$ é verdadeira. A adjacência entre duas observações pode ser definida da seguinte forma:

$$a_{ij} = \begin{cases} h(\boldsymbol{x}_i, \boldsymbol{x}_j), & \text{se } i \neq j; \\ 0, & \text{caso contrário.} \end{cases} \tag{2.20}$$

onde a função $h(\boldsymbol{x}_i, \boldsymbol{x}_j)$ mede a similaridade entre duas observações i e j. Comumente, uma função kernel gaussiana é usada:

$$h(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2\sigma^2}\right), \tag{2.21}$$

onde $d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$ mede a dissimilaridade entre duas observações i e j e σ^2 controla a velocidade de deterioração de h. Esta particular escolha tem a propriedade de que $\boldsymbol{A}$ tem apenas alguns termos significativamente diferentes de 0, ou seja, $\boldsymbol{A}$ é esparsa.

A matriz de grau $\boldsymbol{D}$ é a matriz diagonal cujos elementos são os graus dos nós de G,

$$d_{ii} = \sum_{j=1}^{n} a_{ij}. \tag{2.22}$$

Neste contexto, o problema de agrupamento pode ser visto como um problema de corte (CHUNG, 1996), onde se deseja separar um conjunto de nós $S \subset V$ do conjunto complementar $\bar{S} = V \setminus S$. O problema de corte no grafo pode ser formulado de várias maneiras, dependendo da escolha da função objetivo a otimizar. Uma das mais populares funções para otimizar (CHUNG, 1996) é:

$$cut(S, \bar{S}) = \sum_{v_i \in S, v_j \in \bar{S}} a_{ij}. \tag{2.23}$$

É fácil verificar que a minimização da função objetivo, dada em (2.23) favorece partições

contendo os nós isolados. Para conseguir um melhor equilíbrio na cardinalidade de S e $\bar{S}$, sugere-se a otimização da função de corte normalizada (SHI; MALIK, 2000):

$$Ncut(S,\bar{S}) = cut(S,\bar{S})\left(\frac{1}{assoc(S,V)} + \frac{1}{assoc(\bar{S},V)}\right), \tag{2.24}$$

em que a associação $assoc(S,V)$ é também conhecido como o volume de S:

$$assoc(S,V) = \sum_{v_i \in S, v_j \in V} a_{ij} \equiv vol(S) = \sum_{v_i \in S} d_{ii}. \tag{2.25}$$

Há outras definições de funções para otimizar, por exemplo, a contundência (KANNAN; VEMPALA; VETTA, 2004), a associação normalizada (SHI; MALIK, 2000) e a razão de corte (DHILLON; GUAN; KULIS, 2004).

A complexidade em otimizar essas funções objetivo é muito elevada (por exemplo, a otimização do corte normalizado é *NP-Hard* (SHI; MALIK, 2000; WAGNER; WAGNER, 1993)) e por esta razão, tem-se procurado usar conceitos espectrais de análise de grafos. Estes conceitos podem ser formulados a partir da introdução da matriz Laplaciana (CHUNG, 1996):

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}, \tag{2.26}$$

que pode ser visto como um operador linear em G. Além desta definição de matriz Laplaciana, existem outras definições alternativas:

- Laplaciana Normalizada, dada por $\boldsymbol{L}_N = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}$;
- Laplaciana Generalizada, dada por $\boldsymbol{L}_G = \boldsymbol{D}^{-1}\boldsymbol{L}$;
- Laplaciana Relaxada, dada por $\boldsymbol{L}_\rho = \boldsymbol{L} - \rho\boldsymbol{D}$.

Cada definição é justificada por propriedades especiais desejáveis em um determinado contexto. A decomposição espectral da matriz Laplaciana pode fornecer informações úteis sobre as propriedades do grafo. Em particular, pode ser visto que o segundo menor autovalor de $\boldsymbol{L}$ está relacionado com o corte no grafo (FIEDLER, 1973) e os correspondentes autovetores podem agrupar observações semelhantes (CHUNG, 1996; SHI; MALIK, 2000; BRAND; HUANG, 2003).

O problema da redução de dimensionalidade tem como objetivo encontrar uma representação dimensional adequada de um conjunto de dados em um espaço dimensional elevado. Em Belkin e Niyogi (2003), cada nó no grafo, que representa uma observação, é conectado apenas com nós correspondentes a observações vizinhas e a decomposição

espectral da matriz Laplaciana do grafo obtido permite encontrar uma representação dimensional baixa de $\boldsymbol{X}$.

São inúmeros os algoritmos que foram desenvolvidos para resolver o problema de particionamento de dados por meio de métodos espectrais. Dentre eles, podem ser citados os algoritmo de Shi & Malik (SHI; MALIK, 2000), Ng, Jordan & Weiss, Perona & Freeman, dentre outros (WAGNER; WAGNER, 1993; FILIPPONE et al., 2008). Neste trabalho, focaremos na utilização do algoritmo espectral Ng, Jordan & Weiss, adaptado para agrupar dados de formas e comparando-o com a adaptação para dados de formas do algoritmo k-médias, proposta por Amaral et al. (2010).

2.4.1 Algoritmo de Ng, Jordan & Weiss

O algoritmo de Ng, Jordan & Weiss utiliza autovetores da matriz Laplaciana, além de, através do algoritmo k-médias, obter uma partição do conjunto de observações (MORAIS, 2012). É descrito abaixo, o roteiro do algoritmo para um conjunto de dados $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$

Algoritmo 2: Algoritmo Ng, Jordan & Weiss

Entrada: Conjunto de observações $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ e número de grupos k;

Saída: Grupos $C_r (1 \leq r \leq k)$;

1 Calcule matriz de adjacência $A = [a_{ij}] \in \mathbb{R}^{n \times n}$,

$$a_{ij} = \begin{cases} \exp\left(-\dfrac{d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2\sigma^2}\right), & \text{se } i \neq j; \\ 0, & \text{caso contrário,} \end{cases}$$

onde $d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$ representa a distância euclidiana quadrada entre duas observações i e j. Como um critério para a escolha de σ^2 diversos autores sugerem uma busca em um *grid*;

2 Calcule a matriz Laplaciana normalizada $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$;

3 Encontre os k autovetores de $\boldsymbol{L}$ associados aos seus k maiores autovalores;

4 Construa uma matriz $\boldsymbol{Z}$ concatenando os k autovetores associados aos k maiores autovalores de $\boldsymbol{L}$;

5 Obtenha a matriz $\boldsymbol{Y}$ através de $\boldsymbol{Z}$, aplicando $y_{ij} = z_{ij} / \sum_{r=1}^{k} z_{ir}^2$, fazendo com que todas as linhas de $\boldsymbol{Z}$ tenham norma unitária. Este procedimento mapeia os pontos originais para uma esfera unitária;

6 Usando $\boldsymbol{Y}$, obter os grupos utilizando o algoritmo k-médias clássico;

7 Atribuir o ponto original $\boldsymbol{x}_i$ ao grupo j se e só se a linha i da matriz $\boldsymbol{Y}$ pertencer ao grupo j.

Podendo ser vista como um grafo, a matriz de afinidade liga entre si as suas linhas e as colunas, representando as ligações entre os vértices de um grafo, em que o valor dos campos da matriz representa os próprios vértices.

2.4.2 Agrupamento espectral para dados de formas

No contexto de análise de dados de formas, um kernel de Mercer ou kernel positivo-definido é obtido quando substituímos a distância euclidiana no kernel gaussiano pela distância de procrustes completa, resultando no então chamado kernel gaussiano procrustes (JAYASUMANA et al., 2013).

A função kernel gaussiana procrustes é

$$h(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp\left(-\frac{d_F^2(\boldsymbol{z}_i, \boldsymbol{z}_j)}{2\sigma^2}\right) = \exp\left(-\frac{1 - |(\boldsymbol{z}_i^* \boldsymbol{z}_j)|^2}{2\sigma^2}\right) \tag{2.27}$$

em que d_F é a distância de Procrustes Completa expressa na Equação (2.12) e é um kernel

positivo definido para todo $\sigma^2 \in \mathbb{R}^+$.

A proposta do nosso trabalho consiste na adaptação do cálculo da matriz de adjacência no algoritmo de Ng, Jordan & Weiss com base na função kernel, usando a distância de Procrustes Completa. Além disso, o algoritmo de agrupamento espectral de Ng, Jordan & Weiss, usualmente utilizado em dados clássicos, foi utilizado para análise de dados de formas de objetos projetados no espaço tangente.

CAPÍTULO 3

APLICAÇÕES

3.1 Introdução

Neste capítulo serão apresentadas a descrição de diversos conjuntos de dados de formas extraídos dos pacotes *shapes*, *geomorph* e *momocs*, do software estatístico `R`, junto a uma comparação do método proposto (Ng, Jordan & Weiss para dados de formas planas) com o algoritmo k-médias para formas proposto por Amaral et al. (2010). A Tabela 3.1 apresenta os conjuntos de dados utilizados no estudo, suas dimensões e o número de grupos, a priori, de cada conjunto de dados.

Tabela 3.1: Conjunto de dados utilizados no estudo, formato dos dados e número de grupos de cada conjunto.

Dados	Formato	Número de grupos
Cérebros de esquizofrênicos e não-esquizofrênicos	8×2×30	2
Vértebra de camundongos	6×2×76	3
Crânios de gorilas	8×2×59	3
Crânios de macacos	7×3×18	2
Crânios de grandes primatas	8×2×167	3
Cérebros de adultos saudáveis	24×3×58	2
Crânios de chimpanzés	8×2×54	2
Crânios de orangotangos	8×2×60	2
Grãos de areia	50×2×49	2
Cabeças de salamandras	12×2×40	2
Asas de mosquito	18×2×40	2
Caudas de salamandras larvais	17×2×64	6
Glumas de cereais	21×2×172	3
Corações desenhados a mão	8×2×240	8

Foi realizada uma busca em *grid* para se encontrar o valor ótimo do parâmetro σ^2 do kernel Gaussiano, em cada aplicação, e os melhores valores para esses parâmetros estão apresentados na Tabela 3.2. A definição do *grid* para a busca foi feita de forma empírica para cada conjunto de dados. Por exemplo, para o conjunto de dados de cérebros de esquizofrênicos e de nao-esquizofrênicos, variamos o valor de σ^2 entre 0.001 e 1 por 0.001 para os dois métodos, enquanto que para os dados de crânios de chimpanzés, o melhor valor de $sigam^2$ foi encontrado no intervalo entre 0.1 e 1 por 0.1 para o algoritmo Ng, Jordan & Weiss com a distância de procrustes completa e entre 1 e 10 por 0.1 para o algoritmo Ng, Jordan & Weiss no espaço tangente das formas.

Tabela 3.2: Parâmetros σ^2 do kernel Gaussiano aplicados à cada conjuntos de dados por meio do algoritmo Ng, Jordan & Weiss.

Dados	**Parâmetros da Função kernel**	
	Ng, Jordan & Weiss Procrustes Completo	Ng, Jordan & Weiss Euclid. esp. tang.
Cérebros de esquiz. e não-esquiz.	0.001	0.001
Vértebra de camundongos	0.0001	5
Crânios de gorilas	0.1	0.1
Crânios de macacos	0.001	0.001
Crânios de grandes primatas	0.25	0.25
Cérebros de adultos saudáveis	0.25	0.25
Crânios de chimpanzés	0.5	10
Crânios de orangotangos	0.5	10
Grãos de areia	1	1
Cabeças de salamandras	0.01	0.1
Asas de mosquito	0.1	1
Caudas de salamandras larvais	0.0001	0.0001
Glumas de cereais	1	0.001
Corações desenhados a mão	0.0001	0.0001

3.1.1 Índices de avaliação

Para comparar os métodos de agrupamento considerados neste trabalho, utilizamos o Índice de Rand Ajustado (IRA) (HUBERT; ARABIE, 1985) e a taxa total de erro de alocação (TEA) (BREIMAN et al., 1984).

Seja $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_i, \ldots, \mathcal{P}_c\}$ a partição *a priori* de $\Omega = \{1, \ldots, n\}$ em c classes e seja $P = \{P_1, \ldots, P_k, \ldots, P_K\}$ uma partição rígida de $\Omega = \{1, \ldots, n\}$ em K grupos fornecidos por um algoritmo de agrupamento. As quantidades n_{ik}, $i = 1, \ldots, c$, $k = 1, \ldots, K$, representam o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k e podem ser representadas na forma da Tabela 3.3, denominada matriz de confusão.

Tabela 3.3: Matriz de confusão.

Classes	Grupos P_1	$\cdots$	P_k	$\cdots$	P_K	$\sum$
$\mathcal{P}_1$	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1K}	$n_{1\bullet} = \sum_{k=1}^{K} n_{1k}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\mathcal{P}_i$	n_{i1}	$\cdots$	n_{ik}	$\cdots$	n_{iK}	$n_{i\bullet} = \sum_{k=1}^{K} n_{ik}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\mathcal{P}_c$	n_{c1}	$\cdots$	n_{ck}	$\cdots$	n_{cK}	$n_{c\bullet} = \sum_{k=1}^{K} n_{ck}$
$\sum$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$\cdots$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$\cdots$	$n_{\bullet 1} = \sum_{i=1}^{c} n_{i1}$	$n = \sum_{i=1}^{c} \sum_{k=1}^{K} n_{ik}$

O Índice de Rand Ajustado (IRA) é obtido como

$$CR = \frac{\sum_{i=1}^{c}\sum_{k=1}^{K}\binom{n_{ik}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{c}\binom{n_{i\bullet}}{2}\sum_{k=1}^{K}\binom{n_{\bullet k}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{c}\binom{n_{i\bullet}}{2} + \sum_{k=1}^{K}\binom{n_{\bullet k}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{c}\binom{n_{i\bullet}}{2}\sum_{k=1}^{K}\binom{n_{\bullet k}}{2}}, \tag{3.1}$$

onde $\binom{n}{2} = \frac{n(n-1)}{2}$, n_{ik} representa o número de observações que estão na classe $\mathcal{P}_i$ e no grupo P_k, $n_{i\bullet}$ representa o número de observações na classe $\mathcal{P}_i$, $n_{\bullet k}$ representa o número de observações no grupo P_k, e n é o número total de observações no conjunto de dados.

O IRA avalia o grau de concordância (similaridade) entre uma partição *a priori* e uma partição fornecida por um método de agrupamento. Além disso, o IRA não é sensível ao número de classes nas partições ou à distribuição das observações nos grupos. Finalmente, o IRA assume valores no intervalo $[-1, 1]$, no qual o valor 1 indica concordância perfeita entre as partições, enquanto que valores próximos de zero ou negativos correspondem a concordância entre partições encontrada ao acaso Milligan (1996).

Em problemas de classificação, cada grupo P_k é associado a uma classe *a priori* $\mathcal{P}_i$ e esta associação deve ser interpretada como se a verdadeira classe *a priori* fosse $\mathcal{P}_i$. Dessa forma, para uma observação pertencente a um dado grupo P_k a decisão está correta se a classe *a priori* dessa observação é $\mathcal{P}_i$. Para obter uma taxa de erro de classificação mínima, precisamos encontrar uma regra de decisão que minimize a probabilidade de erro.

Seja $\ell(\mathcal{P}_i, P_k)$ a probabilidade *a posteriori* de que uma observação pertença à classe $\mathcal{P}_i$ quando associado ao grupo P_k. Seja $\ell(P_k)$ a probabilidade de que a observação pertença ao grupo P_k. A função ℓ é conhecida como função de verossimilhança.

A estimativa da máxima probabilidade *a posteriori* é a moda da probabilidade *a*

posteriori $\ell(\mathcal{P}_i, P_k)$ e o índice da classe *a priori* associada a esta moda é dada por

$$MAP(P_k) = \arg\max_{1\leq i\leq c} \ell(\mathcal{P}_i, P_k).$$

A regra de decisão de Bayes que minimiza a probabilidade média de erro é selecionar a classe *a priori* que maximiza a probabilidade *a posteriori*. A taxa de erro de alocação do grupo P_k é igual a $1-\ell(\mathcal{P}_{MAP(P_k)}/P_k)$ e a taxa total de erro de alocação (TEA) é igual a

$$TEA = \sum_{k=1}^{K} \ell(P_k)(1-\ell(\mathcal{P}_{MAP(P_k)}/P_k)).$$

Para uma amostra,

$$\ell(\mathcal{P}_{MAP(P_k)}/P_k) = \max_{1\leq i\leq c} n_{ik}/n_{\bullet k}.$$

A taxa total de erro de alocação (TEA) foi concebida de modo a medir a habilidade de um algoritmo de agrupamento encontrar as classes *a priori* presentes em um conjunto de dados e é calculada da forma:

$$TEA = \sum_{k=1}^{K} \frac{n_{\bullet k}}{n}\left(1-\max_{1\leq i\leq c} n_{ik}/n_{\bullet k}\right) = 1-\frac{\sum_{k=1}^{K}\max_{1\leq i\leq c} n_{ik}}{n}. \tag{3.2}$$

O índice TEA assume valores no intervalo $[0,1]$, no qual valores próximos de zero indicam maior habilidade de um algoritmo na detecção de classes *a priori*.

3.2 Aplicações

Esta seção compreende na comparação dos algoritmos k-médias (4) e Ng, Jordan & Weiss (7) aplicados a conjuntos de dados de formas. Os resultados para 14 conjunto de dados foram apresentados em tabelas, considerando os índices de avaliação IRA e TEA, definidos outrora, para as distâncias de Procrustes, Procrustes Parcial, Procrustes Completa e distância euclidiana no espaço tangente.

Na Tabela 3.4, dispomos um resumo acerca do desempenho dos métodos comparados considerando as possíveis distâncias que os algoritmos permitem utilizar. A partir dos resultados dos dois índices de avaliação, destacamos, com o círculo hachurado ($\bullet$), o método que apresentou valores superiores em relação aos demais. Os métodos (e distâncias) considerados são k-médias de procrustes completo (k_{FP}), k-médias de procrustes parcial (k_{PP}), k-médias de procrustes (k_P), k-médias clássico no espaço tangente das formas (k_{ET}), Ng,

Jordan & Weiss de procrustes completo (Ng_{FP}) e Ng, Jordan & Weiss clássico no espaço tangente das formas (Ng_{ET}).

Tabela 3.4: Resumo dos resultados das aplicações a partir dos índices de avaliação para cada distância considerada aos métodos utilizados neste trabalho.

Dados/Métodos	k_{FP}	k_{PP}	k_P	k_{ET}	Ng_{FP}	Ng_{ET}
Cérebros esquiz. e não-esqui.	○	○	○	○	●	●
Vértebra de camundongos	○	○	○	○	●	○
Crânios de gorilas	○	○	○	○	●	○
Crânios de macacos	○	○	○	○	●	●
Crânios de grandes primatas	○	○	○	○	●	●
Cérebros de adultos saudáveis	○	○	○	○	○	○
Crânios de chimpanzés	○	○	○	○	●	○
Crânios de orangotangos	●	●	●	○	○	○
Grãos de areia	○	○	○	●	○	●
Cabeças de salamandras	○	○	○	●	●	○
Asas de mosquito	○	○	○	○	●	●
Caudas de salamandras larvais	○	○	○	○	●	●
Glumas de cereais	○	○	○	○	●	○
Corações desenhados a mão	○	○	○	○	●	●

A partir destes resultados verifica-se que a adaptação ao algoritmo de Ng, Jordan & Weiss, proposta neste trabalho, findou em melhores resultados, na maioria dos conjuntos de dados considerados. Em apenas uma situação o algoritmo adaptado apresentou resultados inferiores ao algoritmo do método k-médias (Dados: Crânios de orangotangos) e em outras duas situações, ambos os algoritmos, em alguma distância, resultaram em melhores valores. Nota-se também um predomínio de melhores resultados para o algoritmo adaptado de Ng, Jordan & Weiss considerando a distância de procrustes completa, o que, em geral, evidencia a eficiência deste algoritmo adaptado para dados de formas planas de objetos.

Nas subseções, à seguir, serão apresentados valores dos índices de avaliação para cada conjunto de dados considerado, bem como destacamos o método com melhores resultados.

3.2.1 Cérebros de esquizofrênicos e não-esquizofrênicos

Esse conjunto de dados corresponde a uma amostra de 28 configurações contendo 13 marcos anatômicos extraídos de imagens de ressonância magnética de cérebros de 14 pacientes saudáveis e 14 pacientes com esquizofrenia. A Tabela 3.5 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Pode-se observar que o algoritmo espectral de

Ng, Jordan & Weiss obteve resultados superiores ao método k-médias tanto utilizando a distância de procrustes completa quanto a distância euclidiana no espaço tangente.

Tabela 3.5: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de esquizofrênicos e não-esquizofrênicos.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,036**	**0,857**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	**0,036**	**0,857**
k-médias (Procrustes Completa)	0,429	−0,016
k-médias (Parcial Procrustes)	0,429	−0,016
k-médias (Procrustes)	0,429	−0,016
k-médias (Eucl. Esp. Tang.)	0,429	−0,016

3.2.2 Vértebras de camundongos

Num experimento para avaliar os efeitos da seleção para o peso corporal na forma de vértebras de camundongos, foram obtidos três grupos de camundongos: Controle, Grande e Pequeno. O grupo Controle contém camundongos não selecionados, o grupo Grande contém camundongos selecionados a cada geração de acordo com o maior peso corporal e o grupo Pequeno foi selecionado para peso corporal menor. Os ossos fazem parte de um estudo muito maior e esses ossos são da replica E do estudo ((TRUSLOVE, 1976), (BARGER-LUX et al., 1995, 1988), (MARDIA; DRYDEN, 1989)). Consideramos a segunda vértebra torácica T2. Existem 30 ossos controles, 23 ossos grandes e 23 ossos pequenos. O objetivo é avaliar se existe uma diferença de tamanho e forma entre os três grupos e fornecer descrições de quaisquer diferenças. Cada vértebra foi colocada sob um microscópio e digitaliza usando uma câmera de vídeo para dar uma imagem de nível de cinza. O esboço do osso é então extraído utilizando técnicas de processamento de imagem padrão para dar uma corrente de cerca de 300 coordenadas em torno da ferramenta.

A Tabela 3.6 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Onde se observa que o algoritmo espectral de Ng, Jordan & Weiss obteve resultados superiores ao método k-médias considerando apenas a utilização da distância de procrustes completa. O método k-médias, considerando-se as 4 distâncias, resultaram em estimativas melhores, comparado ao algoritmo Ng, Jordan & Weiss com distância euclidiana no espaço tangente.

Tabela 3.6: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de vértebras T2 de camundongos.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,316**	**0,373**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	0,368	0,206
k-médias (Procrustes Completa)	0,329	0,268
k-médias (Parcial Procrustes)	0,329	0,268
k-médias (Procrustes)	0,329	0,268
k-médias (Eucl. Esp. Tang.)	0,329	0,268

3.2.3 Crânios de gorilas

Trata-se de uma investigação para avaliar as diferenças cranianas entre os sexos de gorilas, onde 29 crânios de gorilas adultos machos e 30 crânios de gorilas adultos fêmea, foram tomados. Os marcos são marcos anatômicos e foram localizados por um biólogo especialista. É interessante avaliar se existe uma diferença de tamanho entre os sexos e se existem diferenças de forma entre os sexos nas regiões cerebrais. Um biólogo também estaria interessado em descrições geométricas da diferença de forma, e como a forma se relaciona com o tamanho e outras covariáveis. Apenas as dimensões X e Y foram utilizadas.

A Tabela 3.7 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Onde se observa que o algoritmo espectral de Ng, Jordan & Weiss obteve resultados superiores ao método k-médias considerando apenas a utilização da distância de procrustes completa. O algoritmo Ng, Jordan & Weiss, considerando a distância euclidiana no espaço tangente, apresentou as mesmas estimativas do método k-médias, considerando-se as 4 distâncias.

Tabela 3.7: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de crânios de gorilas.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes completa)	**0,068**	**0,743**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	0,085	0,684
k-médias (Procrustes completa)	0,085	0,684
k-médias (Parcial Procrustes)	0,085	0,684
k-médias (Procrustes)	0,085	0,684
k-médias (Eucl. Esp. Tang.)	0,085	0,684

3.2.4 Crânios de macacos

Busca-se investigar sobre as diferenças no sexo a partir de registros cranianos de uma espécie de macaco *Macaca fascicularis*. Foram obtidas amostras aleatórias de 9 crânios de machos e 9 crânios de fêmeas. Um subconjunto de sete marcos anatômicos foi localizado em cada crânio e as coordenadas em três dimensões de cada ponto foram registradas. A análise utilizando os métodos para dados de formas planas, considerou as dimensões **X** e **Y** do conjunto de dados.

A Tabela 3.8 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Pode-se observar que o algoritmo espectral de Ng, Jordan & Weiss obteve resultados superiores ao método k-médias tanto utilizando a distância de procrustes completa quanto a distância euclidiana no espaço tangente.

Tabela 3.8: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de crânios de macacos.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,167**	**0,412**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	**0,222**	**0,268**
k-médias (Procrustes Completa)	0,389	−0,003
k-médias (Parcial Procrustes)	0,389	−0,003
k-médias (Procrustes)	0,389	−0,003
k-médias (Eucl. Esp. Tang.)	0,333	0,055

3.2.5 Crânios de grandes primatas

Esta aplicação refere-se a um conjunto de dados contendo marcos do crânio de 167 grandes primatas. Neste conjunto são investigados 30 gorilas fêmeas e 29 gorilas machos, 26 chimpanzés fêmeas e 28 chimpanzés machos, 24 orangotangos fêmeas e 30 orangotangos machos. Os grupos, a priori, são as espécies, independentemente do sexo. A Tabela 3.9 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Pode-se observar que o algoritmo espectral de Ng, Jordan & Weiss , considerando as duas distâncias calculadas, obteve resultados superiores ao método k-médias.

Tabela 3.9: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de crânios de grandes primatas.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,030**	**0,912**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	**0,012**	**0,964**
k-médias (Procrustes Completa)	0,042	0,877
k-médias (Parcial Procrustes)	0,042	0,877
k-médias (Procrustes)	0,042	0,877
k-médias (Eucl. Esp. Tang.)	0,042	0,877

3.2.6 Cérebros de adultos saudáveis

Este conjuntos considera 24 marcos localizados em 58 cérebros adultos saudáveis, em que os grupos são definidos pelo sexo. A análise utilizando os métodos para dados de formas planas, considerou as dimensões $\mathbf{X}$ e $\mathbf{Y}$ do conjunto de dados. A Tabela 3.10 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Nesta aplicação, observa-se que as estimativas para ambos os algoritmos apresentaram valores equivalentes, exceto no método k-médias, considerando a distância euclidiana no espaço tangente, onde as taxas foram ainda menores que as demais verificadas.

Tabela 3.10: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de cérebros de adultos saudáveis.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	0,327	0,103
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	0,327	0,103
k-médias (Procrustes Completa)	0,327	0,103
k-médias (Parcial Procrustes)	0,327	0,103
k-médias (Procrustes)	0,327	0,103
k-médias (Eucl. Esp. Tang.)	0,396	0,027

3.2.7 Crânios de chimpanzés

São considerados 8 marcos em duas dimensões para análise do crânio de 54 chimpanzés, sendo 26 crânios de fêmeas e 28 crânios de machos. A Tabela 3.11 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Observa-se que o algoritmo Ng, Jordan

& Weiss, considerando apenas a utilização da distância de procrustes completa, apresentou melhores estimativas comparado ao método k-médias, que culminou em resultados aproximadamente iguais ao algoritmo Ng, Jordan & Weiss , considerando a distância euclidiana no espaço tangente.

Tabela 3.11: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de crânios de chimpanzés.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,278**	**0,182**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	0,352	0,070
k-médias (Procrustes Completa)	0,352	0,071
k-médias (Parcial Procrustes)	0,352	0,071
k-médias (Procrustes)	0,352	0,071
k-médias (Eucl. Esp. Tang.)	0,352	0,071

3.2.8 Crânios de orangotangos

Nesta análise, são considerados 8 marcos em duas dimensões de 60 orangotangos, sendo 30 fêmeas e 30 machos. A Tabela 3.12 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Nesta análise, os resultados apresentados para o algoritmo k-médias, para as 4 distâncias, foram superiores aos resultados do algoritmo Ng, Jordan & Weiss .

Tabela 3.12: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de crânios de orangotangos.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	0,148	0,486
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	0,130	0,540
k-médias (Procrustes Completa)	**0,092**	**0,657**
k-médias (Parcial Procrustes)	**0,092**	**0,657**
k-médias (Procrustes)	**0,092**	**0,657**
k-médias (Eucl. Esp. Tang.)	0,111	0,597

3.2.9 Grãos de areia

Trata-se de 50 pontos, sendo 24 em areia de mar e 25 perfis de grãos de areia de rio em 2 dimensões. Os dados originais foram fornecidos pelo Professor Dietrich Stoyan ((STOYAN; STOYAN, 1994), (BEKLEMISHEV; STOYAN, 1997)). Os 50 pontos em

cada esboço foram extraídos a comprimentos de arco aproximadamente iguais pelo método descrito em York et al. (2000). A Tabela 3.13 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Nota-se que os resultados apresentados pelo algoritmo Ng, Jordan & Weiss, foram superiores aos resultados dispostos por meio do algoritmo k-médias, apenas considerando a distância euclidiana no espaço tangente. As demais distâncias calculadas por meio do algoritmo k-médias apresentaram resultados aproximadamente iguais as estimativas obtidas por meio do algoritmo Ng, Jordan & Weiss considerando a distância de procrustes completa.

Tabela 3.13: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de grãos de areia.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	0,326	0,102
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	**0,265**	**0,204**
k-médias (Procrustes Completa)	0,326	0,103
k-médias (Parcial Procrustes)	0,326	0,103
k-médias (Procrustes)	0,326	0,103
k-médias (Eucl. Esp. Tang.)	**0,286**	**0,166**

3.2.10 Cabeças de salamandras

Este conjunto considera cabeças de salamandras com 12 pontos de referência em 2 dimensões para 40 indivíduos. A Tabela 3.14 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Os resultados apontam que o algoritmo de Ng, Jordan & Weiss, considerando a distância de procrustes completa e o algoritmo k-médias, considerando a distância euclidiana no espaço tangente, apresentaram estimativas equivalentes, e estas foram superiores as estimativas de qualquer outra distância, independente do algoritmo investigado.

Tabela 3.14: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de cabeças de salamandras.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,250**	**0,235**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	0,425	−0,026
k-médias (Procrustes Completa)	0,500	−0,026
k-médias (Parcial Procrustes)	0,500	−0,026
k-médias (Procrustes)	0,500	−0,026
k-médias (Eucl. Esp. Tang.)	**0,250**	**0,235**

3.2.11 Asas de mosquito

Considera-se 18 pontos de referência em 2 dimensões para as asas de 40 mosquitos. A Tabela 3.15 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Verifica-se que o algoritmo Ng, Jordan & Weiss apresentou resultados superiores às estimativas do algoritmo k-médias, considerando qualquer distância analisada.

Tabela 3.15: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de asas de mosquito.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,425**	**−0,003**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	**0,450**	**−0,009**
k-médias (Procrustes Completa)	0,475	−0,024
k-médias (Parcial Procrustes)	0,475	−0,024
k-médias (Procrustes)	0,475	−0,024
k-médias (Eucl. Esp. Tang.)	0,475	−0,024

3.2.12 Caudas de salamandras larvais

Trata-se de dados de referência das caudas de salamandras larvais expostas a diferentes tratamentos de herbicidas com 17 pontos de referência em 2 dimensões em 64 salamandras. A Tabela 3.16 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Os resultados sugerem que o algoritmo Ng, Jordan & Weiss tem melhores resultados, comparado ao algoritmo k-médias para quaisquer distância utilizada por este algoritmo.

Tabela 3.16: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de caudas de salamandras larvais.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,500**	**0,227**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	**0,456**	**0,283**
k-médias (Procrustes Completa)	0,596	0,087
k-médias (Parcial Procrustes)	0,596	0,087
k-médias (Procrustes)	0,596	0,087
k-médias (Eucl. Esp. Tang.)	0,553	0,109

3.2.13 Glumas de cereais

O conjunto em questão refere-se a coordenadas de marcos e semimarcos em glumas de cereais com 21 pontos de referência em 2 dimensões para 172 amostras. A Tabela 3.17 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado(IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Considerando a distância de procrustes completa, para o algoritmo Ng, Jordan & Weiss, observa-se resultados superiores comparados aos resultados dispostos pelo algoritmo k-means que, por sua vez, apresentou estimativas superiores ao algoritmo Ng, Jordan & Weiss, considerando a distância euclidiana no espaço tangente.

Tabela 3.17: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de glumas de cereais.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,048**	**0,884**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	0,333	0,085
k-médias (Procrustes Completa)	0,143	0,646
k-médias (Parcial Procrustes)	0,143	0,646
k-médias (Procrustes)	0,143	0,646
k-médias (Eucl. Esp. Tang.)	0,143	0,646

3.2.14 Corações desenhados a mão

Com esses dados, tem-se um esquema de coordenadas de corações desenhados a mão com 4 pontos de referência com 240 indivíduos. A Tabela 3.18 apresenta a Taxa de Erro de Alocação (TEA) e o Índice de Rand Ajustado (IRA) obtidos pelos métodos considerados aplicados a este conjunto de dados. Os números apontam que o algoritmo Ng, Jordan & Weiss, considerando as duas distâncias utilizadas, apresenta resultados superiores aos resultados obtidos por meio do algoritmo k-médias.

Tabela 3.18: Taxa de Erro de Alocação (TEA) e Índice de Rand Ajustado (IRA) obtidos pelos métodos de agrupamento considerados aplicados aos dados de corações desenhados a mão.

Método	TEA	IRA
Ng, Jordan & Weiss (Procrustes Completa)	**0,483**	**0,391**
Ng, Jordan & Weiss (Eucl. Esp. Tang.)	**0,417**	**0,428**
k-médias (Procrustes Completa)	0,562	0,170
k-médias (Parcial Procrustes)	0,558	0,180
k-médias (Procrustes)	0,558	0,180
k-médias (Eucl. Esp. Tang.)	0,546	0,183

CAPÍTULO 4

CONSIDERAÇÕES FINAIS

4.1 Conclusões

Apresentamos adaptações do algoritmos de agrupamento espectral de Ng, Jordan & Weiss para dados de formas (7) através do uso de um kernel gaussiano apropriado e para a projeção de formas planas no espaço tangente. Os métodos apresentados foram comparados com a versão do algoritmo k-médias para formas (4) proposto por Amaral et al. (2010).

A comparação dos métodos de agrupamento foi em 14 conjuntos de dados reais e, em geral, findou em melhores resultados para o agrupamento de dados de formas realizado por meio do algoritmo adaptado de Ng, Jordan & Weiss, significando que nos dados analisados, este algoritmo apresentou um desempenho superior ao algoritmo k-médias, na atribuição de obter grupos, de acordo com os Índices de Rand Ajustado (IRA) e as Taxas de Erro de Alocação (TEA).

A adaptação do algoritmo espectral de Ng, Jordan & Weiss para dados de formas, nestas aplicações, mostrou-se eficiente segundo os valores obtidos dos índices IRA e TEA, em sua maioria. Observando-se os dados de formas de *vértebras de camundongos*, *crânios de gorilas*, *crânios de chimpanzés*, *cabeças de salamandras* e *glumas de cereais*, o algoritmo adaptado (7), apresentou melhores resultados considerando-se a distância de procrustes completa. Já os dados de formas de *cérebros de esquizofrênicos e não esquizofrênicos*, *crânio de macacos*, *crânio de grandes primatas*, *asas de mosquito*, *caudas de salamandras larvais* e *corações desenhados a mão*, resultaram em melhores índices de avaliação IRA

e TEA, ao algoritmo adaptado, considerando-se as distâncias de procrustes completa e euclidiana no espaço tangente, inferindo na eficiência da versão adaptada do algoritmo Ng, Jordan & Weiss, ante a comparação com o usual método k-médias para dados de formas.

Ambos os algoritmos apresentaram valores iguais para os índices IRA e TEA na análise dos dados de formas de *cérebros de adultos saudáveis*. Dentre as 14 aplicações, a única que desfavoreceu o algoritmo adaptado de Ng, Jordan & Weiss, com valores inferiores dos índices IRA e TEA, em relação ao algoritmo k-médias, foi a análise dos dados de formas de *crânios de orangotangos*.

Portanto, diante da superioridade nos valores dos índices, apresentada pelo algoritmo espectral de Ng, Jordan & Weiss, adaptado para dados de formas, em comparação com o algoritmo k-médias para dados de formas, nas aplicações utilizadas neste trabalho, há evidências de que a adaptação proposta é eficiente para dados dessa natureza. Entretanto, apesar de o algoritmo adaptado de Ng, Jordan & Weiss ter sido superior, em comparação ao algoritmo k-médias, na maioria dos conjuntos de dados, alguns resultados não foram ideais quanto a qualidade do agrupamento, o que abre margem para proposição de novos métodos.

4.2 Trabalhos futuros

- Avaliação do método proposto através de um esquema de Monte Carlo com diferentes cenários de formas simuladas;
- Adaptação de outros algoritmos de agrupamento espectral para dados de formas;
- Comparação com outros métodos de agrupamento para formas, como, por exemplo, o método kernel k-médias;

REFERÊNCIAS BIBLIOGRÁFICAS

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

AMARAL, G. J. A. et al. k-means algorithm in statistical shape analysis. *Communications in Statistics - Simulation and Computation*, v. 39, n. 5, p. 1016–1026, 2010. Disponível em: <http://dx.doi.org/10.1080/03610911003765777>.

BARGER-LUX, M. et al. Vitamin d receptor gene polymorphism, bone mass, body size, and vitamin d receptor density. *Calcified tissue international*, Springer, v. 57, n. 2, p. 161–162, 1995, 1988.

BEKLEMISHEV, M.; STOYAN, T. Sorption–catalytic determination of manganese directly on a paper-based chelating sorbent. *Analyst*, Royal Society of Chemistry, v. 122, n. 10, p. 1161–1166, 1997.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, MIT Press, v. 15, n. 6, p. 1373–1396, 2003.

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.

BRAND, M.; HUANG, K. A unifying theorem for spectral embedding and clustering. In: *AISTATS*. [S.l.: s.n.], 2003.

BREIMAN, L. et al. *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC, 1984.

CHUNG, F. R. Spectral graph theory (cbms regional conference series in mathematics, no. 92). American Mathematical Society, 1996.

COSTA, L. d. F. D.; JR, R. M. C. *Shape analysis and classification: theory and practice*. [S.l.]: CRC Press, Inc., 2000.

DHILLON, I. S.; GUAN, Y.; KULIS, B. *A unified view of kernel k-means, spectral clustering and graph cuts*. [S.l.]: Citeseer, 2004.

DRYDEN, I.; MARDIA, K. *Statistical analysis of shape*. [S.l.]: Wiley, 1998.

FIEDLER, M. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, Institute of Mathematics, Academy of Sciences of the Czech Republic, v. 23, n. 2, p. 298–305, 1973.

FILIPPONE, M. et al. A survey of kernel and spectral methods for clustering. *Pattern recognition*, Elsevier, v. 41, n. 1, p. 176–190, 2008.

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.

GORDON, A. D. Classification. *Monographs on statistics and applied probability*, Chapman & Hall, v. 82, 1999.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, JSTOR, v. 28, n. 1, p. 100–108, 1979.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. ISBN 0132733501.

HÖPPNER, F. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. [S.l.]: John Wiley & Sons, 1999.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, p. 193–218, 1985.

JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM computing surveys (CSUR)*, Acm, v. 31, n. 3, p. 264–323, 1999.

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.

KANNAN, R.; VEMPALA, S.; VETTA, A. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, ACM, v. 51, n. 3, p. 497–515, 2004.

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.

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

LUXBURG, U. von. A tutorial on spectral clustering. *Statistics and Computing*, v. 17, n. 4, p. 395–416, 2007. ISSN 1573-1375. Disponível em: <http://dx.doi.org/10.1007/s11222-007-9033-z>.

MARDIA, K.; DRYDEN, I. The statistical analysis of shape data. *Biometrika*, JSTOR, p. 271–281, 1989.

MARE, R. A. D.; CORSEUIL, E. Morfometria de papilioninae (lepidopter pilioninae (lepidopter pilioninae (lepidoptera, papilionidae) pilionidae) ocorrentes em quatr entes em quatr entes em quatro localidades do rio gr o localidades do rio gr o localidades do rio grande do sul, ande do sul, ande do sul, brasil. iii. análise da f análise da f análise da forma das asas atr ma das asas atr ma das asas através de mar vés de mar vés de marcos anatômicos cos anatômicos. *Revista Brasileira de Zoologia*, SciELO Brasil, v. 21, n. 4, p. 847–855, 2004.

MERCER, J. Functions of positive and negative type and their connection with the theory of integrals equations. In: *Proc. R. Soc. London*. [S.l.: s.n.], 1909. v. 209, p. 415–446.

MILLIGAN, G. W. Clustering validation: results and implications for applied analyses. In: *Clustering and classification*. [S.l.]: World Scientific, 1996. p. 341–375.

MORAIS, F. M. d. *Clustering de dados biomédicos com algoritmos baseados em critérios entrópicos*. Tese (Doutorado) — Instituto Politécnico do Porto. Instituto Superior de Engenharia do Porto, 2012.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On spectral clustering: Analysis and an algorithm. In: *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*. [S.l.]: MIT Press, 2001. p. 849–856.

OLIVEIRA, R. A. d. Algoritmos para determinação do número de grupos em estudos de formas planas. Universidade Federal de Pernambuco, 2016.

ROHLF, F. J.; BOOKSTEIN, F. L. *Proceedings of the Michigan morphometrics workshop*. [S.l.]: University of Michigan Museum of Zoology, 1990.

SHAWE-TAYLOR, C. J.; KANDOLA, J. Spectral kernel methods for clustering. *Advances in Neural Information Processing Systems*, v. 14, 2001.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 22, n. 8, p. 888–905, 2000.

STOYAN, D.; STOYAN, H. *Fractals, random shapes and point fields: methods of geometrical statistics*. [S.l.: s.n.], 1994.

TRUSLOVE, G. M. The effect of selection for body weight on the skeletal variation of the mouse. *Genetical research*, Cambridge Univ Press, v. 28, n. 01, p. 1–10, 1976.

VERMA, D.; MEILA, M. A comparison of spectral clustering algorithms. *University of Washington Tech Rep UWCSE030501*, v. 1, p. 1–18, 2003.

WAGNER, D.; WAGNER, F. Between min cut and graph bisection. *Mathematical Foundations of Computer Science 1993*, Springer, p. 744–750, 1993.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on neural networks*, Ieee, v. 16, n. 3, p. 645–678, 2005.

YORK, D. G. et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, IOP Publishing, v. 120, n. 3, p. 1579, 2000.

ZELDITCH, M. L.; SWIDERSKI, D. L.; SHEETS, H. D. *Geometric morphometrics for biologists: a primer*. [S.l.]: Academic Press, 2012.

APÊNDICE A

IMPLEMENTAÇAO COMPUTACIONAL

Abaixo segue a implementação computacional do algoritmo adaptado de Ng, Jordan & Weiss para dados de formas planas de objetos. Por meio dos seguintes endereços de e-mail é possível estabelecer contato para esclarecimentos e disponibilização do *script* em R das aplicações utilizadas neste trabalho.

Marcelo Rodrigo - *marcelorpf@gmail.com*

Diogo Vasconcelos - *diogovasconcelos.17@gmail.com*

```
library(shapes)

DF = function(M) {
  # Full Procrustes distance
  n = dim(M)[2]
  D = matrix(NA, n, n)
  for (i in 1:n) {
    for (j in 1:n) {
      if (i == j)
        D[i,j] = 0
      else
        D[i,j] = procdist(M[,i], M[,j], type = "full")
    }
  }
  D
```

```
}

ngjordan = function(x, ncluster, subdim, kpar, thrs=1e-10)
{
  dim1 = dim(x)[2]
  D = DF(x)

  # Adjacency matrix
  w = exp(as.matrix(-D^2) / (2 * kpar^2))
  w[w < thrs] = thrs
  diag(w) = 0

  # Degree matrices
  d = matrix(0, nrow=dim1, ncol=dim1)
  diag(d) = apply(w, 1, sum)
  dn = d
  diag(dn) = diag(d)^(-0.5)

  # Laplacian matrix
  L = dn %*% w %*% dn

  # Eigendecomposition
  sv = eigen(L)
  lambda = sv[[1]]
  eigenvect = sv[[2]]

  # New representation
  y = as.matrix(eigenvect[,1:subdim])
  normaliz = sqrt(apply(y^2, 1, sum))
  y = y / normaliz

  res = list()
  res$y = y
  res$lambda = lambda
  res$labels = kmeans(y, ncluster)$cluster
  res
}

ngjordan0 = function(x, ncluster, subdim, kpar, thrs=1e-10)
```

```
{
  dim1 = dim(x)[1]

  # Adjacency matrix
  w = exp(as.matrix(-dist(x)^2) / (2 * kpar^2))
  w[w < thrs] = thrs
  diag(w) = 0

  # Degree matrices
  d = matrix(0, nrow=dim1, ncol=dim1)
  diag(d) = apply(w, 1, sum)
  dn = d
  diag(dn) = diag(d)^(-0.5)

  # Laplacian matrix
  L = dn %*% w %*% dn

  # Eigendecomposition
  sv = eigen(L)
  lambda = sv[[1]]
  eigenvect = sv[[2]]

  # New representation
  y = eigenvect[,1:subdim]
  normaliz = sqrt(apply(y^2, 1, sum))
  y = y / normaliz

  res = list()
  res$y = y
  res$lambda = lambda
  res$labels = kmeans(y, ncluster)$cluster
  res
}
```