LUIZ HENRIQUE GAMA DORE DE ARAÚJO

AGRUPAMENTO EM ANÁLISE ESTATÍSTICA DE FORMAS

RECIFE-PE ï FEVEREIRO/2008.

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

AGRUPAMENTO EM ANÁLISE ESTATÍSTICA DE FORMAS

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Mestre.

Área de Concentração: Modelagem Etatística Computacional (com ênfase nas áreas agrárias, biológicas e humanas)

Orientador(a): Prof. Dr. Borko D. Stosic

Co-orientador(a): Prof. Dr. Getúlio Amaral

Co-orientador(a): Profa. Dra. Rosângela Lessa

RECIFE-PE ï FEVEREIRO/2008.

FICHA CATALOGRÁFICA

A663 l Araújo, Luiz Henrique Gama Dore de
Agrupamento em análise estatística de formas / Luiz Henrique Gama Dore de Araújo. -- 2008.
38 f. : il.

Orientador : Borko D. Stosic
Dissertação (Mestrado em Biometria e Estatística Aplicada) ï Universidade Federal Rural de Pernambuco. Departamento de Es - tatística e Informática.
Inclui apêndice bibliografia.

CDD 574. 018 2

1. Análise estatística
2. Agrupamento
3. K - médias
I. Stosic, Borko D.
II. Título

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

AGRUPAMENTO EM ANÁLISE ESTATÍSTICA DE FORMAS

LUIZ HENRIQUE GAMA DORE DE ARAÚJO

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 27/02/2008 pela Comissão Examinadora.

Orientador:

Prof. Dr. Borko D. Stosic
Universidade Federal Rural de Pernambuco

Banca Examinadora:

Prof. Dr. Cristiano Ferraz
Universidade Federal de Pernambuco

Prof. Dr. Getúlio Amaral
Universidade Federal de Pernambuco

Profa. Dra. Viviane Oliveira
Universidade Federal Rural de Pernambuco

Resumo

Neste trabalho, o algoritmo k-médias proposto por Hartigan e Wong foi adaptado para o caso no qual se tem observações de um elemento aleatório sobre um espaço métrico arbitrário. Resultados de simulações indicam que o desempenho do algoritmo, no caso em que o espaço métrico é o espaço das formas de configurações planas, é invariante com relação às três métricas de forma usuais a saber, as distâncias de Procrustes completa e parcial e a distância de Procrustes. Além disso, a versão modificada do algoritmo, quando aplicada no espaço das formas com qualquer uma destas três métricas, apresenta o mesmo desempenho do algoritmo original aplicado às coordenadas de Procrustes tangentes parciais. Um problema na identificação das espécies de peixes-agulhas *Hemiramphus balao* e *Hemiramphus brasiliensis* motivou este estudo. Atualmente, os parâmetros de identificação utilizados apresentam alguns problemas operacionais os quais permitem, em muitos casos, que peixes-agulha de uma espécie sejam classificados como da outra. O algoritmo foi utilizado para agrupar uma amostra das formas de configurações destes peixes e dois grupos com padrões de forma estatisticamente distintos foram encontrados. Estes grupos apresentaram uma diferença marcante na posição da cabeça com relação ao resto do corpo: no grupo 1 a cabeça é levemente inclinada para cima enquanto que no grupo 2 a cabeça é levemente inclinada para baixo. A observação destas características em fotos de peixes-agulha nas quais as duas espécies foram corretamente identificadas, permitiu constatar que o grupo 1 corresponde à espécie *Hemirapmphus balao* e o grupo 2 à espécie *Hemiramphus brasiliensis*. Dessa maneira, a posição da cabeça com relação ao resto do corpo (a qual é uma informação totalmente baseada na forma do peixe), pode ser utilizada como um parâmetro bastante robusto para identificação de sua espécie.

Abstract

In this work, the k-means algorithm proposed by Hartigan and Wong is adapted to the case of random element observations in general metric space. Simulation results show that the performance of the algorithm in the case when the metric space is the shape space of the plane configurations, is independent on the choice of the usual shape metrics, more precisely the regular, complete and partial Procrustes distance. Besides, this modified version of the algorithm, applied to the shape space with any of the three metrics, exhibits the same performance as the original algorithm applied to the partial tangent Procrustes coordinates. The current study was motivated by the problem of identification of species of half-beak fish *Hemiramphus balao* and *Hemiramphus brasiliensis*.Currently, the parameters used for identification of these species are subject to certain operational difficulties, which often result in erroneous classification of the specimens. The algorithm was used to perform clustering of shape configuration samples, and two groups with statistically distinct shapes have been identified. These groups exhibit a pronounced difference regarding position of the head in relation to the body: for one group the head is slightly inclined upwards, while for the other group the head is slightly inclined downwards. Observation of these characteristics on the photos of fish specimens on which the two species were correctly classified, leads to identification of group 1 as *Hemirapmphus balao* and group 2 as species *Hemiramphus brasiliensis*. Therefore, head position with relation to body (which represents information entirely on the specimen shape) represents a rather robust parameter for identification of species.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A discriminação entre espécies de peixes (em particular espécies semelhantes) representa um fator de alto impacto no manejo e exploração de estoques pesqueiros. Por outro lado, esta atividade pode apresentar diversas dificuldades, e sua efetiva implementação exige investigações de novas metodologias, no sentido de oferecer procedimentos cada vez mais simples e mais confiáveis.

Em particular, a identificação das espécies de peixes-agulha *Hemiramphus balao* e *Hemiramphus brasiliensis* tem sido feita com base em algumas caracetrísticas destes peixes descritas por (COLLETTE, 1965), as quais podem ser vizualizadas na tabela 1 e na figura 1.1.

Tabela 1.1: Características utilizadas na identificação de *Hemiramphus balao* e *Hemiramphus brasiliensis*

	Hemiramphus balao	*Hemiramphus brasiliensis*
Nadadeira Peitoral	Seu tamanho é maior que a distância da base do raio peitoral à margem anterior da fossa nasal	Seu tamanho é menor que a distância da base do raio peitoral à margem anterior da fossa nasal
Cor do Lobo Caudal Superior	Apresenta uma cor azul-violáceo em vida	Apresenta uma cor laranja-avermelhado em vida

Figura 1.1: Parâmetros utilizados na identificação das espécies de peixes-agulha *Hemiramphus Balao* e *Hemiramphus Brasilientsis*. 1-margem anterior da fossa nasal, 2-início da nadadeira peitoral, 3-fim da nadadeira peitoral, 4-lobo caudal superior.

Estes critérios apresentam alguns problemas operacionais que podem diminuir as chances de sucesso na identificação. As variações entre as duas espécies nos comprimentos do maior raio das nadadeiras peitorais são bastante pequenas, tornando difícil a identificação das espécies por este critério. Também é bastante comum que estas nadadeiras apresentem danos ocorridos durante a pesca ou no armazenamento dos peixes, o que torna a discriminação por meio deste critério impossível. A cor do lobo superior da nadadeira caudal, característica de cada espécie, é observável apenas no espécimem vivo. Pouco tempo após sua morta, sua cauda perde esta coloração.

A morfometria geométrica representa uma área de pesquisa relativamente nova, computacionalmente intensiva, cujo desenvolvimento e aplicação tem crescido significativamente nos últimos anos. Ela é baseada no fato de que organismos com diferentes características biológicas geralmente apresentam formas diferentes. Conseqüentemente, devido às incertezas sobre a identificação das espécies de peixes-agulha *Hemiramphus balao* e *Hemiramphus brasiliensis*, o presente trabalho foi motivado pela suposição de que as duas espécies apresentam certa variabilidade de forma, e que a análise de formas pode ser utilizada para identificação destas espécies. Mais precisamente, o objetivo geral é de identificar numa amostra de peixes-agulha, dois grupos que tenham padrões de forma distintos, e verificar se estes grupos correspondem às duas espécies.

Análise de formas em geral baseia-se em métodos de agrupamento, onde varias técnicas já se encontram estabelecidas na literatura. Também, como estes métodos por sua vez baseiam-se no conceito de distância entre pontos num espaço multidimensional, diversas opções existem para escolha de própria definição de distância. Atualmente não é claro na literatura científica quais são as vantagens e desvantagens destas escolhas diferentes, e conseqüentemente neste trabalho é feita uma comparação entre elas, usando dados sintéticos controlados, dados de forma já estudados na literatura, e dados de forma de espécies de peixes-agulha.

Um método de agrupamento bastante utilizado na prática é o método das k-médias (LEMBER, 2003). O método das k-médias consiste na divisão de um conjunto de observações de um elemento aleatório 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. Hartigan e Wong (1979) propõem um algoritmo que garante que o agrupamento obtido produz uma soma de quadrados localmente mínima, no sentido de que ela não pode ser diminuida movendo-se uma observação de um grupo para o outro. No entanto, este algoritmo está escrito para dados euclidianos. Assim, uma versão adaptada para espaços

métricos gerais é desenvolvida neste trabalho.

O espaço das formas de configurações planas (como espaço de formas dos peixes *Hemiramphus*), é um espaço metrizável e sua distância natural é a distância de Procrustes (KENDALL, 1984). No entanto, duas outras distâncias entre formas podem ser utilizadas: a distância de Procrustes completa e a distância de Procrustes parcial. Também, pode-se utilizar a distância euclidiana entre duas pré-formas projetadas sobre o espaço tangente ao espaço de pré-formas sobre a forma média.

Neste trabalho, inicialmente o algoritmo de Hartigan e Wong modificado é aplicado a dados simulados com o intuito de comparar o desempenho do algoritmo quando utilizado com a distância de Procrustes completa, a distância de Procrustes parcial e a distância de Procrustes. O algoritmo original também é aplicado aos mesmos dados simulados projetados no plano tangente e seu desempenho é comparado com o desempenho de sua versão modificada. Em seguida, o algoritmo é aplicado a dois conjuntos de dados conhecidos na literatura. Cada conjunto de dados é uma amostra contendo configurações provinientes de populações com características biológicas distintas. A eficácia do agrupamento em identificar tais populações é avaliada. Por fim, é feito o agrupamento da amostra de peixes-agulha. Os peixes-agulha serão previamente classificados de acordo com os critério de Collette (1965). Os grupos obtidos pelo algoritmo serão cruzados com os grupos obtidos pelo critério de Collette para identificar que grupo representa qual espécie.

2 Metodologia utilizada para análise de formas

Neste capitulo serão descritos os conceitos e as técnicas utilizadas para analise de formas. Primeiro são definidos conceitos básicos: configurações, formas, distâncias entre formas, coordenadas de forma e o espaço das formas. Em seguida são definidas formas médias e são discutidos dois testes para comparação de formas médias, os quais serão utilizados na validação dos agrupamentos. Finalmente, é discutido o método k-medias, o qual será utilizado para o agrupamento de formas.

2.1 Formas, Distâncias entre Formas, Coordenadas de Forma e o espaço das Formas

Uma configuração de um determinado objeto é um conjunto ordenado de pontos localizados sobre este objeto. Uma configuração de um objeto plano é dita ser uma configuração plana. Os elementos de uma configuração plana são pontos no plano e, portanto, uma configuração plana pode pode ser considerada um vetor complexo. Se x denotar uma configuração plana contendo p pontos e (x_{1j}, x_{2j}) for seu j-ésimo ponto, então pode-se escrever

$$x = (x_{11} + ix_{21}, \cdots, x_{1p} + ix_{2p}).$$

Uma configuração plana representa uma figura geométrica plana. Kendall (1984), de maneira informal, define a forma de uma figura como sendo o que resta da figura quando as informações sobre posição, orientação e tamanho são desconsideradas. Isto significa que duas configurações, x_1 e x_2, têm a mesma forma se $\exists\ \gamma \in C,\ \beta \in R^+$ e $\theta \in (0, 2\pi)$ tais que

$$x_2 = T_{(\gamma,\beta,\theta)}(x_1) = \gamma 1_p + \beta e^{i\theta} x_1,$$

onde, $1_p^T = (1, \cdots, 1)$ é um vetor p-dimensional.

$x_1 + \gamma 1_p$ é a translação de x_1 pelo vetor determinado por γ, βx_1 é a dilatação de x_1 pelo fator β e $e^{i\theta}x_1$ é a rotação de x_1 por um ângulo θ. Estas três transformações são responsáveis por mudanças na posição, no tamanho e na orientação de x_1, respectivamente, e $T_{(\gamma,\beta,\theta)}(x_1)$ é dito ser uma transformação de similaridade de x_1 (DRYDEN; MARDIA, 1998).

Diz-se que uma configuração plana está centrada quando a soma de seus pontos é igual a zero. Denotando por $< x, y >$ o produto interno hermitiano canônico entre dois vetores complexos p-dimensionais x e y, o qual é dado por $y^\star x = \sum_{j=1}^{p} \bar{y_j} x_j$, tem-se que uma configuração plana x contendo p marcos está centrada quando $< 1_p, x >= 0$.

Para comparar as formas de x_1 e x_2, é necessário estabelecer uma medida de dissimilaridade entre formas. Por definição, $T_{(\gamma,\beta,\theta)}(x_1)$ e x_1 têm a mesma forma. Portanto, uma medida de dissimilaridade entre as formas de x_1 e x_2 pode ser obtida encontrando-se uma transformação de similaridade $T_{(\gamma,\beta,\theta)}$ que torne $T_{(\gamma,\beta,\theta)}(x_1)$ o mais próximo possível de x_2. Como $T_{(\gamma,\beta,\theta)}(x_1)$ e x_1 têm a mesma forma, a diferença entre $T_{(\gamma,\beta,\theta)}(x_1)$ e x_2 indicará a magnitude da diferença entre as formas de x_1 e x_2. Considerando o modelo

$$x_2 = T_{(\gamma,\beta,\theta)}(x_1) + \epsilon,$$

tem-se que esta medida de dissimilaridade entre formas é obtida calculando-se os valores de γ, β e θ que minimizam o comprimento do vetor $\epsilon = x_2 - T_{(\gamma,\beta,\theta)}(x_1)$. Ou seja, deve-se encontrar $\hat{\gamma},\ \hat{\beta}$ e $\hat{\theta}$ tais que

$$||\hat{\epsilon}|| = ||x_2 - T_{(\hat{\gamma},\hat{\beta},\hat{\theta})}(x_1)|| = \inf_{\gamma,\beta,\theta} ||x_2 - T_{(\gamma,\beta,\theta)}(x_1)||. \tag{2.1}$$

Se x_1 e x_2 são centradas, então $\hat{\gamma} = 0$, $\hat{\beta} = \frac{|<x_1,x_2>|}{<x_1,x_1>}$ e $\hat{\theta} = -\arg(< x_1, x_2 >)$ e

$$||\hat{\epsilon}|| = \sqrt{< x_2, x_2 > - \frac{< x_1, x_2 >< x_2, x_1 >}{< x_1, x_1 >}} \text{ (DRYDEN; MARDIA, 1998).} \tag{2.2}$$

$||\hat{\epsilon}||$ dado por (2.2) é a medida de dissimlaridade procurada. Supondo-se $||x|| = ||y|| = 1$, tem-se que

$$||\hat{\epsilon}|| = \sqrt{1 - < x_1, x_2 >< x_2, x_1 >}. \tag{2.3}$$

(2.3) é chamada distância de Procrustes completa entre as formas de x_1 e x_2, e é denotada por $d_C(x_1, x_2)$.

$T_{(\hat{\gamma},\hat{\beta},\hat{\theta})}(x_1)$, denotada por x_1^C, é dada por

$$x_1^C = T_{(\hat{\gamma},\hat{\beta},\hat{\theta})}(x_1) =< x_2, x_1 > x_1 \text{ (DRYDEN; MARDIA, 1998).}$$

x_1^C é chamada ajuste de Procrustes completo de x_1 sobre x_2. As coordenadas de x_1^C são chamadas coordenadas de Procrustes completas de x_1.

Diz-se que uma função $D : C^p \rightarrow C^p$ retorna coordenadas de forma de configurações planas contendo p pontos se

$$\forall x \text{ e } y \in C^p,\ D(x) = D(y) \Leftrightarrow\ x \text{ e } y \text{ têm a mesma forma.}$$

Nota-se que as coordenadas de Procrustes completas são coordenadas de forma.

No cálculo da distância de Procrustes completa (2.3), admitiu-se que as configurações envolvidas no cálculo eram centradas e normalizadas. Para tornar uma configuração centrada, basta subtrair seus pontos pelo seu centróide enquanto que a normalização é feita dividindo-se seus pontos pela sua norma. Assim, se x_1 e x_2 não são centradas nem normalizadas, z_1 e z_2, dados por

$$z_1 = \frac{x_1 - c_1 1_p}{||x_1 - c_1 1_p||} \text{ e } z_2 = \frac{x_2 - c_2 1_p}{||x_2 - c_2 1_p||}$$

onde $c_1 = \frac{1}{p}\sum_{j=1}^{p} x_{1j}$ e $c_2 = \frac{1}{p}\sum_{j=1}^{p} x_{2j}$, respectivamente, os são.

Kent (1994) sugere que uma configuração plana seja centrada pré-multiplicando-a pela sub-matriz de Helmert, a qual é a matriz de Helmert (LANCASTER, 1965) sem a primeira linha . A sub-matriz de Helmert, denotada por H, é uma matriz $(k-1) \times k$, cuja j-ésima linha é dada por

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

onde o número de elementos nulos nesta linha é $k - j - 1$ e $j = 1, \cdots, k-1$.

Assim, z_1 e z_2 dadas por

$$z_1 = \frac{Hx_1}{||Hx_1||} \text{ e } z_2 = \frac{Hx_2}{||Hx_2||} \tag{2.4}$$

são configurações centradas e normalizadas. O procedimento adotado para o cálculo das pré-formas é o dado por 2.4.

Se x_1 e x_2 têm a mesma forma, então z_1 e z_2 diferem apenas em orientação. Isto significa que $\exists\ \theta \in (0, 2\pi);\ z_2 = e^{i\theta} z_1$. Kendall (1984) nomeou z_1 e z_2 de pré-formas de x_1 e x_2, respectivamente, pois das três informações contidas na configuração que são indesejáveis à análise de formas, a pré-forma contém apenas a orientantação.

As operações utilizadas no cálculo de z_1 e z_2 são translações e dilatações. Portanto, z_1 e z_2 têm a mesma forma de x_1 e x_2, respectivamente, e, logo, não faz diferença utilizar z_1

e z_2 ou x_1 e x_2. Portanto, o estudo das formas de configurações planas pode ser reduzido ao estudo das formas de suas pré-formas.

Como as pré-formas são vetores complexos unitários, o espaço das pré-formas de configurações planas contendo p pontos é uma esfera complexa unitária de dimensão $p-1$, a qual é denotada por CS^{p-1}.

Seja z a pré-forma de uma configuração x contendo p pontos. O conjunto de todas as pré-formas que têm a mesma forma de x, denotado por $[x]$, é dado por

$$[x] = \{y;\ y = e^{i\theta}z,\ \theta \in (0, 2\pi)\}.$$

$[x]$ é dito ser uma fibra de CS^{p-1}.

O fato de que o espaço de pré-formas é uma esfera complexa unitária permite o uso de duas outras medidas de distância entre formas: a distância de Procrustes parcial e a distância de Procrustes.

A distância de Procrustes parcial entre x_1 e x_2, denotada por $d_P(x_1, x_2)$, é a distância euclidiana entre x_2 e o elemento de $[x_1]$ mais próximo de x_2, segundo a distância euclidiana.

Logo, $d_P(x_1, x_2)$ é dada por

$$d_P(x_1, x_2) = ||z_2 - e^{i\hat{\theta}}z_1|| = \inf_{\theta} ||z_2 - e^{i\theta}z_1||. \tag{2.5}$$

Pode-se notar que o problema de minimizar (2.5) é idêntico ao (2.1), exceto pelo fato de que em (2.5), o parâmetro correspondente à dilatação não é considerado (isto justifica o uso dos termos completa e parcial). Como os valores críticos dos parâmetros em (2.1) são calculados independentemente uns dos outros, $\hat{\theta}$ em (2.5) é o mesmo que $\hat{\theta}$ em (2.1).

A pré-forma $x_1^P = e^{i\hat{\theta}}z_1$ é chamada ajuste de Procrustes parcial de x_1 sobre x_2. Pode-se verificar que as coordenadas de x^P são coordenadas de forma.

Utilizando-se algumas relações trigonométricas, pode-se mostrar que o ângulo entre x_1^P e x_2 é dado por

$$2\,\text{arcsen}\left(\frac{1}{2}d_P(x_1, x_2)\right) = 2\,\text{arcsen}\left(\frac{1}{2}\sqrt{2(1- < x_1, x_2 >)}\right).$$

Este ângulo é a distância de Procrustes entre x_1 e x_2, a qual é denotada por $d(x_1, x_2)$.

O ângulo entre as pré-formas z_1 e $e^{i\theta}z_2$ é dado por $2\,\text{arcsen}\left(\frac{1}{2}||z_1 - e^{i\theta}z_2||\right)$. Ou seja,

o ângulo entre z_1 e $e^{i\theta}z_2$ é uma função monótona de θ. Logo, pode-se concluir que

$$\begin{aligned}\inf_{\theta} 2\,\text{arcsen}\left(\frac{1}{2}||z_1 - e^{i\theta}z_2||\right) &= 2\,\text{arcsen}\left(\frac{1}{2}\inf_{\theta}||z_1 - e^{i\theta}z_2||\right)\\ &= 2\,\text{arcsen}\left(\frac{1}{2}d_P(x_1,x_2)\right)\\ &= d(x,y).\end{aligned}$$

Portanto, x_1^P é o elemento de $[x_1]$ cujo ângulo formado com x_2 é o menor possível e a medida deste ângulo é $d(x_1, x_2)$.

Todos os elementos de uma fibra têm a mesma forma e elementos pertencentes a fibras distintas tem formas distintas. Assim, a própria fibra pode ser considerada uma forma.

Nota-se que as fibras são classes de equivalência defindas pela relação de equivalência $\sim$ que associa duas pré-formas z_1 e z_2 se $\exists\ \theta \in (0, 2\pi);\ z_2 = e^{i\theta}z_1$. Assim, cada fibra no espaço de pré-formas é um ponto no espaço quociente

$$\Sigma_2^p = CS^{p-1}/\sim .$$

Σ_2^p com a topologia quociente, a qual considera $A \subset \Sigma_2^p$ aberto se $\bigcup_{[x]\in A}[x] \subset CS^{p-1}$ é aberto, é chamado espaço das formas das configurações planas contendo p pontos (KENDALL, 1984).

Kendall (1984) mostra que Σ_2^p é uma variedade riemmanniana compacta cuja distância riemanniana é a distância distância de Procrustes.

2.2 Formas médias

Seja (M, ρ) um espaço métrico, X um elemento aleatório em M com distribuição F e $S = \{X_1, \cdots, X_n\}$ uma amostra de X. (ZIEZOLD, 1994) A média de Fréchet de X é qualquer ponto μ que satisfaça

$$\int_M \rho^2(x,\mu)dF(x) = \inf_{y\in M}\int_M \rho^2(x,y)dF(x), \tag{2.6}$$

e a média de Fréchet de S é qualquer ponto $\hat{\mu}$ que satisfaça

$$\sum_{i=1}^{n}\rho^2(x,\hat{\mu}) = \inf_{y\in M}\sum_{i=1}^{n}\rho^2(x,y). \tag{2.7}$$

Como o cálculo da média num espaço métrico geral é um problema de minimzação, a existência e unicidade da média não são garantidas. Também as equações acima podem não ter forma fechada.

No caso em que $M = \Sigma_2^p$ e ρ é qualquer uma das três distâncias entre formas já mencionadas, a existência é garantida, pois estes espaços métricos são compactos (KENDALL, 1984). Resta apenas saber como proceder com o cáculo para cada uma das três distâncias entre formas.

Se M é uma variedade riemanniana, ρ é a distância gerada pela métrica riemmaniana em M e S está contida numa região fortemente convexa de M, o algoritmo A3 (ver apêndice), proposto por Pennec (1994), pode ser utilizado no cálculo da média.

Como Σ_2^k é uma variedade riemmaniana e a distância de procrustes é a distância gerada pela métrica riemmaniana em Σ_2^k (KENDALL, 1984), a média gerada pela distância de procrustes pode ser calculada utilizando-se o algoritmo de Pennec.

No caso em que a distância entre formas é a distância de Procrustes completa, Kent (1994) mostra que a média é única e é dada pelo autovetor correspondente ao maior autovalor da matriz $P = \sum_{i=1}^{n} z_i z_i^\star$, onde z_i é a pré-forma de X_i.

A forma média definida pela distância de Procrustes parcial é chamada forma média parcial. O cálculo da forma média parcial pode ser feito utilizando-se o algoritmo A4 (ver apêndice) proposto por Ziezold (1994). Este algoritmo baseia-se na seguinte proposição:

Proposição 1. *Se $[\hat{\mu}]$ é uma forma média parcial de S, então $\hat{\mu} = \sum_{j=1}^{n} T_{\hat{\mu}}(X_j)$, na qual, $T_{\hat{\mu}}(X_j)$ denota o ajuste de Procrustes parcial de X_i sobre $\hat{\mu}$. Além disso, se $\hat{\mu}_1 = \frac{1}{n}\sum_{j=1}^{n} T_{\hat{\mu}_0}(X_i)$, tem-se que*

$$\hat{\mu}_1 \neq \hat{\mu}_0 \Rightarrow [\hat{\mu}_1] \neq [\hat{\mu}_0] \text{ e } \sum_{j=1}^{n} d_P^2(X_j, \hat{\mu}_1) < \sum_{j=1}^{n} d_P^2(X_j, \hat{\mu}_0). \tag{2.8}$$

Assim, se $\hat{\mu}_0$ denotar uma aproximação inicial para $\hat{\mu}$, o algoritmo fornece como nova aproximação de $\hat{\mu}$ a média aritimética dos ajustes de Procrustes parciais de das configurações de S sobre $\hat{\mu}_0$. Denotanto esta nova aproximação por $\hat{\mu}_1$, o algoritmo fornece como outra aproximação a média aritimética dos ajustes de Procrustes parciais das configurações de S sobre $\hat{\mu}_1$. Se depois de m repetições deste processo, $\hat{\mu}_m$ for sufucientemente próxima de $\hat{\mu}_{m-1}$, considera-se que houve convergência do algoritmo e é assumido que $\hat{\mu} = \hat{\mu}_m$.

Em geral, pode-se garantir apenas que $\sum_{j=1}^{n} d_P^2(X_j, \hat{\mu})$, com $\hat{\mu}$ tendo sido obtido pelo algoritmo $A3$, é um mínimo local. Como uma tentativa de contornar este problema, Ziezold

(1994) sugere que várias formas médias parciais, correspondentes a várias estimativas iniciais, sejam calculadas e que $\hat{\mu}$ seja escolhida como sendo aquela que produzir o menor valor da soma de quadrados.

2.3 Testes para comparação de formas médias

Teste de Hotelling - Sejam X e Y variáveis aleatótias reais p-dimensionais tais que $X \sim N(\mu_1, \Sigma)$ e $Y \sim N(\mu_2, \Sigma)$. Sejam $S_1 = \{X_1, \cdots, X_{n_1}\}$ e $S_2 = \{Y_1, \cdots, Y_{n_2}\}$ amostras aleatórias de X e Y, respectivamente, tais que Cov$(X_i, X_j) =$ Cov$(Y_i, Y_j) = 0$, $i \neq j$, e Cov$(X_i, Y_j) = 0$.

Denote as médias amostrais de S_1 e S_2 por $\bar{X}$ e $\bar{Y}$, respectivamente. A distância de Mahalanobis entre $\bar{X}$ e $\bar{Y}$ é dada por

$$D(\bar{X}, \bar{Y}) = \sqrt{(\bar{X} - \bar{Y})^T \hat{\Sigma}^{-1} (\bar{X} - \bar{Y})}$$

onde $\hat{\Sigma} = \frac{n_1\hat{\Sigma}_1 + n_1\hat{\Sigma}_2}{n_1+n_2-2}$ e $\hat{\Sigma}_i$ é matriz de covariância amostral de $S_i,\ i = 1, 2$.

Deseja-se testar a hipótese $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$.

(HOTELLING, 1931) Sob H_0, tem-se que

$$T = \frac{n_1 n_2 (m_1 + n_2 - p - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)p} \cdot D^2(\bar{X}, \bar{Y}) \sim F_{(p, n_1+n_2-p-1)}.$$

O teste definido por T é conhecido como teste de Hotelling. Ao nível $100 \cdot \alpha\%$ de significância, H_0 é rejeitada se $P(F_{(p,n_X 1+n_2-p-1)} \geq T) \leq \alpha$.

Na análise estatística da forma, o teste de Hotelling é aplicado às coordenadas de Procrustes tangentes parciais. Se os ajustes de Procrustes parciais de uma amostra de configurações constituem um conjunto de dados concentrados, as médias amostrais das coordenadas de Procrustes tangentes parciais de dois grupos deste conjunto de dados são aproximadente as formas médias destes grupos. Portanto, se as coordenadas de Procrustes tangentes parciais satisfazem as suposições impostas pelo teste de Hotelling, este teste pode ser utilizado para testar a igualdade entre as formas médias de duas populações.

Se as configurações são compostas por p marcos em 2 dimensões, tem-se que

$$T = \frac{n_1 n_2 (m_1 + n_2 - M - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)M} \cdot D^2(\bar{v}, \bar{w})$$

onde $M = 2(k-2)$ e $\bar{v}$ e $\bar{w}$ são as médias das coordenadas tangentes das duas amostras,

respectivamente.

Teste de Goodall - Sejam $e_1, \cdots, e_{n_1}$ e $e'_1, \cdots, e'_{n_2}$ vetores complexos k-dimensionais cujas partes reais e imaginárias de suas coordenadas são observações de uma variável aleatória normalmente distribuída com média 0 e variância σ^2. Sejam $X_1, \cdots, X_{n_1}$ e $Y_1, \cdots, Y_{n_2}$ trasnformações de similaridade de $e_1, \cdots, e_{n_1}$ e $e'_1, \cdots, e'_{n_2}$, respectivamente, tais que

$$X_i = \beta_i e^{i\theta_i}(\mu_1 + e_i) + \gamma_i 1_k^T \text{ e } Y_j = \beta'_j e^{i\theta'_j}(\mu_2 + e'_j) + \gamma'_j 1_k^T.$$

Seja $\hat{\mu}_0$ a forma média de Procrustes completa de $\{X_1, \cdots, X_{n_1}, Y_1, \cdots, Y_{n_2}\}$.

Deseja-se testar a hipótese $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$.

(GOODALL, 1991) Sob H_0, com σ pequeno, as quantidades T_1, T_2 e T_3 definidas abaixo distribuem-se, aproximadamente, como

$$T_1 = \sum_{i=1}^{n_1} d_F^2(X_i, \hat{\mu}_1) \sim \tau_0 \chi^2_{(n_1-1)M},$$

$$T_2 = \sum_{i=1}^{n_2} d_F^2(Y_i, \hat{\mu}_2) \sim \tau_0 \chi^2_{(n_2-1)M},$$

$$T_3 = d_F^2(\hat{\mu}_1, \hat{\mu}_2) \sim \tau_0 (\frac{1}{n_1} + \frac{1}{n_2}) \chi^2_M,$$

nas quais $\tau_0 = \frac{\sigma}{\delta_0}$, $\delta_0 = S(\hat{\mu}_0) = ||H^T \hat{\mu}_0||$, H é a submatriz de Helmert de ordem k, $\hat{\mu}_1$ é a forma média de Procrustes completa de $\{X_1, \cdots, X_{n_1}\}$ e $\hat{\mu}_2$ é a forma média de Procrustes completa de $\{Y_1, \cdots, Y_{n_2}\}$. Além disso, T_1 e T_2 são independentes e, T_i e T_3, $i = 1, 2$, são aproximadamente independentes. Portanto, sob H_0 tem-se

$$T = \left(\frac{n_1 + n_2 - 2}{\frac{1}{n_1} + \frac{1}{n_2}} \right) \frac{T_3}{T_1 + T_2} \sim F_{[M,(n_1+n_2-2)M]}$$

O teste definido pela estatística T é conhecido como teste de Goodall. Ao nível de $100 \cdot \alpha\%$ de significância, H_0 é rejeitada se $P(F_{[M,(n_1+n_2-2)M]} \geq T) \leq \alpha$.

Pode-se mostrar que o teste de Goodall é idêntico ao teste de Hotelling sob a suposição de isotropia da distribuição de X e Y (DRYDEN; MARDIA, 1998). Assim, quando as suposições do teste de Goodall são válidas, este teste se torna mais poderoso que o teste de Hotelling pois menos graus de liberdade são utilizados na estimação da matriz de covariância (DRYDEN; MARDIA, 1998).

2.4 k-médias

Seja (M, ρ) um espaço métrico, X um elemento aleatório em M e $S = \{X_1, \cdots, X_n\}$ uma amostra de X.

Uma k-partição de S é uma classe de subconjuntos de S, $P(k) = \{C_1, \cdots, C_k\}$, tal que

$$\bigcup_{i=1}^{n} C_i = S \text{ e}$$
$$C_i \cap C_j = \emptyset.$$

Seja P^k o conjunto de todas as k-partições de S. O método k-médias consiste em encontrar $P_0(k) = \{C_{01}, \cdots, C_{0k}\} \in P^k$ tal que

$$V_k(S) = SQ[P_0(k)] = \inf_{P(k) \in P^k} SQ[P(k)], \tag{2.9}$$

na qual $SQ[P(k)] = \sum_{i=1}^{n} \sum_{j=1}^{k} I_{(x_i \in C_j)} \rho^2(x_i, \hat{\mu}_j)$, $I_{(x \in C)} = 1$ se $x \in C$, $I_{(x \in C)} = 0$ se $x \notin C$ e $\hat{\mu}_i$ é a média de Fréchet amostral de C_i (2.7). $P_0(k)$ é dito ser uma k-partição globalmente ótima de S e $V_k(S)$ é chamada de k-variância amostral de S.

Para que se tenha certeza de que uma k-partição seja globalmente ótima, é necessário que o valor de SQ avaliado nesta k-partição seja menor ou igual ao valor de SQ avaliado em todas as outras k-partições em P^k. No entanto, o número muito grande de k-partições em P^k torna esta comparação impraticável. Ao invés de se buscar uma k-partição globalmente ótima, algoritmos iterativos são utilizados para encontrar uma k-partição localmente ótima.

Define-se uma vizinhança de k-partições para cada k-partição. Começando de uma k-partição inicial, a k-partição localmente ótima é encontrada movendo-se de uma k-partição para outra em sua vizinhança, de acordo com alguma regra de movimentação, até que a movimentação seja encerrada, segundo algum critério de parada. O ponto no qual a movimentação é encerrada é considerado uma k-partição localmente ótima. As regras de movimentação e de parada são determinadas pelos algoritmos iterativos.

Hartigan e Wong (1979) propõem um algoritmo, para o caso no qual $M = R^p$ e $\rho(x, y) = ||x - y||$, que tem o objetivo de encontrar uma k-partição cuja soma de quadrados não pode ser reduzida transferindo-se um elemento de um grupo para outro. Este algoritmo considera que a vizinhaça de $P(k)$ é o conjunto das k-partições que podem ser obtidas movendo-se um elemento de um grupo de $P(k)$ para outro grupo. Deve-se mo-

ver de uma k-partição $P_1(k)$ para uma k-partição $P_2(k)$ se $SQ(P_1(k)) < SQ(P_2(k))$ e o movimento é encerrado quando se atinge uma k-partição a qual, dentre as suas vizinhas, apresenta a menor soma de quadrados. O algoritmo de Hartigan e Wong encontra-se descrito no apêndice 1 (algoritmo A1).

Para elaborar uma versão do algoritmo A1 para espaços métricos gerais, deve-se fazer algumas observações sobre as quantiades $R1$ e $R2$ neste algoritmo.

Sejam C_1 e C_2 grupos numa k-partição $P(K)$ e C_1' e C_2'' os grupos obtidos transferindo-se uma obsevação X no grupo C_1 para o grupo C_2. Ou seja,

$$C_1' = C_1 - \{X\} \text{ e } C_1'' = C_2 \cup \{X\}.$$

denotando por $\bar{C}$ a média do grupo C, tem-se que

$$\begin{aligned}
||X - \bar{C_1}||^2 &= ||X - \frac{SC_1}{N_{C_1}}||^2 = \frac{1}{N_{C_1}^2}||N_{C_1}X - SC_1||^2 \\
&= \frac{1}{N_{C_1}^2}||N_{C_1}X - X - (SC_1 - X)||^2 \\
&= \frac{1}{N_{C_1}^2}||X(N_{C_1} - 1) - SC_1'||^2 \\
&= \frac{(N_{C_1} - 1)^2}{N_{C_1}^2}||X - \frac{SC_1'}{(N_{C_1} - 1)}||^2 \\
&= \frac{(N_{C_1} - 1)^2}{N_{C_1}^2}||X - \bar{C'}_1||^2,
\end{aligned}$$

e

$$\begin{aligned}
||X - \bar{C_2}||^2 &= ||X - \frac{SC_2}{N_{C_2}}||^2 = \frac{1}{N_{C_2}^2}||N_{C_2}X - SC_2||^2 \\
&= \frac{1}{N_{C_2}^2}||N_{C_2}X + X - (SC_2 + X)||^2 \\
&= \frac{1}{N_{C_2}^2}||X(N_{C_2} + 1) - SC_2''||^2 \\
&= \frac{(N_{C_2} + 1)^2}{N_{C_2}^2}||X - \frac{SC_2''}{(N_{C_2} + 1)}||^2 \\
&= \frac{(N_{C_2} + 1)^2}{N_{C_2}^2}||X - \bar{C''}_2||^2,
\end{aligned}$$

nas quais $SC = \sum_{i=1}^{n} I_{(x_i \in C)} x_i$. Logo, $||X - \bar{C'}_1||^2$ e $||X - \bar{C''}_2||^2$ são dados por

$$\frac{N_{C_1}^2}{(N_{C_1} - 1)^2} ||X - \bar{C}_1||^2 \text{ e } \frac{N_{C_2}^2}{(N_{C_2} + 1)^2} ||X - \bar{C}_2||^2,$$

respectivamente.

As quantidades $R1$ e $R2$ no algoritmo $A1$ são, portanto,

$$R1 = ||X - \bar{C'}_1|| \cdot ||X - \bar{C}_1|| \text{ e } R2 = ||X - \bar{C''}_2|| \cdot ||X - \bar{C}_2||.$$

No caso geral, tem-se

$$R1 = \rho(X, \bar{C'}_1) \cdot \rho(X, \bar{C}_1) \text{ e } R2 = \rho(X, \bar{C''}_2) \cdot \rho(X, \bar{C}_2), \tag{2.10}$$

na qual ρ é a métrica adotada.

Para que o algoritmo $A1$ possa ser aplicado em espaços métricos gerais, o cálculo de $R1$ e $R2$ deve ser feito seguindo-se os passos abaixo:

1. Transfira uma observação do seu grupo atual para o grupo desejado;

2. Atualize as médias dos dois grupos;

3. Calcule $R1$ e $R2$ de acordo com (2.10).

O algoritmo $A2$ no apêndice 1 corresponde à versão do algortimo $A1$ para espaços métricos arbitrários.

3 Resultados e Discussão

Neste capítulo encotram-se apresentados os resultados da análise de formas, aplicando o algoritmo de Hartigan e Wong, modificado neste trabalho para espaços métricos gerais.

A Comparação da performance para diversas escolhas da métrica de forma é feita primero usando dados sintéticos controlados, em seguida dois conjuntos de dados de forma já estudados na literatura, e finalmente dados de forma de espécies de peixes-agulha.

3.1 Avaliação da performance do agrupamento de formas em dados simulados

Foram consideradas nas simulações dois tipos de configurações planas: um quadrado, denotado por μ_1, e um octógono, denotado por μ_2. Duas outras configurações, denotadas por μ_1' e μ_2', foram contruídas pertubando-se o segundo vértice de μ_1 e μ_2, respectivamente. As figuras 3.1 e 3.2 exibem as coordenadas de Procrustes parciais de μ_1' com relação à μ_1 e μ_2' com relação à μ_2, respectivamente.

Amostras de Monte Carlo foram geradas segundo o modelo $X(\mu) = \mu + e$, no qual μ é uma configuração e $e = (e_1, \cdots, e_p)$ com $e_j = e_{1j} + ie_{2j}$ e $e_{ij} \sim N(0, \sigma^2)$.

Foram gerados três tipos de amostras: uma contendo 30 observações de $X(\mu_1)$ e 30 observações de $X(\mu_1')$, uma contendo 30 observações de $X(\mu_1)$ e 15 observações de $X(\mu_1')$, e uma contendo 30 observações de $X(\mu_2)$ e $X(\mu_2')$. Três valores de σ^2 diferentes foram utilizados: $0,01$, $0,1$ e $0,5$, nos casos em que $\mu = \mu_1$ e $\mu = \mu_1'$, e $0,001$, $0,005$ e $0,01$ nos casos em que $\mu = \mu_2$ e $\mu = \mu_2'$.

Na figura 3.3 encontra-se uma amostra simulada correspondente a cada um dos casos descritos acima para cada caso citado acima.

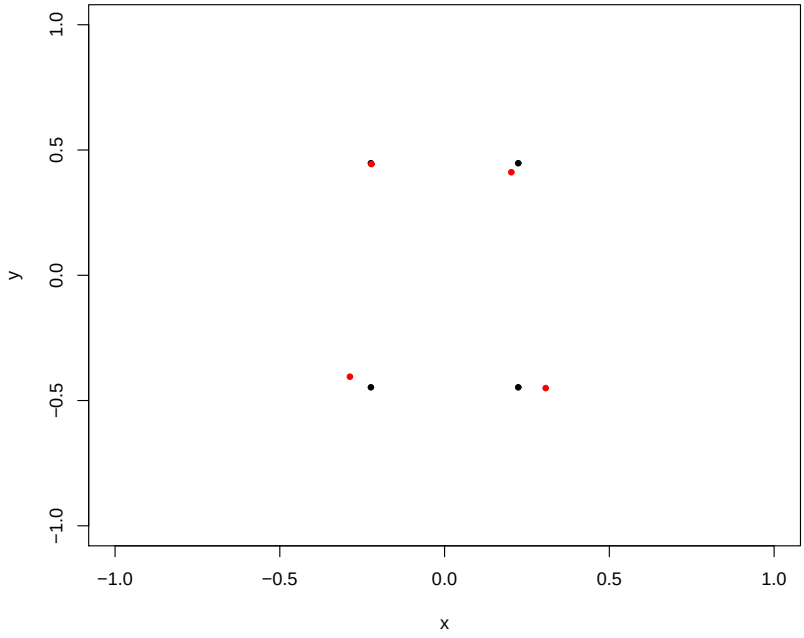

Figura 3.1: Ajuste parcial de μ_1' (vermelho) sobre μ_1 (preto).

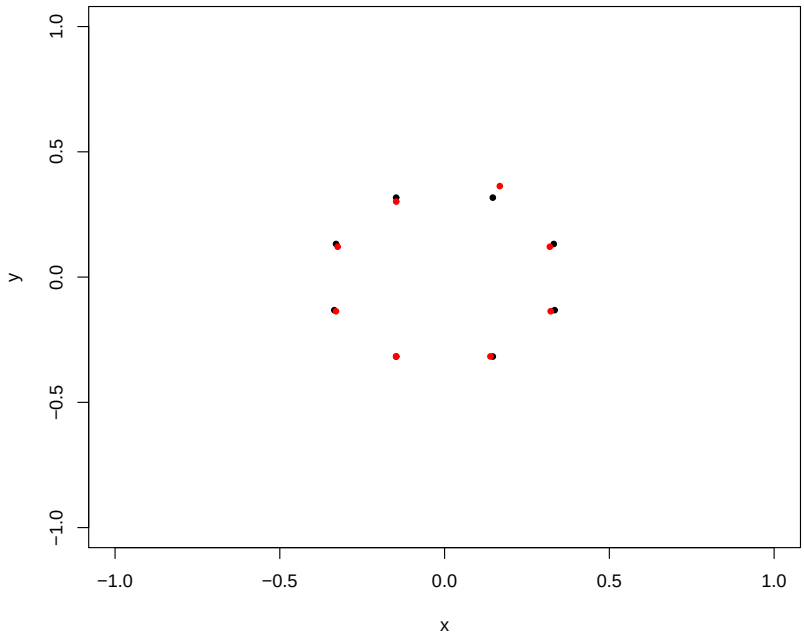

Figura 3.2: Ajuste parcial de μ_2' (vermelho) sobre μ_2 (preto).

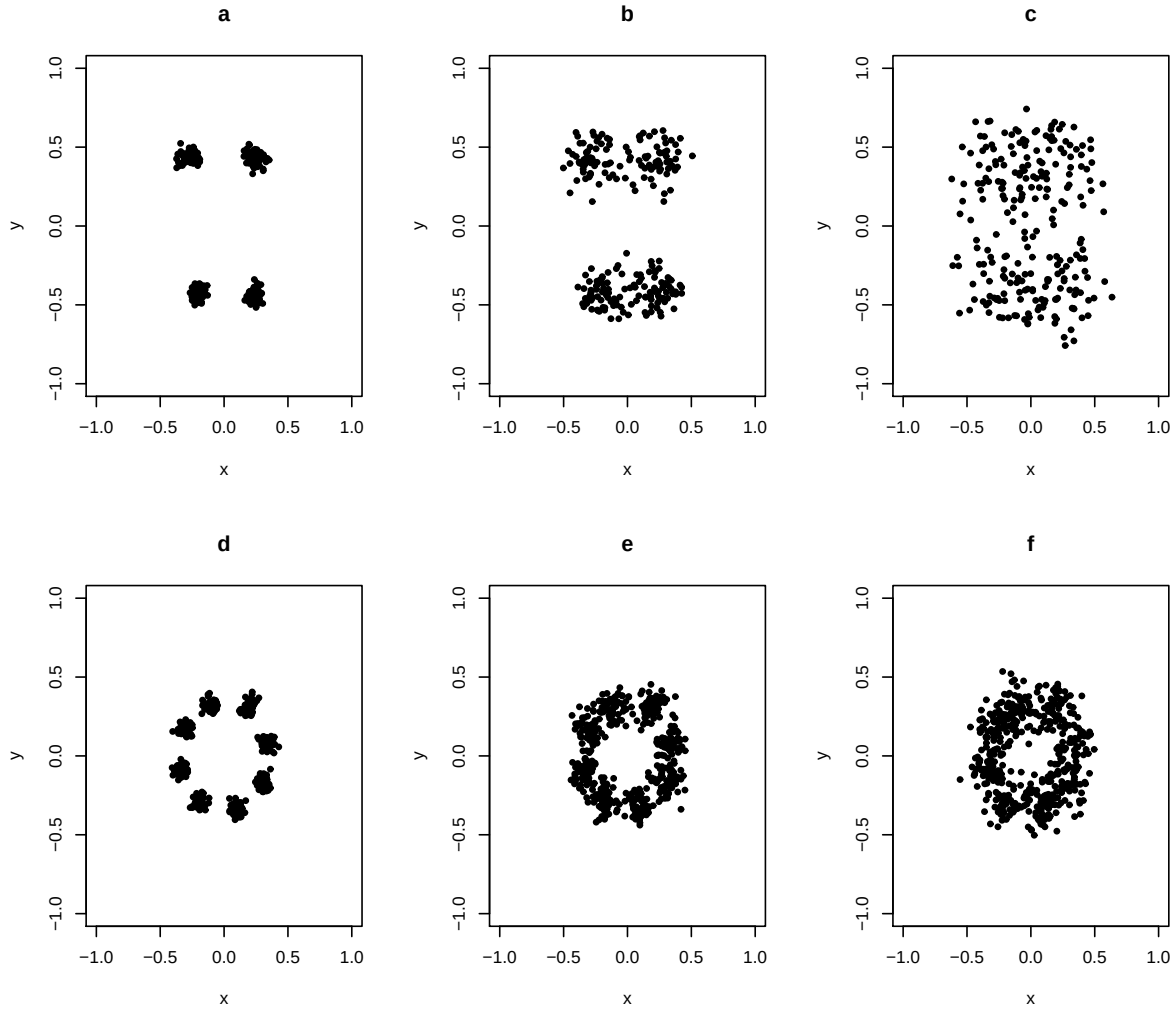

Figura 3.3: Exemplos do primeiro e do terceiro tipo de amostra gerada.

Para cada um destes três tipos, e para cada valor de σ, 1.000 amostras de Monte Carlo foram geradas.

O algoritmo $A2$ com cada uma das três distâncias de Procrustes e o algoritmo $A1$ foram aplicados a cada amostra para cada um destes casos. As médias das taxas de alocação e das k-variâncias sobre cada cada conjunto de 1.000 amostras foram calculadas e encontram-se nas tabelas 3.1, 3.2 e 3.3.

As k-variâncias foram calculadas, em todos os casos, utilizando-se a distância de Procrustes, para que a performance dos agrupamentos podessem ser comparadas. Em todos os casos, tanto as k-variâncias como as taxas de alocação são praticamente as mesmas para as três distâncias e para o agrupamento sobre as coordenadas tangentes. As taxas de alocação decaem enquanto as k-variâncias aumentam com o aumento de σ. A homogeneidade na performance nas quatro diferentes maneiras de se realizar o agrupamento

de formas não parece depender da variabilidade dos dados e nem do número de marcos nas configurações consideradas. Também parece que esta homogeneidade não é afetada pelo fato de uma amostra não ser balanceada, isto é, conter mais observações de um grupo do que de outro.

Tabela 3.1: Médias das taxas de alocação e a média das k-variâncias (entre parênteses) dos agrupamentos das amostras de Monte Carlo correspondentes a μ_1 $(n = 30)$ e μ_1' $(n = 30)$.

	Completa	Procrustes	Parcial	Tangente
$\sigma = 0,01$	0,8897 (0,3870)	0,8896 (0,3869)	0,8898 (0,3869)	0,8896 (0,3869)
$\sigma = 0,1$	0,6014 (3,2830)	0,6013 (3,2832)	0,6014 (3,2833)	0,6 (3,2826)
$\sigma = 0,5$	0,5617 (13,1085)	0,5617 (13,1123)	0,5619 (13,1090)	0,5606 (13,1189)

Tabela 3.2: Médias das taxas de alocação e a média das k-variâncias (entre parênteses) dos agrupamentos das amostras de Monte Carlo correspondentes a μ_1 $(n = 30)$ e μ_1' $(n = 15)$.

	Completa	Procrustes	Parcial	Tangente
$\sigma = 0,01$	0,8498 (0,2966)	0,8497 (0,2967)	0,8497 (0,2967)	0,8498 (0,2967)
$\sigma = 0,1$	0,594 (2,4777)	0,5934 (2,478)	0,5935 (2,4777)	0,592 (2,4764)
$\sigma = 0,5$	0,5637 (9,8116)	0,564 (9,8292)	0,5646 (9,8271)	0,5645 (9,8319)

Tabela 3.3: Médias das taxas de alocação e a média das k-variâncias (entre parênteses) dos agrupamentos das amostras de Monte Carlo correspondentes a μ_2 $(n = 30)$ e μ_2' $(n = 30)$.

	Completa	Procrustes	Parcial	Tangente
$\sigma = 0,001$	0,74 (0,661)	0,74 (0,661)	0,74 (0,661)	0,739 (0,661)
$\sigma = 0,005$	0,59 (3,065)	0,589 (3,064)	0,589 (3,064)	0,59 (3,065)
$\sigma = 0,01$	0,566 (5,970)	0,566 (5,901)	0,567 (5,9)	0,568 (5,899)

Como o uso da distância Euclidiana permite simplificações consideráveis no algoritmo $A2$, tornando o algoritmo mais rápido, pode-se concluir que a melhor alternativa dentre as quatro mencionadas para execução do agrupamento de formas utilizando-se o método k-médias é a aplicação do algoritmo $A1$ sobre as coordenadas tangentes.

3.2 Configurações sobre imagens do cérebro de esquizofrênicos e não-esquizofrênicos

O conjunto de dados corresponde a uma amostra de 28 configurações contendo 13 marcos anatômicos extraídos de imagens de ressonância magnética de cérebros de 14 indivíduos com esquizofrenia e 14 indivíduos saudáveis. Estes dados foram coletados e analisados com o objetivo de identificar diferenças na estrutura cerebral de esquizofrênicos e não-esquizofrênicos (DEQUARDO; BOOKSTEIN, 1996) (BOOKSTEIN, 1996). A figura 3.4a exibe os marcos anatômicos selecionados.

O resultado dos agupamentos podem ser vistos na tabela 3.4.Pode-se observar que os agrupamentos obtidos utilizando-se o algoritmo $A2$ com cada uma das três métricas de forma e utilizando-se o algoritmo $A1$ sobre as coordenadas tangentes são iguais. Este resultado, portanto, concorda com os resultados obtidos na simulação. O valor da taxa de alocação foi relativamente baixo, $57,14\%$.

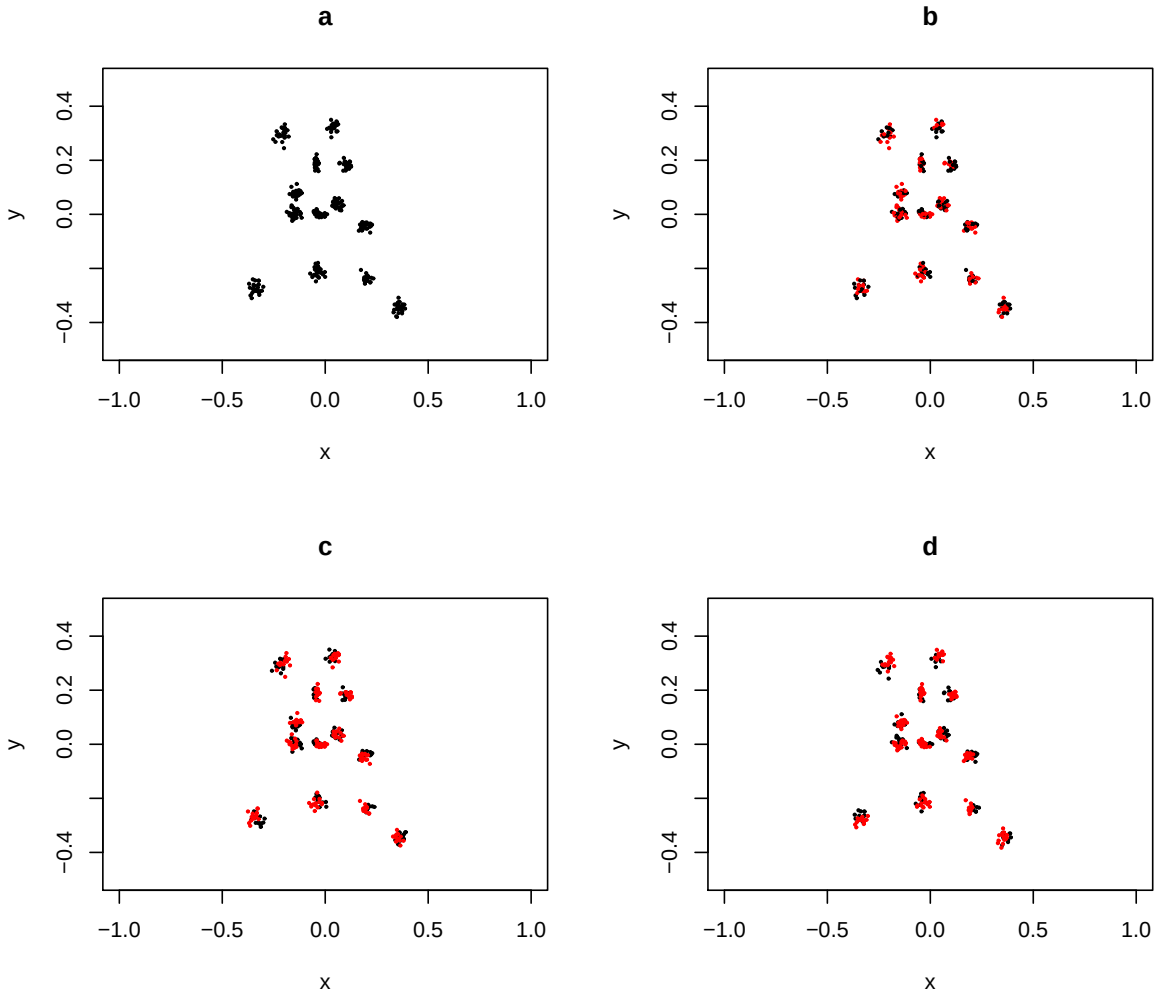

Figura 3.4: Coordenadas de Procrustes da amostra contendo os dois casos (esquizofrênicos e não-esquizofrênicos) (a e b), das amostras de cada caso (vermelho-esquizofrênico e preto-não-esquizofrênico) (c) e dos grupos obtidos pelo k-médias (d).

Tabela 3.4: Agrupamento da amostra de configurações de esquizofrênicos e não-esquizofrênicos. Taxa de alocação e as respectivas k-variâncias.

	Completa	Procrustes	Parcial	Tangente
taxa	0,5714	0,5714	0,5714	0,5714
ssq	0,1254	0,1254	0,1254	0,1254

Tabela 3.5: Estatísticas dos testes de Hotelling e de Goodall para igualdade entre as formas médias de esquizofrênicos e não-esquizofrênicaos e entre as formas médias dos grupos de pacientes obtidos pelo k-médias. Entre parênteses os p-valores.

	Hotelling	Goodall
Esquizofrênicos e não esquizofrênicos	0,834 (0,6579)	1,9036 (0,008)
Grupo vermelho e grupo preto	3,4727 (0,0854)	3,2942 (0)

3.3 Configurações sobre crânios de gorilas macho e fêmea

Amostra é composta por 59 configurações contendo 8 marcos anatômicos situados nos crânios de 29 gorilas machos e 30 gorilas fêmeas de acordo com a figura 3.5. A análise das formas destas configurações teve como objetivo detectar e descrever possível dimorfismo sexual entre gorilas (O'HIGGINS; DRYDEN, 1993).

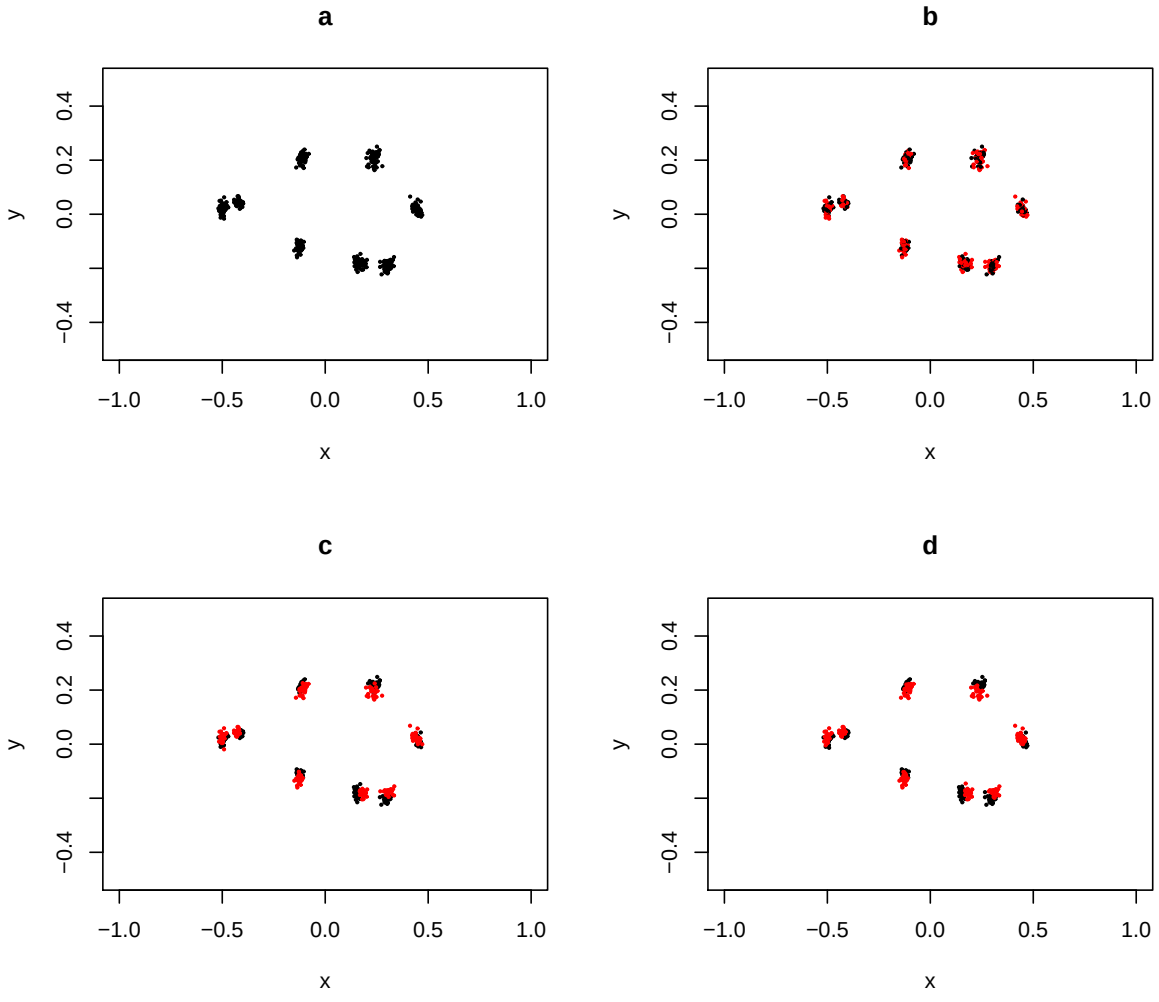

Figura 3.5: Coordenadas de Procrustes da amostra de configurações contendo gorilas dos dois sexos (a e b), das amostras de cada sexo (macho-preto e fêma-vermelho) (c) e dos grupos obtidos pelo k-médias (d).

O resultado do agupamento destes dados encontra-se na tabela 3.6. Os agrupamentos resultantes das três distâncias entre formas são idênticos entre si e quando comparados com o agrupamento das coordenadas tangentes parciais. A taxa de alocação neste agrupamento foi alta, 91.53%, o que é reflexo da diferença siginificativa entre as formas médias

dos machos e das fêmeas, conforme indicam os resultados dos testes de Hotelling e de Goodall na tabela 3.7.

Tabela 3.6: Agrupamento da amostra de configurações de gorilas machos e fêmeas. Taxas de alocação e as respectivas k-variâncias.

	Completa	Procrustes	Parcial	Tangente
taxa	0,9153	0,9153	0,9153	0,9153
V_k	0,1247	0,1247	0,1247	0,1247

Tabela 3.7: Estatísticas dos testes de Hotelling e de Goodall para igualdade entre as formas médias de gorilas macho e fêmea e entre as formas médias dos grupos de gorilas obtidos pelo k-médias. Entre parênteses os p-valores.

	Hotelling	Goodall
Machos e fêmeas	26,4704 (0)	22,29 (0)
Grupo vermelho e grupo preto	14.11986 (0)	25.5099 (0)

3.4 Aplicação do k-médias na identificação de peixes-agulha das espécies *Hemirmaphus balao* e *Hemiramphus brasiliensis*

A amostra consiste de 49 observações das quais 11 são da espécie *Hemiramphus balao* e 38 da espécie *Hemiramphus brasiliensis*. Os espécimens foram fotografados e as coordenadas de 11 marcos foram extraídas das fotografias digitalizadas utilizando-se o programa tpsDig (colocar referência). Estes 11 marcos encontram-se na figura 3.6.

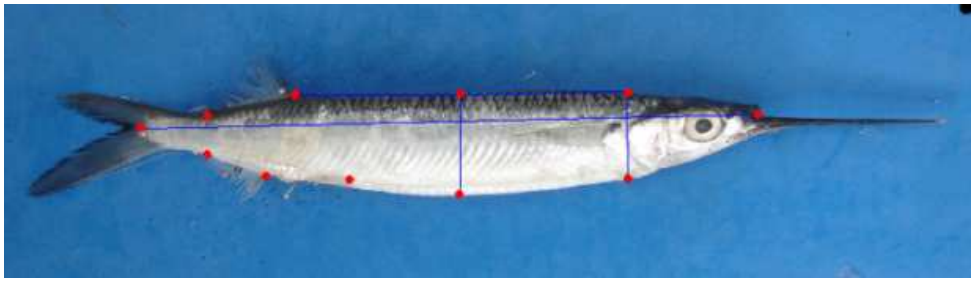

Figura 3.6: Espécimem *Hemiramphus brasiliensis com os marcos selecionados.*

Os resultados das aplicações dos testes de Hotelling e Goodall para avaliar a igualdade entre as formas médias das espécies *Hemiramphus Balao* e *Hemiramphus Brasiliensis*, identificadas de acordo com as características descritas na tabela 1, encontram-se na tabela 3.8.

Tabela 3.8: Estatísticas dos testes de Hotelling e de Goodall para igualdade entre as formas médias de peixes-agulha das espécies *Hemiramphus Balao Hemiramphus Brasiliensis* e entre as formas médias dos grupos de peixes-agulha obtidos pelo k-médias. Entre parênteses os p-valores.

	Hotelling	Goodall
Hemiramphus balao e *Hemiramphus brasiliensis*	1,3626 (0,2208)	6,6273 (0)
Grupos vermelho e preto	6,4673 (0)	37.6159 (0)

De acordo com o teste de Hotelling, as formas médias destas duas espécies podem ser concideradas iguais enquanto pelo teste de Goodall, pode-se concluir que a hipótese de igualdade entre as formas médias destas espécies deve ser rejeitada.

A figura 3.7 contém os ajustes de Procrustes parciais de cada espécie, enquanto a figura 3.8 contém os ajustes de Procrustes parciais dos grupos obtidos pelo k-médias.

Os espécimens no grupo preto apresentam a cabeça, com relação ao corpo, levemente inclinada para cima, enquanto os do grupo vermelho apresentam a cabeça, com relação ao corpo, levemente inclinada para baixo. Esta diferença na posição das cabeças pode ser observada, com menos nitidez, na figura 3.7, na qual os grupos correspondentes às espécies *Hemiramphus balao* e *Hemiramphus brasiliensis* assumem os papéis dos grupos preto e vermelho na figura 3.8, respectivamente. Esta perda de nitidez na vizualização destas diferenças pode ser atribuída a possíveis erros de classificação das espécies.

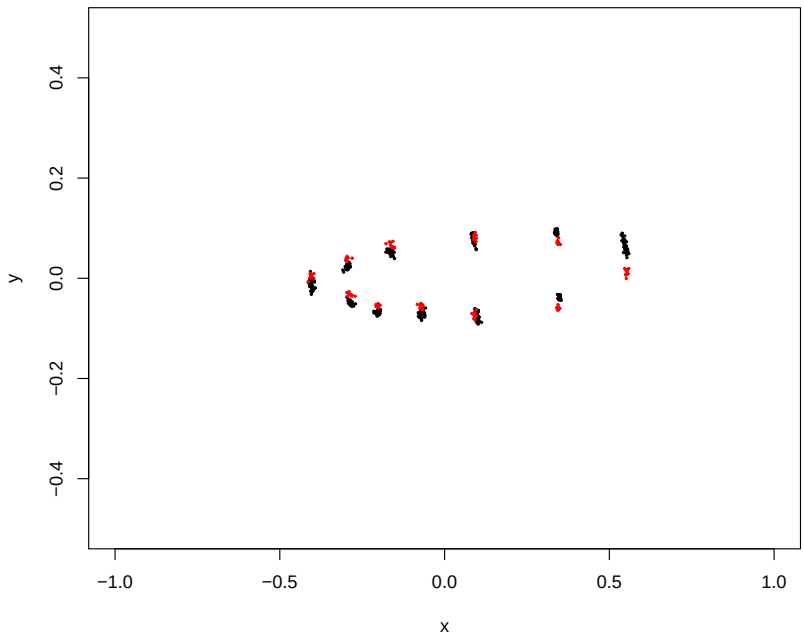

Figura 3.7: Coordenadas de Procrustes das configurações de *Hemiramphus Balao* (preto) *Hemiramphus Brasiliensis* (vermelho).

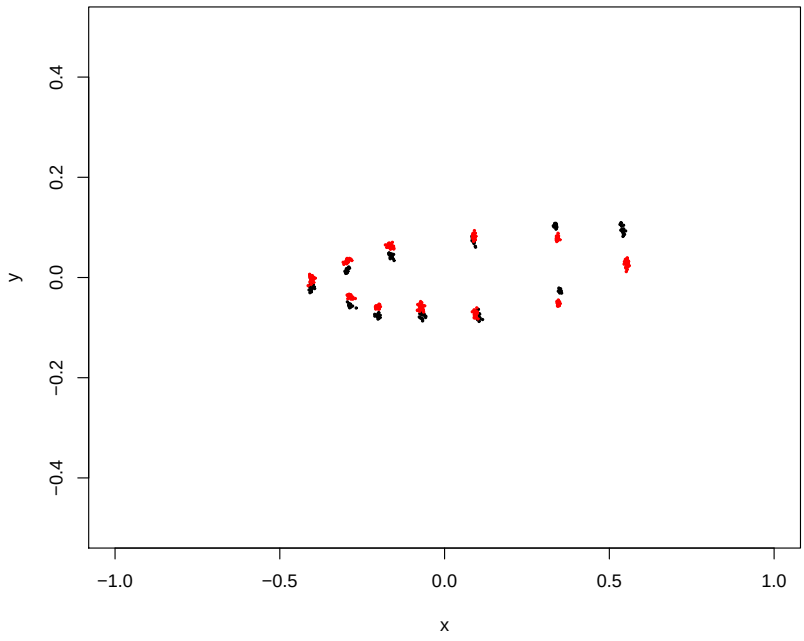

Figura 3.8: Coordenadas de Procrustes das configurações dos grupos obtidos pelo k-médias.

As diferenças exibidas pelos dois grupos na figura 3.8 também são aparentes nas fi-

guras 3.9 e 3.10, as quais são imagens de espécimens de *Hemiramphus brasiliensis* e *Hemiramphus balao*, respectivamente. Nota-se que o peixe-agulha *Hemiramphus brasiliensis* exibe características, com relação à posição da cabeça, semelhantes às do grupo vermelho enquanto o peixe-agulha da espécie *Hemiramphus balao* se assemelha mais aos do grupo preto. Isto evidencia o fato de que o grupo preto corresponde à espécie *Hemiramphus balao* e o grupo vermelho à espécie *Hemiramphus brasiliensis* e que, portanto, a posição da cabeça do peixe-agulha com relação ao corpo pode servir como parâmetro de identificação destas duas espécies.

Figura 3.9: Foto de um espécimem *Hemiramphus brasiliensis*.

Figura 3.10: Foto de um espécimem *Hemiramphus balao*.

4 Conclusão

A adaptação do algoritmo de Hartigan e Wong que foi feita neste trabalho lida com o caso no qual se tem observações de um elemento aleatório sobre um espaço métrico arbitrário, e os resultados das simulações indicam que o desempenho do algoritmo, no caso em que o espaço métrico é o espaço das formas de configurações planas, é invariante com relação às três distâncias de Procrustes. Além disso, a versão modificada do algoritmo, quando aplicada no espaço das formas com qualquer uma destas três métricas, apresenta o mesmo desempenho do algoritmo original aplicado às coordenadas de Procrustes tangentes parciais.

O problema na identificação das espécies de peixes-agulhas *Hemiramphus balao* e *Hemiramphus brasiliensis* que motivou este estudo foi solucionado utilizando o algoritmo proposto para agrupar uma amostra das formas de configurações destes peixes em dois grupos com padrões de forma estatisticamente distintos. Estes grupos apresentaram uma diferença marcante na posição da cabeça com relação ao resto do corpo: no primeiro grupo a cabeça é levemente inclinada para cima enquanto que no segundo grupo a cabeça é levemente inclinada para baixo. Foi constatado que o primeiro grupo corresponde à espécie *Hemirapmphus balao* e o segundo grupo à espécie *Hemiramphus brasiliensis*. Dessa maneira, a posição da cabeça com relação ao resto do corpo pode ser utilizada como um parâmetro identificação de sua espécie.

Referências

BOOKSTEIN, F. L. Biometrics, biomathematics and the morphometric sythesis. **Bulletin of Mathematical Biology**, v. 58, p. 313–365, 1996.

COLLETTE, B. B. Hemiramphidae (pisces, synentognathi) from tropical west africa. **Atlantide Reports**, v. 8, p. 217–235, 1965.

DEQUARDO, J. R.; BOOKSTEIN, F. L. Spatial relationships of neuroanatomic landmarks in schizophrenia. **Psychiatry Reaserch: Neuroimaging.**, v. 67, p. 81–95, 1996.

DRYDEN, I.; MARDIA, K. **Statistical Shape Analysis.** [S.l.: s.n.], 1998.

GOODALL, C. R. Procrustes methods in the statistical analysis of shape. **Journal of the Royal Statistical Society, Series B.**, v. 53, p. 285–339, 1991.

HARTIGAN, J. A.; WONG, M. A. Algorithm, as136: A k-means clustering algorithm. **Applied statistics**, England, v. 28, p. 100–108, 1979.

HOTELLING, H. The generalization of student's ratio. **The Annals of Mathematical Statistics.**, v. 2, p. 360–378, 1931.

KENDALL, D. G. Shape manifolds, procrustean metrics and complex projective spaces. **Bull. of London Math. Soc.**, v. 16, p. 81–121, 1984.

KENT, J. T. The complex bingham distribution and shape analysis. **Journal of the Royal Statistical Society Series B**, v. 56, p. 285–299, 1994.

LANCASTER, H. O. The helmert matrices. **American Mathematical Monthly**, v. 72, n. 1, p. 4–12, 1965.

LEMBER, J. On minimizing sequences of k-centres. **Journal of Approx. Theory**, v. 120, p. 20–35, 2003.

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

PENNEC, X. Probabilities and statistics on riemannian manifolds: basic tools for geometric measurements. **IEEE Workshop on Nonlinear Signal and Image Processing.**, 1994.

ZIEZOLD, H. Mean figures and mean shapes applied to biological figure and shape distributions in the plane. **Biomatrical Journal**, v. 36, n. 4, p. 491–510, 1994.

APÊNDICE A – Algoritmos

ALGORITMO A1

Considere um conjunto de dados contendo M observações e o número de grupos é K.

Seja $NC(L)$ o número de elementos no grupo L e $D(I, L)$ a distância entre a observação I e a média do grupo L.

Forneça um conjunto de K vetores n-dimensionais como valores iniciais para as K-médias.

Passo 1. Para cada $I(I = 1, \cdots, M)$, encontre a sua média mais próxima e sua segunda média mais próxima, $IC1(I)$ e $IC1(I)$, respectivamente. Atribua o ponto I ao grupo $IC1(I)$.

Passo 2. Atualize as médias dos grupos para serem as médias dos pontos contidos dentro deles.

Passo 3. Inicialmente, todos os grupos pertencem ao conjunto ativo.

Passo 4.(*Optimal transfer stage*): Considere cada ponto $I(I = 1, \cdots, M)$. Se o grupo $L(L = 1, \cdots, M)$ foi atualizado no passo 6, então ele pertence ao conjunto ativo. Caso contrário, em cada passo, ele não está no conjunto ativo se ele não foi atualizado nos últimos M passos do passo 4. Seja $L1$ o grupo do ponto I. Se $L1$ está no conjunto ativo, vá para o passo 4a. Casso contrário, vá para o passo 4b.

Passo 4a. Calcule o mínimo da quantidade $R2 = \frac{NC(L)D(I,L)^2}{NC(L)+1}$, sobre todos os grupos $L(L \neq L1,\ L = 1, \cdots, K)$. Seja $L2$ o grupo com menor $R2$. Se este valor é maior que ou igual a $\frac{NC(L1)D(I,L)^2}{NC(L1)-1}$, realocação não é necessária e $L2$ é o novo $IC2(I)$. (Note que $\frac{NC(L1)D(I,L)^2}{NC(L1)-1}$ é relembrado e permanecerá o mesmo para o ponto I até que $L1$ seja atualizado)

Caso contrário, o ponto I é alocado ao grupo $L2$ e $L1$ é o novo $IC2(I)$. As médias dos grupos são atualizadas para serem as médias dos pontos atribuídos a eles se realocação

tem ocorrido. Os dois pontos envolvidos na trasnferência do ponto I neste passo estão agora no conjunto ativo.

Passo 4b. Este passo é idêntico ao 4a, exceto que o mínimo de $R2$ é calculado somente sobre os grupos no conjunto ativo.

Passo 5. Pare se o conjunto ativo estiver vazio. Caso contrário, vá para o passo 6.

Passo 6.(*Quick transfer stage*): Considere cada ponto $I(I = 1, \cdots, M)$. Faça $L1 = IC1(I)$ e $L2 = IC2(I)$. Não é necessário checar o ponto I se ambos os grupos $L1$ e $L2$ não mudaram nos últimos M passos. Calcule os valores $R1 = \frac{NC(L1)D(I,L)^2}{NC(L1)-1}$ e $R2 = \frac{NC(L)D(I,L)^2}{NC(L)+1}$. (como notado anteriormente, $R1$ é relembrado e permanecerá o mesmo até que $L1$ seja atualizado).

Se $R1$ é menor que $R2$, o ponto I permanece no grupo $L1$. Caso contrário, troque $IC1(I)$ com $IC2(I)$ e atualize as médias dos grupos $L1$ e $L2$. Os dois grupos são também notados por seu envolvimento numa transferência neste passo.

Passo 7. Se nos últimos M passos nenhuma transferência foi realizada, vá para o passo 4. Caso contrário, vá para o passo 6.

ALGORITMO A2

Considere um conjunto de dados contendo M observações sobre um espaço métrico arbitrário e o número de grupos é K.

Denote por L_I^- o grupo L sem a observação I, supondo que esta obervação pertence ao grupo L, e por L_I^+ o grupo L com a observação I, supondo que esta observação não pertence ao grupo L. Isto é. $L_I^- = L - \{I\}$ e $L_I^+ = L \cup \{I\}$.

Seja $NC(L)$ o número de elementos no grupo L e $D(I, L)$ a distância entre a observação I e a média do grupo L.

Forneça um conjunto de K vetores n-dimensionais como valores iniciais para as K-médias.

Passo 1. Para cada $I(I = 1, \cdots, M)$, encontre a sua média mais próxima e sua segunda média mais próxima, $IC1(I)$ e $IC1(I)$, respectivamente. Atribua o ponto I ao grupo $IC1(I)$.

Passo 2. Atualize as médias dos grupos para serem as médias dos pontos contidos dentro deles.

Passo 3. Inicialmente, todos os grupos pertencem ao conjunto ativo.

Passo 4.(*Optimal transfer stage*): Considere cada ponto $I(I = 1, \cdots, M)$. Se o grupo $L(L = 1, \cdots, M)$ foi atualizado no passo 6, então ele pertence ao conjunto ativo. Caso contrário, em cada passo, ele não está no conjunto ativo se ele não foi atualizado nos últimos M passos do passo 4. Seja $L1$ o grupo do ponto I. Se $L1$ está no conjunto ativo, vá para o passo 4a. Casso contrário, vá para o passo 4b.

Passo 4a. Calcule o mínimo da quantidade $R2 = D(I, L_I^+) \cdot D(I, L_I)$ sobre todos os grupos $L(L \neq L1,\ L = 1, \cdots, K)$. Seja $L2$ o grupo com menor $R2$. Se este valor é maior que ou igual a $R1 = D(I, L1_I^-) \cdot D(I, L1_I)$, realocação não é necessária e $L2$ é o novo $IC2(I)$. (Note que $D(I, L1_I^-)$ é relembrado e permanecerá o mesmo para o ponto I até que $L1$ seja atualizado) Caso contrário, o ponto I é alocado ao grupo $L2$ e $L1$ é o novo $IC2(I)$. As médias dos grupos são atualizadas para serem as médias dos pontos atribuídos a eles se realocação tem ocorrido. Os dois pontos envolvidos na trasnferência do ponto I neste passo estão agora no conjunto ativo.

Passo 4b. Este passo é idêntico ao 4a, exceto que o mínimo de $R2$ é calculado somente sobre os grupos no conjunto ativo.

Passo 5. Pare se o conjunto ativo estiver vazio. Caso contrário, vá para o passo 6.

Passo 6.(*Quick transfer stage*): Considere cada ponto $I(I = 1, \cdots, M)$. Faça $L1 = IC1(I)$ e $L2 = IC2(I)$. Não é necessário checar o ponto I se ambos os grupos $L1$ e $L2$ não mudaram nos últimos M passos. Calcule os valores

$$R1 = D(I, L1_I^-) \cdot D(I, L1_I) \text{ e } R2 = D(I, L2_I^+) \cdot D(I, L2_I). \tag{A.1}$$

(como notado anteriormente, $R1$ é relembrado e permanecerá o mesmo até que $L1$ seja atualizado.) Se $R1$ é menor que $R2$, o ponto I permanece no grupo $L1$. Caso contrário, troque $IC1(I)$ com $IC2(I)$ e atualize as médias dos grupos $L1$ e $L2$. Os dois grupos são também notados por seu envolvimento numa transferência neste passo.

Passo 7. Se nos últimos M passos nenhuma transferência foi realizada, vá para o passo 4. Caso contrário, vá para o passo 6.

ALGORITMO A3

y_0: valor inicial para a média;

$\{x_1, \cdots, x_n\}$: amostra de configurações centradas e normalizadas;

e: erro na aproximação da forma média.

Passo 1 Atribua a y a configuração inicial y_0.

Passo 2 Atribua a Δy o vetor $\frac{1}{n}\sum_{i=1}^{n} \text{Log}_{y_0}(x_i)$.

Passo 3 Atribua a y o vetor $\frac{1}{n}\sum_{i=1}^{n} \text{Exp}_{y_0}(\Delta y)$.

Passo 4 Se $||\Delta y|| < e$ pare. caso contrário, atribua a y_0 o vetor y e repita os passos 2, 3 e 4.

ALGORITMO A4

y_0: valor inicial para a média;

$\{x_1, \cdots, x_n\}$: amostra de configurações centradas e normalizadas;

e: erro na aproximação da forma média.

Passo 1 Atribua a y a configuração inicial y_0.

Passo 2: De i=1 até n faça:

Se $< y, x_i > \neq 0$, então $u_i(y) = \frac{<y,x_i>}{|<y,x_i>|}$. Caso contrário, $u_i(y) = 1$.

Passo 3: Atribua a $T(y)$ o vetor $\frac{1}{n}\sum_{i=1}^{n} u_i(y)x_i$.

Passo 4 Se $||y - y_0|| < e$, pare. Caso contrário, atribua a y_0 o vetor y e repita os passos 2, 3 e 4.