UMA INTRODUÇÃO A ANÁLISE ESTATÍSTICA DE FORMA E CLASSIFICAÇÃO NÃO SUPERVISIONADA NO CONTEXTO NÃO EUCLIDIANO.

Jerfson Bruno do Nascimento Honório[1]
Getúlio José Amorim Do Amaral[2]

RESUMO

O objetivo deste artigo é dá uma introdução sobre análise estatística de forma e propor um método de classificação não supervisionado (K-médias) para dados de tamanho e forma considerando imagens bidimensionais (formas planas).

Conduzimos dados simulados em dois cenários para avaliar o desempenho do método proposto. Os resultados numéricos provam que para os conjuntos de dados, em que os tamanhos dos centroides se diferem o algoritmo ganha mais performance. Para as bases de dados real, vértebras de camundongo e ressonância magnética de pessoas com esquizofrenia, o método, novamente, se prova ser eficiente.

Palavras-chave: Marcos Anatômicos, Formas, Pré-Formas, Tamanho-e-Forma, K-médias.

1 INTRODUÇÃO

Com os avanços da tecnologia, a captura de imagens bidimensionais e tridimensionais tem se tornado cada vez mais comum no nosso cotidiano. Essas imagens fornecem diversas informações para estudos estatísticos, sendo essa área chamada de morfometria. A morfometria é uma das maneiras de estudar estas imagens que se encontram bem consolidadas com diversas aplicações, tais como: Medicina, Zoologia, Biologia e outros. Nesse contexto, existem estudos que tratam da forma, e estudos que tratam o tamanho e forma, dos objetos capturados nas imagens. No caso de forma, os efeitos de locação, escala e rotação são removidos. No caso de tamanho e forma, o efeito de escala não é removido.

No século atual, o amadurecimento da análise estatística de forma como uma área teórica e aplicada vem crescendo, uma vez que no atual século a maioria das tecnologias usam reconhecimento facial, ou seja, propriedades geométricas de tamanho e forma. As aplicações da análise estatística de forma se estendem por quase todas as áreas científicas e tecnológicas aplicadas, das menores às maiores escalas.

A análise de de forma ou tamanho e forma dos objetos, pode ser útil para a tomada de importantes decisões, como a de um médico que precisa decidir se um câncer é maligno ou benigno, baseado em uma ressonância magnética digitalizada.

[1] Doutorando em Estatística na Universidade Federal de Pernambuco-UFPE, Bolsista FACEPE, jerfson35@gmail.com;
[2] Professor Orientador: PhD em Estatística, Universidade Federal de Pernambuco-UFPE, gjaa@de.ufpe.br.

Em diversas ocasiões, em análise estatística de forma, é necessário agrupar um conjunto de dados em grupos, de tal maneira, que se tenha grupos com características mais homogêneas. Com isso, ter algoritmos que trabalhem no espaço não euclidiano, como é no caso dos dados de forma ou tamanho e forma, é necessário.

O trabalho atual foi motivado pelos poucos trabalhos em análise estatística de forma e classificação não supervisionado aplicados a essa área. Dos poucos trabalhos, temos o de Amaral et al. (2010), que adaptou o método K-médias, proposto por Macqueen (1967), para dados de pré-formas, com isso construir um algoritmo para tratar dados de tamanho e forma também é necessário.

2 METODOLOGIA

Neste trabalho foram consideradas algumas estratégias para avaliar o algoritmo proposto. Os dados de tamanho e forma foram criados artificialmente através de uma normal complexa central.

Para simular a distribuição normal complexa central, usamos o método proposto por Nascimento et al. (2016). O algoritmo 1 mostra o passo a passo de como são gerados os dados.

Algoritmo 1: Simulação da distribuição normal complexa central.

1 Considere uma matriz definida positiva Hermitiana de ordem p, sendo ela, definida como $\Sigma = \mathbf{R} + i\mathbf{I}$ em que $(i = \sqrt{-1})$ e sejam $\mathbf{R}$ e $\mathbf{I}$ partes reais e imaginárias, respectivamente;

2 Gere k normais multivariadas de ordem 6, $[y_1, ..., y_6]^\top \sim N_6(\mathbf{0}, \Sigma)$, em que

$$\Sigma = \begin{bmatrix} \mathbf{R} & -\mathbf{I} \\ \mathbf{I} & \mathbf{R} \end{bmatrix}.$$

3 Os k dados de tamanho e forma, são gerados a partir da equação $W_k = [y_1\ y_2\ y_3]^\top + i\,[y_4\ y_5\ y_6]^\top$

3 DESENVOLVIMENTO

3.1 ANÁLISE ESTATÍSTICA DE FORMA

Formas de objetos estão disponíveis em todos os lugares, seja em uma pesquisa na internet, uma ressonância magnética que você faz, ou até mesmo no desbloqueio de um celular. Essas tomadas de decisões servem para pesquisar, identificar, classificar e agrupar informações. A análise estatística de formas é um tópico relativamente recente e está relacionada com características e comparações de formas de objetos. Dryden e Mardia (2016) descrevem com mais detalhes os conceitos sobre análise de formas em seu livro.

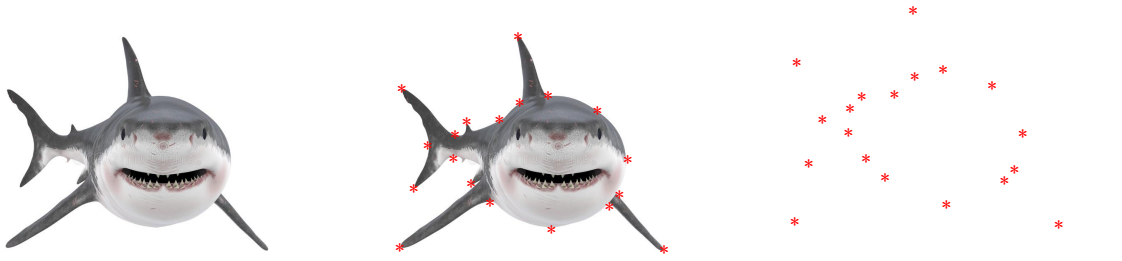
Uma maneira de descrever uma forma é indicar um subconjunto finito de pontos no contorno do objeto. O número de pontos não seguem uma restrição teórica segundo Kendall (1984). Esse número finito de pontos de cada objeto é conhecido como marcos.

Segundo Dryden e Mardia (2016) existe três tipos marcos, são eles:

- Marcos matemáticos: são pontos alocados em um objeto de acordo com propriedades matemáticas ou geométricas, por exemplo, pontos de máximo, mínimo, inflexão entre outros;
- Marcos anatômicos: são pontos atribuídos por um especialista que correspondem a alguma característica biologicamente significativa;
- Pseudo-Marcos: são pontos construídos em um objeto, localizados no contorno ou entre os marcos anatômicos e/ou matemáticos.

A Figura 1 mostra um exemplo da utilização desses marcos:

Figura 1 – Representação da forma baseada em marcos



(a) Objeto (b) Objeto com os marcos (c) 19 marcos para a configuração da forma do objeto

Fonte: Autor

A partir da colocação dos marcos, e obtendo suas coordenadas, podemos construir a matriz de configuração, e assim obter a forma do objeto. Para isso, devemos realizar transformações matemáticas para remover efeitos de escala, locação e rotação. Pois segundo Kendall (1977), formas são definidas como toda informação geométrica que permanece quando os efeitos de locação, escala e rotação são retirados de um objeto.

Seja $\mathbf{Y}$ uma matriz de configuração $k \times m$, em que k representa o número de marcos e m as dimensões. A matriz de configuração é definida como:

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,m} \\ y_{2,1} & \cdots & y_{2,m} \\ \vdots & \ddots & \vdots \\ y_{k,1} & \cdots & y_{k,m} \end{bmatrix} \tag{1}$$

Para iniciar as transformações em busca da forma do objeto, o primeiro passo é remover o efeito de locação. Esta transformação pode ser feita de diversas maneiras dependendo do sistema de coordenadas utilizados, mas aqui serão usadas as coordenadas de Kendall (1984).

Kendall (1984) usa a sub-matriz de Helmert $\mathbf{H}$ para remover os efeitos de locação. A matriz de Helmert completa $\mathbf{H}^F$ é uma matriz ortogonal em que a primeira linha tem todos os elementos iguais a $1/\sqrt{k}$. Já a j-ésima linha possui $j+1$ elementos, iguais a:

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

com $j=1,...,k-1$. Além disso, temos o número de elementos zeros na linha $j+1$ igual a $k-j-1$.

Um exemplo da matriz de Helmert dado $k=4$, é descrita como:

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

e sua sub-matriz é dada por:

$$\mathbf{H}=\begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{bmatrix} \tag{4}$$

Eliminamos a primeira linha de $\mathbf{H}^F$ para que o $\mathbf{HY}$ transformado não dependa da locação original da configuração. Observe também que $\mathbf{H}^\top\mathbf{H}=\mathbf{C}$, em que $\mathbf{C}$ é a matriz de centralização de $\mathbf{H}$. Para mais detalhes ver Dryden e Mardia (2016).

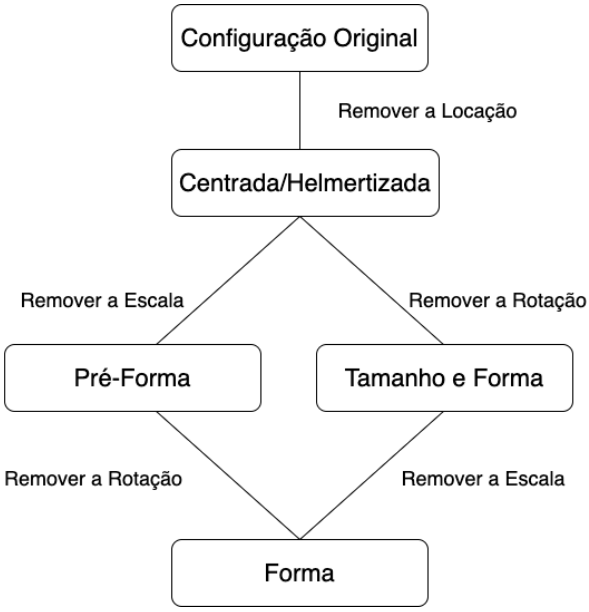
Com isso, a retirada dos efeitos de locação é dada por:

$$\mathbf{Y_C}=\mathbf{H}^\top\mathbf{HY}=\mathbf{CY} \tag{5}$$

em que $\mathbf{Y_C}$ é uma matriz $k\times m$, e é chamada de matriz de configuração centrada.

A partir da remoção desse efeito, é onde entramos nos três tipos de configurações mais usados; são eles: Pré-Forma, Forma e Tamanho-e-forma. A Figura 2 representa em diagrama essas três configurações, e como elas se relacionam entre sí:

Figura 2 – Diagrama dos vários espaços de estatística de forma.



Fonte: Autor

3.2 PRÉ-FORMA

A pré-forma de uma matriz de configuração $\mathbf{Y}$ tem todas as informações sobre locação e escala removidas.

A pré-forma de uma matriz de configuração $\mathbf{Y}$ é dada por:

$$\mathbf{Z}_C = \frac{\mathbf{CY}}{||\mathbf{CY}||}, \tag{6}$$

que é invariável sob a locação e escala da configuração original $\mathbf{Y}$.

3.3 TAMANHO E FORMA

Se a locação e a rotação forem removidas, mas não a escala, teremos o tamanho e a forma da matriz de configuração $\mathbf{Y}$.

O tamanho e a forma de uma matriz de configuração $\mathbf{Y}$ são todas as informações geométricas sobre $\mathbf{Y}$ que são invariáveis sob locação e rotação, e isso pode ser representado pelo conjunto $[\mathbf{Y}]_C$ dado por:

$$[\mathbf{Y}]_C = \{\mathbf{CY}\Gamma : \Gamma \in SO(m)\} \tag{7}$$

em que Γ é uma matriz de rotação.

Uma matriz de rotação tem $\frac{1}{2}m(m-1)$ graus de liberdade. Para $m = 2$ dimensões, a matriz de rotação pode ser parametrizada por um único ângulo $\theta, -\pi \leq \theta \leq \pi$ em radianos para

girar no sentido horário sobre a origem:

$$\Gamma = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Para mais detalhes ver Dryden e Mardia (2016)

4 CLASSIFICAÇÃO NÃO SUPERVISIONADA

Algoritmos de agrupamento por particionamento encontram uma partição que maximiza ou minimiza algum critério numérico. Este trabalho tenta adaptar para os dados de tamanho e forma, o algoritmo K-médias proposto por Macqueen (1967).

Uma descrição do algoritmo K-médias introduzido por Macqueen (1967), usa uma partição inicial e depois o valor médio dos objetos em um grupo como o protótipo de partição.

Ao trabalharmos com dados de tamanho e forma, a distância, normalmente Euclidiana, é substituída pela distância de procrustes no espaço de tamanho e forma, e ela é dada pela Equação 8:

$$d_S(Y_1, Y_2) = \sqrt{S_1^2 + S_2^2 - 2S_1S_2\cos\rho(Y_1, Y_2)} \tag{8}$$

em que S_1, S_2 são os tamanhos dos centroides de Y_1, Y_2, e ρ é a distância Riemannian. Para mais detalhes ver Dryden e Mardia (2016).

A medida de dissimilaridade utilizada no K-médias, é a soma dos quadrados totais, dada por Landau et al. (2011) em seu livro. A mesma modificada para os dados de tamanho e forma é definida como:

$$J = \sum_{m=1}^{\pi_m} \sum_{n=1}^{n_m} d_S(Y_{mn}, \mu_m) \tag{9}$$

em que Y_{mn} é o n-ésima observação do grupo m e μ é a média do grupo. A média para uma amostra aleatória de dados de tamanho e forma é definida como média de Fréchet ou Karcher:

$$\hat{\mu}_s = \arg\inf_{\mu} \frac{1}{n} \sum_{i=1}^{n} d_S^2(\mathbf{Y_i}, \mu) \tag{10}$$

em que d_S é a distância de tamanho e forma dada na Equação 8, e μ, nesta situação, é uma configuração $\mathbf{Y}$ escolhida e proposta com média inicial, $E[\mathbf{Y}] = \mu$.

Com isso, o algoritmo K-médias para tamanho e forma tem os seguintes passos:

Algoritmo 2: K-médias modificado para análise estatística de tamanho e forma.

Passo 1: Dados de Entrada

1.1 Seja $\Omega = 1, ..., n$ um conjunto de dados descrito pela matriz de configuração $\mathbf{Y}$.

1.2 Calcule os dados de tamanho e forma $\mathbf{Y_C}$ definido na Equação 5.

1.3 Escolha K centroides e gere uma partição inicial π_K

1.4 Fixe o número de iterações $T = 100$ e um erro $\varepsilon > 0$; Faça $t = 1$.

1.5 Calcule as médias μ_{K_t} dos grupos π_{K_t}, definido na Equação 10.

Passo 2: Atribuindo os objetos aos grupos

2.1 Calcule a distância de cada objeto em relação a média μ_{K_t} e atribua esse objeto ao novo grupo $\pi_{K_{t+1}}$ mais próximo.

2.1 Calcule as médias dos novos grupos $\pi_{K_{t+1}}$ formados.

Passo 3: Critério de Parada

3.1 Se $||\pi_{K_{t+1}} - \pi_{K_t}|| < \varepsilon$ ou $t > T$, então **Pare**. Caso contrário, faça $t = t + 1$ e volte para o passo 2.

5 VALIDAÇÃO DO CLUSTER

A validade de um estudo refere-se a quão bem os resultados encontrados representam resultados verdadeiros para indivíduos semelhantes fora do estudo. Este conceito de validade se aplica a todos os tipos de estudos, seja ele clínicos, sobre prevalência, associações, intervenções e diagnóstico. A validade de um estudo de pesquisa inclui dois domínios: a validade interna e a validade externa.

5.1 ÍNDICE INTERNO

A validade interna é definida como a extensão em que os resultados observados representam uma possível verdade para a população. Em nosso trabalho, usaremos o índice residual de Procrustes, já que o mesmo é muito utilizado em análise estatística de forma. Para mais detalhes ver Dryden e Mardia (2016).

O Índice Residual Procrustes é útil para encontrar o número ideal de grupos, e também avaliar a qualidade do ajuste em um conjunto de dados em análise estatística de formas.

Após obter a alocação dos indivíduos do conjunto de dados pelo o algoritmo K-médias. Calcula-se a norma quadrática dos resíduos de cada indivíduo dentro do seu grupo (r_{in}) e fora do seu grupo (r_{out}).

Com isso, o índice residual Procrustes $pr(i)$ para cada indivíduo do conjunto de dados é dado como

$$pr(i) = \frac{r_{out}(i) - r_{int}(i)}{\max(r_{out}(i), r_{int}(i))} \tag{11}$$

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

Para se obter um índice de validação geral de Procrutes, é feito a média de todos os $pr(i)$:

$$PR = \frac{1}{n}\sum_{i=1}^{n} pr(i) \tag{12}$$

Uma Tabela de interpretação para os resultados do PR, é baseada na tabela de interpretação do índice silhueta do livro de Izenman (2008).

5.2 ÍNDICE EXTERNO

Feito por Rand (1971), o índice Rand ou medida Rand em estatística, é uma medida da similaridade entre agrupamentos de dados. Do ponto de vista matemático, o índice de Rand está relacionado à precisão, mas é aplicável mesmo quando os rótulos de classe não são usados.

Denotado por R, o índice Rand é calculado como:

$$R = \frac{a+b}{\binom{n}{2}} \tag{13}$$

em que, a, é o número de vezes que um par de elementos pertence ao mesmo cluster; b é o número de vezes que um par de elementos pertence a grupos de diferentes; e $\binom{n}{2}$: é número de pares não ordenados em um conjunto de n elementos.

Valores próximos de 1 indicam que o agrupamento está bem definido, logo foi agrupado apropriadamente. Valores próximos de 0 indicam que o agrupamento pode ter sido definido de forma errada.

5.3 ACURÁCIA

Quando se fala em tecnologia atualmente, um termo que tem se tornado cada vez comum é o da acurácia. Usada muitas vezes quase como um sinônimo de precisão e eficiência. Em outras palavras, a acurácia serve para ver a porcentagem de acertos do algoritmo de agrupamento.

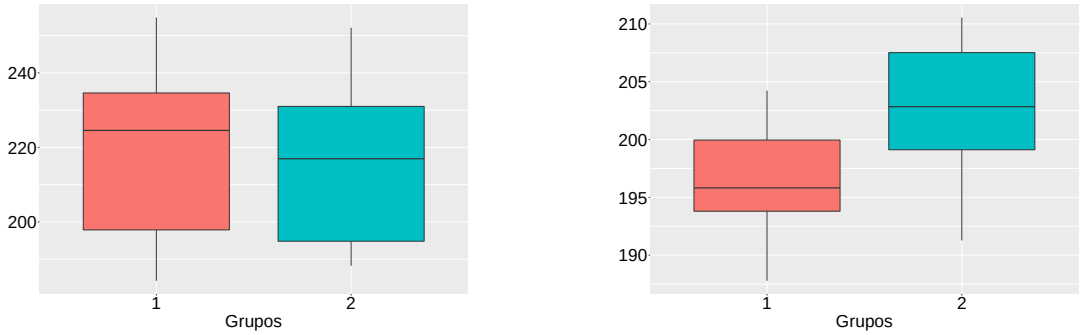
6 APLICAÇÃO E RESULTADOS

Para verificar o comportamento do algoritmo proposto, foram criados dados artificiais de tamanho e forma contendo dois grupos. Nesses dados simulados, gerados a partir da distribuição normal complexa central, propomos dois cenários possíveis: O primeiro para grupos homogêneos[3]; o segundo para dois grupos heterogêneo[4].

[3] Apresenta uma grande semelhanças na estrutura dos dados.
[4] Apresenta uma grande diferença na estrutura dos dados.

São apresentados na Figura 3, os boxplots dos tamanhos dos centroides para cada cenário.

Figura 3 – Boxplots dos tamanhos dos centroides para os dados de cada cenário proposto.



(a) Cenário 1 (b) Cenário 2

Fonte: Autor

6.1 CENÁRIO 1: GRUPOS HOMOGÊNEOS.

No primeiro cenário, ao construirmos o tamanho dos centroides de cada grupo, Figura 3(a), podemos ver o quanto eles são semelhantes. Neste cenário, o algoritmo teve um pouco de dificuldade para realizar o agrupamento. Porém, seus índices interno e externos, além da acurácia, comprovam que o algoritmo teve boa performance.

Tabela 1 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados simulados do cenário 1.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,809	0,495	89,58%

Fonte: Autor

6.2 CENÁRIO 2: GRUPOS HETEROGÊNEO.

No segundo cenário, ao construirmos o tamanho dos centroides de cada grupo, Figura 3(b), podemos ver que os grupo são bastantes diferentes. Nesse cenário, o algoritmo teve melhor desempenho se comparado ao primeiro cenário.

Tabela 2 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados simulados do cenário 2.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,958	0,627	97,92%

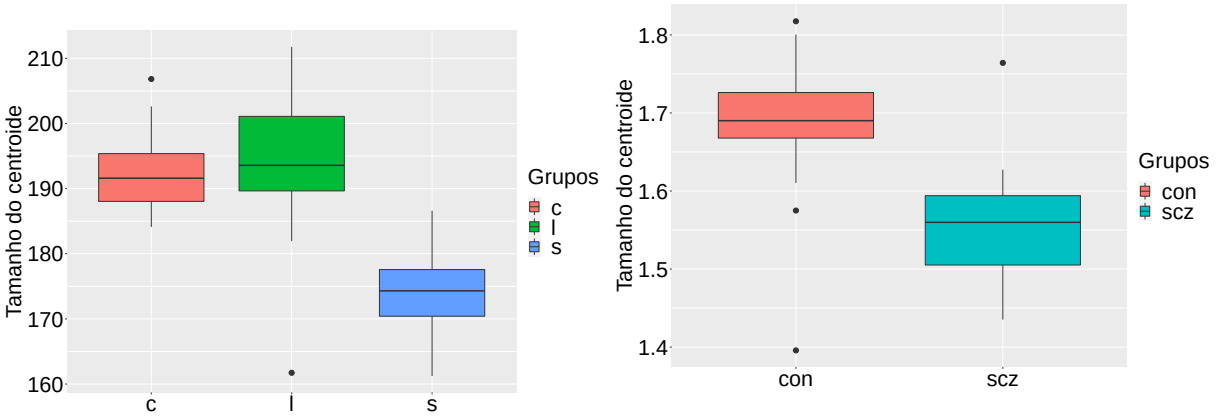
Fonte: Autor

Com isso, concluímos que à medida que o tamanho dos centroides dos grupos diferem, o algoritmo vai tendo mais facilidade em realizar o agrupamento. Além disso, provamos que o algoritmo é eficiente para tratar dados de tamanho e forma.

7 DADOS REAIS

Após a validação com os dados artificiais, aplicaremos o algoritmo *K*-médias modificado para duas bases de dados reais. As mesmas podem ser encontradas no pacote `shapes` do *software* R Core Team (2021)

Figura 4 – Boxplots dos tamanhos dos centroides para os dados reais do pacote `shapes`.



(a) dados das vértebras de camundongos.

(b) dados das ressonâncias magnéticas de pessoas com esquizofrenia.

Fonte: Autor

7.1 VÉRTEBRAS TORÁCICAS T2 DE CAMUNDONGOS.

A base de dados das vértebras torácicas T2 de camundongos possui 6 marcos em 2 dimensões, e possui um total de 76 indivíduos. Esses indivíduos foram classificados em 3 grupos: controle(**c**)=30, grandes(**l**)=23 e pequenos(**s**)=23. Os 6 pontos de referência foram obtidos usando um método semi-automático em pontos de alta curvatura, em que, é mostrado com mais detalhes em Dryden e Mardia (2016).

Para ter uma ideia inicial de como os algoritmos se comportarão, foi feito o boxplot do tamanho dos centroides de cada grupo. Pela Figura 4(a), podemos ver que apenas um grupo difere dos demais, além da presença de *outliers* em dois grupos. Com isso, podemos supor a partir dos resultados dos dados simulados, que o algoritmo terá uma boa performance para realizar os agrupamentos.

Tabela 3 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados das vértebras de camundongos.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,737	0,592	75,0%

Fonte: Autor

A Tabela 3 mostra os resultados do algoritmo, com ela, podemos ver que os resultados a performance de cada índice. De modo geral, os índices interno e externo de validação tiveram bons resultados, indicando assim que os agrupamentos foram bem definidos e posteriormente, validando os agrupamentos.

7.2 RESSONÂNCIA MAGNÉTICA DE PESSOAS COM ESQUIZOFRENIA

A esquizofrenia é um transtorno mental grave que muda o modo como a pessoa pensa, sente e se comporta socialmente. A base de dados das ressonâncias magnéticas de pessoas com esquizofrenia possui 13 marcos em 2 dimensões, e possui um total de 28 indivíduos. Esses indivíduos foram classificados em 2 grupos: controle(**con**)=14 e esquizofrênico(**scz**)=14. Para mais detalhes sobre a base de dados, ver Dryden e Mardia (2016).

Novamente, para ter uma ideia inicial de como os algoritmos se comportarão, foi feito o boxplot do tamanho dos centroides de cada grupo. Pela Figura 4(b), podemos ver que os dois grupos estão bem separados, porém, existe a presença de outliers nos dois grupos. Com isso, podemos supor a partir dos resultados dos dados simulados, que os algoritmos terão uma boa performance para realizar os agrupamentos.

Tabela 4 – Resultado dos índices internos, externos e acurácia para validação dos algoritmos propostos, aplicados aos dados das ressonâncias magnéticas de pessoas com esquizofrenia.

	Índice de Rand	**Índice de Procrustes**	**Acurácia**
K-médias	0,696	0,419	82,1%

Fonte: Autor

A Tabela 4 mostra os resultados do algoritmo, com ela, novamente, podemos ver que os resultados e a performance de cada índice. Os índices interno e externo de validação tiveram bons resultados, indicando assim que os agrupamentos foram bem definidos e posteriormente, validando os agrupamentos.

8 CONCLUSÕES

Neste trabalho, apresentamos uma introdução a análise estatística de formas e classificação não supervisionada no espaço não euclidiano.

No geral, os resultados obtidos mostraram que o algoritmo proposto obteve bom desempenho, para realizar os agrupamentos, usando dados de tamanho e forma.

Se tratando dos dados simulados e reais, o algoritmo foi eficiente para todos os cenários propostos. Para o cenário em que os os grupos eram mais homogêneos, o algoritmo perdeu um pouco de eficiência, para grupos mais heterogêneos o algoritmo foi mais eficiente.

Assim, este trabalho contribuiu para a literatura teórica dos métodos de agrupamento para dados de análise estatística de tamanho e forma.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL, G. J. et al. K-means algorithm in statistical shape analysis. **Communications in Statistics—Simulation and Computation®**, Taylor & Francis, v. 39, n. 5, p. 1016–1026, 2010.

DRYDEN, I. L.; MARDIA, K. V. **Statistical Shape Analysis, with Applications in R. Second Edition.** Chichester: John Wiley and Sons, 2016.

IZENMAN, A. J. **Modern Multivariate Statistical Techniques**. [S.l.]: Springer New York, 2008.

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

KENDALL, D. G. Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. **Bulletin of the London Mathematical Society**, v. 16, n. 2, p. 81–121, 03 1984. ISSN 0024-6093.

LANDAU, S. et al. **Cluster Analysis**. [S.l.]: Wiley, 2011. (Wiley Series in Probability and Statistics). ISBN 9780470978443.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **In 5-th Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.: s.n.], 1967. p. 281–297.

NASCIMENTO, A. et al. Influential observation in complex normal data for problems in allometry. **Communications in Statistics - Theory and Methods**, Taylor and Francis, v. 45, n. 9, p. 2714–2729, 2016.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021.

RAND, W. M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical Association**, [American Statistical Association, Taylor and Francis, Ltd.], v. 66, n. 336, p. 846–850, 1971. ISSN 01621459.