UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

SAUL DE AZEVÊDO SOUZA

MÉTODOS ESTATÍSTICOS PARA DADOS EM ESPAÇOS NÃO EUCLIDIANOS

Recife

2021

SAUL DE AZEVÊDO SOUZA

MÉTODOS ESTATÍSTICOS PARA DADOS EM ESPAÇOS NÃO EUCLIDIANOS

Tese apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de doutor em Estatística.

Área de Concentração: Estatística Aplicada

Orientador (a): Abraão David Costa do Nascimento

Coorientador (a): Getúlio José Amorim do Amaral

Recife

2021

Catalogação na fonte
Bibliotecário Cristiano Cosme S. dos Anjos, CRB4-2290

S729 Souza, Saul de Azevêdo
Métodos estatísticos para dados em espaços não euclidianos / Saul de Azevêdo Souza. – 2021.
109 f.: il., fig.,tab.

Orientador: Abraão David Costa do Nascimento.
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN, Estatística, Recife, 2021.
Inclui referências e apêndices.

1. Estatística Aplicada. 2. Dados direcionais. 3. Dados axiais. 4. Distância estocástica. I. Nascimento, Abraão David Costa do (orientador). II. Título.

310 CDD (23. ed.) UFPE- CCEN 2021 - 120

SAUL DE AZEVEDO SOUZA

MÉTODOS ESTATÍSTICOS PARA DADOS EM ESPAÇOS NÃO EUCLIDIANOS

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Estatística.

Aprovada em: 01 de Julho de 2021.

BANCA EXAMINADORA

Prof. Dr. Abraão David Costa do Nascimento
UFPE

Profª. Drª Lúcia Pereira Barroso
USP

Profª Drª Wenia Valdevino Félix de Lima
UEPB

Profª Drª Maria do Carmo Soares de Lima
UFPE

Prof. Dr. Alex Dias Ramos
UFPE

AGRADECIMENTOS

Agradeço primeiramente à DEUS, pelo seu cuidado para comigo, conduzindo e guiando-me para o melhor caminho. Por me proteger, ajudar-me a enfrentar os obstáculos que a vida me propôs e a tomar as melhores decisões.

Aos meus pais, Sílvio e Maria, pelo cuidado, apoio e dedicação, aos quais dedico todas as minhas conquistas.

Ao meu irmão Sílvio Jr., aos meus avós Lourival e Herenilde, aos meus tios, Sérgio e Sátiro, as minhas tias Marluce e Sônia, e aos meus primos, Aline, Samantha, Sarah, Silvia e Sátiro, pela amizade, apoio e incentivo.

Agradeço a minha namorada Ana Luiza, por seu cuidado, companheirismo, palavras de encorajamento, enfim, por compreender meus momentos de desespero, obrigado.

Aos meus amigos, Jodavid, Daniel, Cesar, Pedro, Alisson, Marina, Adenice, Anny Kerollayny, Wendel, Diogenes, pela amizade, companheirismo, descontração, incentivo e paciência. Por tornar meus dias mais agradáveis e divertidos.

Aos professores Abraão e Getúlio, pela orientação constante, paciência, confiança, amizade e respeito. Pelos valiosos ensinamentos e direcionamentos para a conclusão desta tese.

À Valéria e Michelle, secretárias do Programa de Pós-Graduação em Estatística, pela amizade e orientação em burocracias acadêmicas.

Aos professores Hemílio, Tatiene, Ulisses, Caliandra, Eufrásio, João Agnaldo, pelos conselhos e amizade. Por engrandecer meus conhecimentos ao longo desses anos.

Aos docentes da Banca Examinadora pelas sugestões e contribuições, desde a qualificação do projeto de pesquisa.

A todos os Professores do DE-UFPE, por contribuírem para minha formação acadêmica.

A todos os funcionários do Departamento de Estatística.

À CAPES, pelo apoio financeiro.

RESUMO

Esta tese objetiva primeiramente estudar concentração em dados esféricos através de um novo paradigma, a saber, reduzir o problema na esfera real para o intervalo $[0, 1]$. Os dados esféricos endereçados são em duas frentes: fenômenos direcionais e axiais. Para alcançar o objetivo em questão, duas distribuições são propostas a partir de transformações baseadas em distâncias sobre as distribuições von Mises-Fisher (caso direcional) e Watson (caso axial) reais. As distribuições são denotadas como primeira transformação baseada em distância $(\text{TD}_1(\kappa))$ e segunda transformação baseada em distância $(\text{TD}_2(\kappa))$ para os casos direcional e axial, respectivamente, em que κ é um parâmetro que herda a relação com a concentração dos dados das distribuições esféricas. Algumas propriedades matemáticas para as distribuições TD_1 e TD_2 são discutidas: função geradora de momentos, momentos, curtose, assimetria e matriz de informação de Fisher. Além disso, discussões sobre inferência (pontual e teste de hipóteses) para os parâmetros dos novos modelos são realizadas. Uma vez propostas e estudadas as distribuições, elas são utilizadas como elementos centrais no desenvolvimento de estatísticas de testes para dados direcionais e axiais.Distribuições exatas para estas estatísticas são derivadas. Estudos numéricos, para as distribuições TD_1 e TD_2, apontam que as estimativas de máxima verossimilhança para κ apresentam bons desempenhos mesmo em pequenas amostras. Para o modelo TD_1, verificam-se que: (i) os testes de hipóteses clássicos (escore, Wald e razão de verossimilhanças) são, em geral, conservadores quanto ao nível pré-especificado em altas concentrações; (ii) o teste escore foi o mais conservador; (iii) o teste Wald foi o mais liberal para pequenos valores de κ. Para o modelo TD_2, observam-se que: (i) o teste da razão de verossimilhanças tende a ser mais liberal para $\kappa > 1$; (ii) os testes Wald e escore são mais conservadores para $\kappa > 0$. Duas aplicações são feitas para ilustrar as propostas em dados esféricos. Resultados mostram que o uso dos paradigmas propostos conseguem detectar de modo simples (isto é, transferindo o problema de uma esfera real para o intervalo $[0, 1]$) e eficiente alta concentração em amostras esféricas. É sabido que a média é uma medida de locação influenciada por valores destoantes do conjunto tanto no contexto uni quanto multivariado em espaços Euclidianos. Esse problema também se verifica para variedades estocásticas, como o espaço das pré-formas ou a hiperesfera complexa. A segunda parte desta tese se dedica a proposta de uma alternativa robusta a média extrínseca de Fréchet, que tem fórmula analítica intratável. Fórmulas matemáticas para computar a mediana extrínseca projetada e procedimentos para detecção de outliers, baseados nessa medida, são apresentados. Estudos

numéricos por simulação de Monte Carlo são realizados para quantificar a robustez da nova mediana em termos da distribuição Bingham complexa para o caso de formas planares. Os resultados mostraram que a mediana proposta é mais robusta que a forma média, principalmente para pequenos tamanhos de amostras. Uma aplicação aos dados de microfósseis ilustra o uso da mediana que propomos.

Palavras-chaves: Dados direcionais; dados axiais; distância estocástica; testes de hipóteses; critérios de concentração; variedade; medida extrínseca; medida intrínseca.

ABSTRACT

Firstly, this thesis aims to study high phenomena concentration on real spherical data through a new paradigm, say to reduce the sphere problem to the interval $[0, 1]$. The spherical data are commonly addressed on two branchs: directional and axial phenomena. To achieve the paradigm we propose, two distributions are proposed in terms of distance-based transformations from the real von Mises-Fisher (directional case) and Wason (axial case) distributions. These new distributions are denoted as the first distance-based transformation ($\text{TD}_1(\kappa)$) and second distance-based transformation ($\text{TD}_2(\kappa)$) for projected directional and axial data, respectively, being κ a parameter that inherits the relationship with the concentration of data from spherical distributions. Some mathematical properties for the TD_1 and TD_2 distributions are discussed: moment generating function, moments, kurtosis and asymmetry and Fisher information matrix. In addition, essays about statistical inference (in the pontual and hypothesis test contexts) for the parameters for the new models are carried out. Once the distributions have been studied and proposed, they are used as core parts in the development of test statistics for directional and axial data The exact distributions of these statistics are derived from the TD_1 and TD_2 laws. Numerical studies for both distributions show that the maximum likelihood estimates for κ achieve good performances, even in small samples. For the TD_1 model, it is found that: (i) the classic hypothesis tests (score, Wald and likelihood ratio) are generally conservatives with respect to the pre-specified nominal level in high concentration; (ii) the score tests is the most conservative; (iii) the Wald test is the most liberal for small values of κ. For the TD_2 law, it is noticeable that: (i) the likelihood ratio test tends to be more liberal for $\kappa > 1$: (ii) the Wald and score tests are more conservatives for $\kappa > 0$. Two applications to real data are made to illustrate the proposals in spherical data. Results show that the use of the proposed paradigms is able to detect in a simple (transferring the problem of a real sphere to the interval $[0, 1]$) and efficient way high concentration in spherical samples. It is known that the mean is a location measure that is influenced by the different values of the set (say outliers), in both the uni and multivariate contexts of the Euclidean spaces. This problem also occurs for stochastic manifolds, such as the pre-shape space or the complex hypersphere. Second, this thesis also proposes an alternative to Fréchet extrinsic mean, which is analytically intractabel. Mathematical formulae to compute the projected extrinsic median are presented. Additionaly, a method for detecting outliers is introduced, based on the projected extrinsic median. Numerical studies by Monte Carlo experiments are performed to quantify the robustness of the median

measure in terms of the complex Bingham distribution for plannar shapes.Results show that the proposed median is more robust than the mean shape, mainly for small sample sizes. Finally, an application to microfossil data is made to illustrate the proposed median.

Keywords: Direction data; axial data; stochastic distance; hypothesis tests; concentration criteria; manifold; extrinsic measure; intrinsic measure.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

$\mathsf{vMF}(\boldsymbol{\mu}, \kappa)$	Distribuição von Mises-Fisher
$\mathsf{TD}_1(\kappa)$	Primeira distribuição de probabilidade baseada em distância
$\mathsf{TD}_2(\kappa)$	Segunda distribuição de probabilidade baseada em distância
$\mathsf{W}(\boldsymbol{\mu}, \kappa)$	Distribuição Watson
U_S	Distribuição uniforme na esfera
$\mathbb{N}$	Números naturais ou inteiros positivos
$\mathbb{R}$	Numeros reais
$\mathbb{R}^+$	Números reais positivos
$\boldsymbol{X}$	Vetor unitário aleatório
$f(x)$	Função densidade de probabilidade
$F(x)$	Função de distribuição acumulada
$M_X(t)$	Função geradora de momento
$\mathcal{M}$	Variedade (Manifold)
$B(\cdot, \cdot)$	Função beta
$f_a(\cdot)$	Função angular
$I_p(\cdot)$	Função Bessel modificada do primeiro tipo de ordem p
$\Gamma(\cdot)$	Função gama
$\mathsf{erf}(\cdot)$	Função de erro
$M(\cdot, \cdot, \cdot)$	Função de Kummer
$\mathrm{vec}(\cdot)$	Operador de vetorização

SUMÁRIO

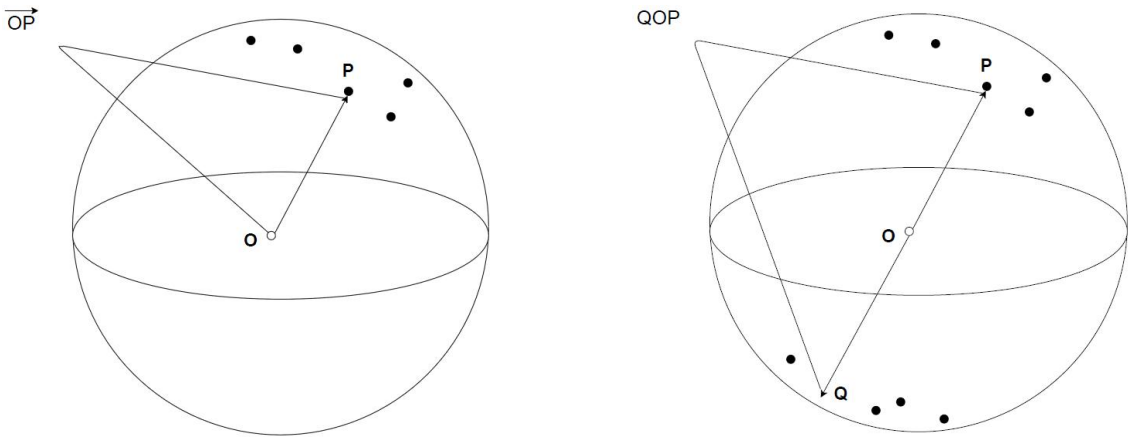
1 INTRODUÇÃO

A análise estatística de dados esféricos é amplamente utilizada em diversas áreas da ciência, tais como: Astronomia (HUO; ZHOU, 2013), Climatologia (AZNAR et al., 2012), Astrofísica (ZHANG et al., 2018), Oceanografia (MUJICA; NAVA; VARGAS, 2014), entre outras . Tratam-se de observações que são direções ou eixos, tendo como suporte uma hiperesfera unitária em $\mathbb{R}^p$ para $p \geq 3$ (LEY; VERDEBOUT, 2017). São exemplos de dados esféricos: (i) direções de ventos (AMIRI ABOUBACAR; THIAM; VERDEBOUT, 2016); (ii) direções de eixos ópticos de cristais de quartzo (FISHER; LEWIS; EMBLETON, 1993). A depender de sua natureza, estes dados podem ser classificados em dois tipos:

- Dados direcionais: denominado linha dirigida ou vetor direcionado, o vetor $\overrightarrow{OP}$ é determinado pela representação de um ponto P na superfície de uma esfera unitária centrada em O (Figura 1a) Fisher, Lewis e Embleton (1993);
- Dados axiais: denominado linha não dirigida ou eixo, os pares de pontos P e Q são definidos em extremidades opostas distanciadas pelo diâmetro da esfera (Figura 1b) Fisher, Lewis e Embleton (1993).

Figura 1 – Geometria dos dados direcionais e axiais.

(a) Dados direcionais. (b) Dados axiais.



Fonte: O autor (2021).

Existem muitos sistemas de coordenadas para representar pontos esféricos e métodos para obter suas projeções sobre o plano. Devido a variedade de aplicações, cada situação parece exigir uma espécie de metodologia de representação diferente. Na prática, a análise é feita a

partir de dados reais que são transformados em sistemas de coordenadas polares ou em seus respectivos cossenos direcionais (FISHER; LEWIS; EMBLETON, 1993).

As análises estatísticas de dados esféricos ganharam muito destaque com os avanços ocorridos na área de Paleomagnetismo (CARDOZO; ALLMENDINGER, 2013), responsável por estudar o campo geomagnético registrado na magnetização das rochas. Como o campo magnético gerado é um campo vetorial é possível representar suas medições por vetores unitários desenhados nos três eixos ortogonais, centrados na esfera unitária. Seja devido a composição ou a heterogeneidade de condições ocorridas no processo de formação dessas rochas, o magnetismo remanescente apresenta uma dispersão tão alta que são necessários métodos especiais para seu estudo (FISHER, 1953).

Dessa forma, para lidar com este tipo de dado, Fisher (1953) sugeriu uma densidade de probabilidade na esfera, proporcional a

$$\exp\left[\kappa\,\cos(\theta)\right],$$

em que θ é o deslocamento angular a partir da direção média verdadeira e κ é o parâmetro de concentração. A concentração dos pontos esféricos, controlada pelo parâmetro κ, é um dos fenômenos mais importantes na teoria de dados esféricos. Por exemplo, Watson (1984) estudou ensaios de eventos de concentração na distribuição von Mises-Fisher (vMF), apresentando resultados assintóticos e testes de hipóteses para alta concentração. Ko (1992) propôs um processo de estimação robusto baseado em desvios medianos para o parâmetro de concentração da vMF. Assumindo alta concentração, Chikuse (2003) derivou e avaliou o comportamento assintótico de algumas estatísticas de teste baseadas na matriz de médias amostrais para a distribuição vMF. Figueiredo (2006) considerou a ANOVA bidirecional aninhada para a distribuição Watson (W) sob alta concentração. Figueiredo (2009) sugeriu testes de hipóteses para igualdade de parâmetros direcionais para diferentes concentrações da distribuição W. Fisher (1986) tratou de testes de hipóteses para comparar a dispersão de dados na esfera para as distribuições W e vMF.

Outro tipo de dado presente em problemas reais, tal como em reconhecimento de padrões (PLAMONDON; SRIHARI, 2000), está no espaço das pré-formas, a saber, uma hiperesfera complexa. As principais referências para análise de forma são os livros de Mardia e Jupp (1999), Small (1990) e Dryden e Mardia (2016) que detalham todo o marco teórico para os métodos da segunda parte desta tese. Esta abordagem se refere a um conjunto de procedimentos para lidar com dados geométricos. Em geral, a análise começa registrando a imagem de um objeto

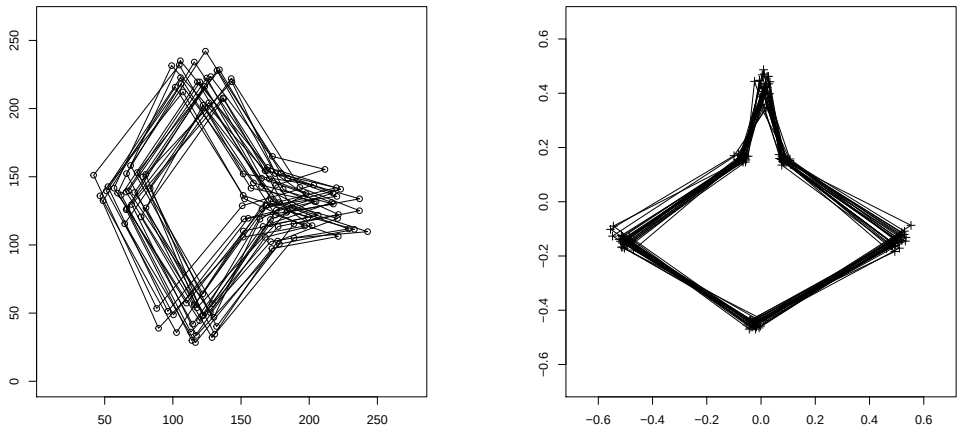
e então se definem pontos no contorno dele que são chamados de marcos anatômicos (*landmarks*). Estes objetos podem ser, por exemplo, exames médicos, imagens de animais e muitos outros (MARDIA; JUPP, 1999). Como ilustração, a Figura 2a apresenta as vértebras torácicas de camundongos definidas a partir de 6 *landmarks*. Por outro lado, na Figura 2b são removidos os efeitos de locação e translação, resultando em observações centralizadas (DRYDEN; MARDIA, 2016).

Definição 1 (Dryden e Mardia (2016)) *A forma de um objeto é definida como toda informação geométrica que permanece quando os efeitos de translação, rotação e escala são removidos.*

Um dos modelos mais importantes para descrever dados de pré-forma (ou planar) é a distribuição Bingham complexa. Ela é um caso especial da distribuição Bingham real, que por sua vez possui uma relação com a distribuição Watson (KENT, 1994). As duas últimas distribuições reais são comumente utilizadas para modelar dados axiais.

Figura 2 – Vértebras torácicas de camundongos. São observados 6 *landmarks* em cada vértebra.

(a) Possui rotação, escala e translação. (b) Mantém a rotação.



Fonte: O autor (2021).

1.1 PROBLEMÁTICA

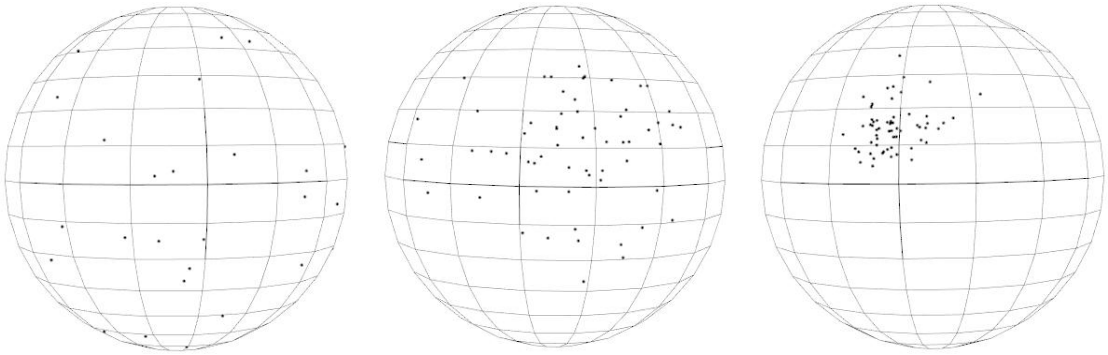
O primeiro problema que esta tese trata é a identificação de que uma aplicação envolvendo dados esféricos consiste em um estado de alta concentração. Para ilustrar, a Figura 3 mostra

o comportamento de 60 observações geradas pela distribuição von Mises-Fisher, com média direcional $\boldsymbol{\mu} = (0.4150,\ 0.3016,\ 1.1082)$ e concentração $\kappa \in \{0,\ 10,\ 50\}$. O parâmetro de concentração κ é responsável pela dispersão dos pontos esféricos. Quando κ se aproxima de zero, a distribuição dos pontos tende a uma distribuição uniforme na esfera. Em termos interpretativos, κ indica a proximidade dos resultados da distribuição de Fisher ($\mathcal{F}$). Além disso, boa parte dos resultados assintóticos da teoria de dados esféricos assumem que a condição de "alta concentração" é verificada (MARDIA; JUPP, 1999).

Na prática, o problema surge sobre o que significa um κ grande tal que o fenômeno de alta concentração é verificado. Adicionalmente, as análises visuais são fortemente influenciadas pela percepção do indivíduo. Logo, é importante saber qual critério pode ser utilizado para categorizar uma base de dados como altamente concentrada.

Figura 3 – Gráfico esférico de 60 observações provenientes da distribuição von Mises-Fisher.

(a) $\kappa = 0$. (b) $\kappa = 10$. (c) $\kappa = 50$.



Fonte: O autor (2021).

O segundo problema diz respeito ao processo de estimação da média extrínseca de Fréchet e de mecanismos para detectar *outliers* em dados de pré-forma (isto é, na esfera complexa). É conhecido que a média é uma medida de locação bastante influenciada por *outliers* no conjunto de dados, tanto no contexto univariado quanto no contexto multivariado em espaços Euclidianos. Além disso, esse problema persiste para variedades estocásticas, como o espaço das pré-formas ou a hiperesfera complexa. Dessa forma, surge a necessidade de procurar por abordagens que forneçam melhores estimativas e métodos para detectar *outliers*.

1.2 OBJETIVO

Os objetivos dessa tese são divididos em duas partes, referentes as análises de dados nas esferas real e complexa. Primeiro, no âmbito da esferea real, é proposta uma maneira mais simples de estudar o fenômeno da concentração em dados esféricos. Nesta proposta, ao invés de realizar uma análise multivariada dos dados, envolvendo vetores unitários no espaço tridimensional, é construída uma medida de distância capaz de capturar a concentração destes pontos.

Aqui, lidamos com a análise univariada de uma medida de distância compacta, definida em $[0, 1]$. Para tanto, são propostas duas distribuições de probabilidade baseadas em distância para descrever a dispersão dos pontos esféricos. O primeiro modelo, é construído sob a suposição de distribuição von Mises-Fisher, muito comum no estudo de dados direcionais, permitindo descrever observações que se concentram nos polos. Por outro lado, o segundo modelo, é introduzido sob a suposição de distribuição Watson que, diferente da von Mises-Fisher, permite modelar observações que se concentram nos polos e na região equatorial.

Como objetivos específicos desta primeira parte da tese, têm-se:

- ✓ Construir duas novas distribuições de probabilidade baseadas em distância, sob a suposição de distribuição von Mises-Fisher ou Watson, denotadas como $\mathrm{TD}_1(\kappa)$ (dados direcionais) e $\mathrm{TD}_2(\kappa)$ (dados axiais), respectivamente;
- ✓ Definir uma medida de distância compacta em $[0, 1]$ capaz de capturar o grau de concentração em pontos esféricos;
- ✓ Realizar estudos de simulação envolvendo estimação e testes de hipóteses sobre o parâmetro κ dos modelos $\mathrm{TD}_1(\kappa)$ e $\mathrm{TD}_2(\kappa)$;
- ✓ Definir estatísticas para testar o fenômeno de alta concentração em dados direcionais e axiais e derivar suas distribuições exatas.

Também, temos como meta desenvolver uma alternativa robusta, de implementação simples, para a média extrínseca de Fréchet, que é uma medida de locação comum ao tratar com variedades estocásticas em espaços não Euclidianos. Para tanto, em virtude da intratabilidade analítica e computacional da média extrínseca, é proposto um método baseado na mediana extrínseca de Fréchet. Este último, por sua vez, é um método robusto bastante utilizado em

análise multivariada. Adicionalmente é apresentado um método de detecção de *outliers*, que utiliza a mediana extrínseca projetada (proposta nesta tese) como referência.

Como objetivos específicos para esta segunda metade da tese, têm-se:

- ✓ Propor uma alternativa robusta à média extrínseca de Fréchet e realizar estudos numéricos para comparar as duas abordagens;
- ✓ Propor um método para detectar *outliers* que usa a mediana extrínseca proposta como referência.

O restante da tese está organizado da seguinte maneira. No Capítulo 2, são apresentados os referenciais teóricos para análise de dados na esfera real e complexa. Aqui, são destacadas as principais distribuições de probabilidade e ferramentas, descritivas e gráficas, utilizadas nestas duas metodologias. No Capítulo 3, é introduzida uma nova abordagem estatística para detectar alta concentração em dados direcionais. Especificamente, é apresentado o novo modelo $\mathrm{TD}_1(\kappa)$. São realizados estudos numéricos, envolvendo estimação pontual e testes de hipóteses. No Capítulo 4, é discutida a nova abordagem estatística para dados axiais. O modelo $\mathrm{TD}_2(\kappa)$ é apresentado e são realizados estudos de simulação, envolvendo estimação pontual e testes de hipóteses. No Capítulo 5, é apresentada a proposta de uma medida mais robusta para dados de pré-forma na esfera complexa, baseada na mediana extrínseca de Fréchet. Adicionalmente, é abordado um novo método para detectar *outliers* baseado nos procedimentos de Hoaglin e Iglewicz (1987). Finalmente, o Capítulo 6 apresenta as conclusões e as pesquisas futuras associadas aos temas desta tese.

1.3 CONTRIBUIÇÕES

Esta tese avança no tratamento de dados esféricos (direcionais e axiais) através dos seguintes resultados:

- ✓ Como primeiro produto desta tese, submetemos um artigo intitulado *A new distance-based distribution: Detecting concentration in directional data* à revista Anais da Academia Brasileira de Ciências. Uma versão resumida deste trabalho foi aceita pelo CNMAC 2021 (XL Congresso Nacional de Matemática Aplicada e Computacional).

- ✓ Como segundo produto desta tese, submetemos um artigo intitulado *Detecting high concentration on Watson axial data from a distance-based transformation* à revista Journal of Computational and Applied Mathematics.

- ✓ Como terceiro produto desta tese, estamos trabalhando num artigo no tema *Novos métodos baseados na mediana extrínseca projetada de Fréchet*. Este tópico tem sido desenvolvido em colaboração com uma equipe do Australian National University (College of Business and Economics).

1.4 PLATAFORMA COMPUTACIONAL

O ambiente R é uma plataforma computacional de código aberto (open source) que permite: construção de gráficos de alta qualidade, implementação (ou uso) de técnicas estatísticas bem definidas e calcular funções especiais, como as funções de Bessel e hipergeométricas, por meio de pacotes que são atualizados continuamente. Além disso, o R recebe contribuições de diversos colaboradores ao redor do mundo, o que diversifica as áreas do conhecimento (R Core Team, 2013).

Os procedimentos computacionais usados nesta tese foram desenvolvidos utilizando o software R. Foram desenvolvidas funções específicas (rotinas) para atingir os objetivos propostos. Adicionalmente, utilizamos o SageMath e o wxMaxima para resolver algumas expressões complicadas ao longo desta tese. O primeiro, é um sistema de software matemático de código aberto e apresenta uma linguagem comum baseada em Python. O segundo, é uma interface baseada em documento para o sistema de álgebra computacional.

2 REFERENCIAL TEÓRICO

Neste Capítulo, são introduzidas ferramentas básicas para explorar e visualizar dados em espaços não euclidianos, com ênfase nas esferas real e complexa. É apresentado o conceito de variedade e são introduzidas as abordagens de análise de dados esféricos e de forma, destacando as principais distribuições de probabilidade.

2.1 INTRODUÇÃO A DADOS NÃO EUCLIDIANOS: DADOS DIRECIONAIS E AXIAIS COMO VARIEDADES

Ao longo desta tese, serão apresentados métodos para análise estatística de dados que surgem em variedades. Uma variedade, $\mathcal{M}$, trata-se de um espaço, com uma estrutura diferenciável, que pode ser visto localmente como um espaço Euclidiano de dimensão d (BHATTACHARYA; BHATTACHARYA, 2012). Seja $\gamma(t) \in \mathcal{M}$ uma curva diferenciável em $\mathcal{M}$ com $t \in \mathbb{R}$ e $\gamma(0) = p$ (BHATTACHARYA; BHATTACHARYA, 2012). O vetor tangente em p é dado por

$$\boldsymbol{\gamma}'(0) = \lim_{t \to 0} \frac{\partial \gamma}{\partial t}.$$

O espaço tangente de $\mathcal{M}$ em p, denotado por $T_p(\mathcal{M})$, é o conjunto de todos os vetores tangentes $\boldsymbol{\gamma}'(0)$ para todas as curvas que passam por p. Naturalmente, temos certa dificuldade em visualizar a curvatura de conjuntos ou espaços que não podem ser aplicados no espaço Euclidiano tridimensional. Contudo, as variedades diferenciaveis não sofrem com esse problema, pois elas se assemelham localmente ao espaço Euclidiano em três dimensões (BHATTACHARYA; BHATTACHARYA, 2012). Um ponto interessante é que geodésicas (ou seja, a menor distância que une dois pontos) podem ser construídas se a variedade $\mathcal{M}$ apresentar conexão afim, isto é, pode-se introduzir mecanismo para conectar espaços tangentes próximos. Logo, uma variedade conectada que possui produto interno positivo definido em cada espaço tangente $T_p(\mathcal{M})$ é chamado de variedade Riemanniana $\mathcal{M}$ (BHATTACHARYA; BHATTACHARYA, 2012). Aqui, o estudo da concentração dos pontos se torna relevante, pois em qualquer variedade diferenciável pode existir um confundimento entre pontos que estão sobre a variedade, $\mathcal{M}$, e pontos que estão no espaço tangente, $T_p(\mathcal{M})$, aonde os métodos tradicionais para análise de dados em espaços Euclidianos, como testes de hipóteses, tendem a funcionar bem.

Por exemplo, o espaço de formas de k *landmarks* em duas dimensões (ilustrando um objeto em uma imagem com k pontos destacados em seu contorno) é tratado como uma

variedade Riemanniana. O livro de Bhattacharya e Bhattacharya (2012) é uma das principais referências para o uso de métodos não paramétricos em variedades. Os autores apresentam abordagens para calcular a média extrínseca e intrínseca de Fréchet. Uma das variedades mais importantes e talvez mais simples se refere ao espaço de todas as possíveis direções em $\mathbb{R}^{p+1}$. Naturalmente, este espaço pode ser identificado como uma esfera unitária. Aqui, as análises de dados direcionais, axiais e de forma encontram diversas aplicações (BHATTACHARYA; BHATTACHARYA, 2012).

A análise estatística de dados esféricos possui uma série de ferramentas e metodologia própria para lidar com variáveis aleatórias assumindo valores em um espaço p-dimensional (SAU; RODRIGUEZ, 2017). Como o espaço amostral é uma esfera, a principal dificuldade ao tratar estes dados é lidar com a sua curvatura. Isto porque a esfera é um exemplo de variedade não linear. Logo, os métodos tradicionais para analisar dados univariados ou multivariados não podem ser empregados (LEY; VERDEBOUT, 2017; MARDIA; JUPP, 1999).

Considere um caso mais simples como o círculo e tome as medidas circulares $10°$, $30°$ e $350°$. Ser (2014) mostrou que a média aritmética desses ângulos, igual a $130°$, sugere a direção geográfica sudeste. Por outro lado, quando os pontos são localizados no círculo, estes ângulos indicam uma direção média na direção norte. Vale destacar que a teoria de estatísticas circulares é análoga para estatísticas esféricas, sendo este um exemplo mais prático. Portanto, os procedimentos de análise de dados esféricos são mais adequados para lidar com medidas angulares.

2.1.1 Análise de dados esféricos

Sejam $P_1, \ldots, P_n$ uma amostra aleatória de pontos na superfície da esfera unitária centrada em O. Na prática, os dados esféricos buscam resumir as orientações médias de um vetor no espaço, denotado por $\overrightarrow{OP}$. Este tipo de representação requer o uso de um sistema de coordenadas polares, dado por

$$(\theta, \phi),$$

em que $0° \leq \theta \leq 180°$ é o ângulo de colatitude e $0° \leq \phi < 360°$ é o ângulo de longitude. Os pontos esféricos podem ser representados nos três eixos ortogonais $(x,\ y,\ z)$, definidos pelos cossenos direcionais

$$x = \text{sen}(\theta)\cos(\phi), \quad y = \text{sen}(\theta)\text{sen}(\phi) \quad \text{e} \quad z = \cos(\theta),$$

que são as contribuições de cada eixo para a formação de uma vetor unitário, diga-se $\boldsymbol{p}_i^\top = (x_i,\ y_i,\ z_i)$. É possível definir um vetor resultante aos n vetores unitários $\overrightarrow{OP_1}, \ldots, \overrightarrow{OP_n}$ como

$$(\hat{x}, \hat{y}, \hat{z}) = \left(\frac{S_x}{R}, \frac{S_y}{R}, \frac{S_z}{R}\right),$$

em que $R = \sqrt{S_x^2 + S_y^2 + S_z^2}$, $S_x = \sum_{i=1}^{n} x_i$, $S_y = \sum_{i=1}^{n} y_i$ e $S_z = \sum_{i=1}^{n} z_i$.

O ângulo ψ (em radianos) formado entre os vetores unitários $\overrightarrow{OP_1}$ e $\overrightarrow{OP_2}$ é obtido a partir do produto interno

$$\cos\psi = x_1x_2 + y_1y_2 + z_1z_2 = (x_1\ y_1\ z_1)\begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix},$$

com $0^\circ \leq \psi \leq 180^\circ$ (FISHER; LEWIS; EMBLETON, 1993).

As coordenadas polares, $(\hat{\theta}, \hat{\phi})$, do vetor resultante podem ser obtidas através de transformações trigonométricas dos cossenos direcionais, $\hat{x} = \text{sen}(\hat{\theta})\ \cos(\hat{\phi})$, $\hat{y} = \text{sen}(\hat{\theta})\,\text{sen}(\hat{\phi})$ e $\hat{z} = \cos(\hat{\theta})$, definidas como

$$\hat{\phi} = \arctan(\hat{y}/\hat{x}) \quad \text{e} \quad \hat{\theta} = \arccos(\hat{z}),$$

em que $(\hat{\theta}, \hat{\phi})$ é a média direcional. Fisher, Lewis e Embleton (1993) descrevem o comportamento do ângulo $\hat{\phi}$ para diferentes valores dos cossenos direcionais do vetor resultante. É mostrado que

$$\begin{aligned}
0^\circ < \hat{\phi} < 90^\circ \quad &\text{se} \quad \hat{x} > 0,\ \hat{y} > 0; \\
90^\circ < \hat{\phi} < 180^\circ \quad &\text{se} \quad \hat{x} < 0,\ \hat{y} > 0; \\
180^\circ < \hat{\phi} < 270^\circ \quad &\text{se} \quad \hat{x} < 0,\ \hat{y} < 0; \\
270^\circ < \hat{\phi} < 360^\circ \quad &\text{se} \quad \hat{x} > 0,\ \hat{y} < 0.
\end{aligned}$$

Sejam $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ uma amostra aleatória proveniente de uma distribuição esférica tal que $\boldsymbol{p}_i = [x_i, y_i, z_i]^\top$. Então, nos próximos capítulos desta tese, construiremos uma metodologia estatística sobre uma medida de distância $d(\cdot, \cdot)$ dada por

$$d_i = d_i(\boldsymbol{p}_i, \widehat{\boldsymbol{\mu}}) = 1 - (\mathbf{p}_i^\top \widehat{\boldsymbol{\mu}}\ \widehat{\boldsymbol{\mu}}^\top \boldsymbol{p}_i), \tag{2.1}$$

em que

$$\boldsymbol{p}_i = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} \quad \text{e} \quad \hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix}$$

são, respectivamente, os cossenos direcionais dos vetores unitários e do vetor resultante. A medida de distância definida em (2.1) assume valores em $[0, 1]$ e quanto mais próximo de zero maior é a proximidade dos pontos esféricos sobre uma direção preferida. Discutiremos que é possível estudar fenômenos de alta concentração em dados esféricos através desta transformação e que ela tem uma relação geométrica com rotação simétrica e coeficientes do vetor de projeção.

2.2 NATUREZA DE DADOS DE PRÉ-FORMA

A análise estatística de formas se refere a um conjunto de procedimentos utilizados para lidar com dados geométricos. Esta abordagem permite aos pesquisadores realizar estudos sobre a forma ou tamanho-e-forma de objetos. As aplicações podem ser abordadas em áreas como: Geologia, Farmácia, Medicina e análise de imagens (DRYDEN; MARDIA, 2016). Esta abordagem, por exemplo, pode ser empregada rotineiramente na tomada de decisões sobre reconhecimentos faciais ou diagnósticos de doenças a partir de imagens (DRYDEN; MARDIA, 2016).

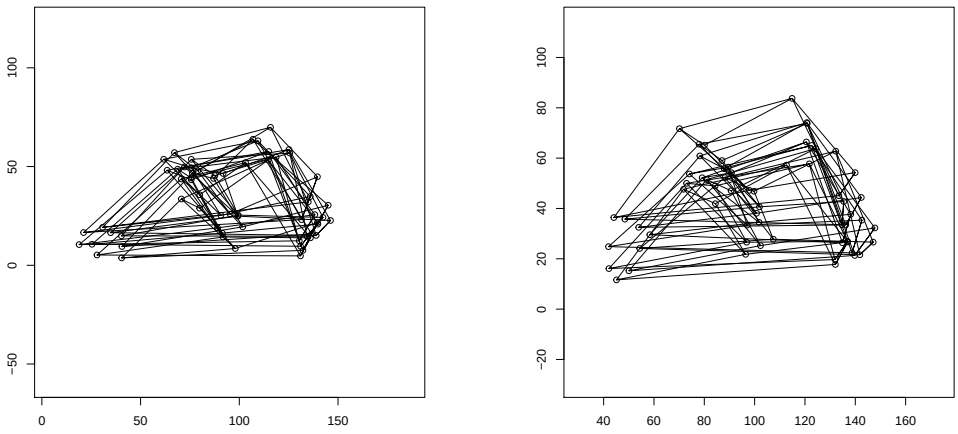
Kendall (1977) definiu a palavra forma, utilizada no cotidiano, para designar o contorno de objetos. Ou seja, toda informação geométrica que permanece ao remover os efeitos de locação, rotação e escala de um objeto. Por exemplo, na biologia a forma dos crânios de macacos consiste de toda informação geométrica que permanece quando o crânio é transladado, redimensionado e rotacionado em um sistema de coordenadas arbitrárias. Logo, dois crânios têm a mesma forma se corresponderem exatamente ou se as diferenças forem estatisticamente negligíveis. (DRYDEN; MARDIA, 2016).

As Figuras 4a e 4b se referem, respectivamente, a 9 crânios de macacos machos e fêmeas, provenientes da espécie *Macaca fascicularis*, representados a partir de 7 *landmarks*. Esse é um exemplo de dado geométrico, disponível no pacote shapes do software R. Eles foram inicialmente coletados para estudar as diferenças significativas do sexo na formação dos crânios de macacos dessa espécie (DRYDEN; MARDIA, 2016). Aqui, os *landmarks* são um conjunto finito

de pontos em um objeto, definidos segundo um sentido biológico. Basicamente, eles permitem descrever a forma de um objeto.

Figura 4 – Dados do crânio dos macacos: 7 *landmarks* de 18 indivíduos (9 crânios de macho e fêmea).

(a) Crânio do macho. (b) Crânio da fêmea.



Fonte: O autor (2021).

2.2.1 Análise estatística de pré-formas em 2 dimensões

Qualquer objeto com q dimensões pode ser representado por um conjunto de k *landmarks* apropriados. Basicamente, existem 3 tipos de *landmarks*: científico, matemático e pseudo-*landmarks*. Eles são comumente utilizados em diferentes áreas da ciência (MARDIA; JUPP, 1999).

- *Landmark* científico: são pontos de referência, com correspondência significativa entre os objetos, atribuído por um especialista.
- *Landmark* matemático: são pontos determinados a partir de propriedades matemáticas ou geométricas.
- Pseudo-*landmarks*: são pontos localizados entre os *landmarks* matemáticos e científicos.

Quando os objetos são identificados em duas dimensões, ou seja, $q = 2$, as informações dos *landmarks* podem ser armazenadas em uma matriz $\boldsymbol{Y}$ de coordenadas cartesianas, dada

por

$$\boldsymbol{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} \\ \vdots & \vdots \\ y_{k,1} & y_{k,2} \end{pmatrix},$$

em que $\boldsymbol{Y}$ é a matriz de configuração com dimensão $k \times 2$. Neste cenário, é possível expressar a matriz de configuração por um vetor complexo

$$\boldsymbol{z}^0 = (y_{1,1} + iy_{1,2}, \ldots, y_{k,1} + iy_{k,2})^\top,$$

o que simplifica os cálculos (AMARAL; WOOD, 2010). Aqui, $\boldsymbol{z}^0$ se refere as coordenadas complexas dos *landmarks*.

O efeito de translação ou locação é determinado pela adição de um vetor constante a cada coordenada de um *landmark*. Uma maneira de remover este efeito é pré-multiplicar o vetor complexo $\boldsymbol{z}^0$ pela submatriz de Helmert, $\boldsymbol{H}$, em que sua i-ésima linha é dada por $(h_j, \cdots, h_j, -jh_j, 0, \cdots, 0)$, para uma j-ésima coluna, e $h_j = -\left[j(j+1)\right]^{-1/2}$.

A submatriz de Helmert, $\boldsymbol{H}$, é definida como a matriz de Helmert, $\boldsymbol{H}^F$, de ordem $(k-1)\times k$ desconsiderando a primeira linha. A matriz $\boldsymbol{H}^F$ possui na primeira linha todos os elementos iguais a $1/\sqrt{k}$ (MARDIA; JUPP, 1999). Para $k = 3$, $\boldsymbol{H}^F$ é dado por

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

Logo, é possível calcular

$$\tilde{\boldsymbol{z}} = \boldsymbol{H}\boldsymbol{z}^0,$$

em que $\tilde{\boldsymbol{z}}$ é a coordenada dos *landmarks* Helmetizados. Perceba que $\tilde{\boldsymbol{z}}$ é uma matriz $(k-1)\times 1$.

A remoção do efeito de escala é obtida após normalização de $\tilde{\boldsymbol{z}}$,

$$\boldsymbol{z} = \frac{\tilde{\boldsymbol{z}}}{|\tilde{\boldsymbol{z}}|} = \frac{\tilde{\boldsymbol{z}}}{(\tilde{\boldsymbol{z}}^*\tilde{\boldsymbol{z}})^{1/2}} = \frac{\boldsymbol{H}\boldsymbol{z}^0}{\{(\boldsymbol{H}\boldsymbol{z}^0)^*\boldsymbol{H}\boldsymbol{z}^0\}^{1/2}},$$

em que $|\cdot|$ denota a norma de um vetor complexo, * denota o transposto conjugado complexo e $\boldsymbol{H}$ refere-se a submatriz de Helmert. Por definição, a pré-forma, $\boldsymbol{z}$, de uma matriz de configuração $\boldsymbol{Y}$ pode ser obtida após a remoção das informações de locação e escala. Ou

seja, a pré-forma é uma forma que mantém a informação de rotação. Logo, a forma, $[z] = \left\{e^{i\theta}\boldsymbol{z} : 0 \leq \theta < 2\pi\right\}$, da pré-forma $\boldsymbol{z}$ é definida como uma classe de equivalência (AMARAL; WOOD, 2010), não sendo uma informação manipulável na prática. Assim, costuma-se trabalhar com as pré-formas.

Um problema importante em análise estatística de forma é definir a forma média de uma amostra aleatória de configurações (AMARAL; WOOD, 2010). Considere a configuração aleatória $\boldsymbol{z}^0$ correspondente a pré-forma $\boldsymbol{z}$, então, a forma média amostral *full Procrustes*, que também é uma forma média extrínseca, é dada por

$$[\hat{\boldsymbol{m}}^{\mathbf{0}}] = \left\{e^{i\theta}\hat{\boldsymbol{m}}_0 : 0 \leq \theta < 2\pi\right\},$$

em que $\hat{\boldsymbol{m}}^0$ é o maior autovetor correspondente ao maior autovalor da matriz produto $\hat{\boldsymbol{S}} = \sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i^{\star}$ (AMARAL; WOOD, 2010).

É importante destacar que os dados de forma possuem uma certa conexão com os dados axiais. Uma pré-forma, $\boldsymbol{z}$, bidimensional com k *landmarks* pode ser aplicada na esfera $\boldsymbol{S}^{2k}$, enquanto que os dados axiais podem ser definidos como vetores unitários não alinhados na esfera unitária $\pm\boldsymbol{X} \in \boldsymbol{S}^k$.

Para o caso $q = 2$, configurando um objeto bidimensional, o uso de vetores complexos simplificam os cálculos para distâncias. Dessa forma, se $\boldsymbol{p}$ e $\boldsymbol{w}$ são dois vetores complexos de dimensão $(k-1)$, a distância Riemanniana é dada por

$$\rho(\boldsymbol{p}, \boldsymbol{w}) = \arccos|\boldsymbol{p}^*\boldsymbol{w}|,$$

em que $|\boldsymbol{z}| = \sqrt{\boldsymbol{z}^*\boldsymbol{z}}$ e $0 \leq \rho \leq \pi/2$. A distância Riemanniana é o comprimento do menor arco definido por dois pontos na casca da esfera, relacionados a dois vetores, $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$, no espaço de pré-formas (DRYDEN; MARDIA, 2016).

2.2.2 Distância Riemanniana

A distância Riemannian é considerada uma distância intrínseca, pois é dada pelo comprimento do arco da geodésica de minimização entre dois pontos, definido como

$$L = \int_a^b ||\boldsymbol{\gamma}'(t)||_g \, dt,$$

em que L é o tamanho da curva parametrizada $\{\boldsymbol{\gamma}(t); t \in [a, b]\}$ e $||d\boldsymbol{u}||_g = \left\{\sum_i \sum_j g_{ij} du_i du_j\right\}$ é a norma de um vetor $d\boldsymbol{u}$ induzido pelo produto interno g. Aqui, $g = \{g_{ij}\}$ é um tensor positivo definido que determina o produto interno em cada espaço tangente (DRYDEN; MARDIA,

2016). Ou seja, se forem consideradas as seguintes coordenadas $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, a métrica no espaço pode ser definida como

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

Uma alternativa à distância Riemanniana é utilizar a distância extrínseca, que calcula a distância dentro de um espaço de aplicação de dimensão superior. Para tanto, é necessário recorrer a uma função de aplicação, denotada por $j(\boldsymbol{X})$ com $\boldsymbol{X} \in \mathcal{M}$, em conjunto com uma projeção única de espaço de aplicação de volta a variedade, denotada por $P(\cdot)$ (veja Bhattacharya e Bhattacharya (2012), para maiores detalhes).

2.3 SISTEMAS DE COORDENADAS

Considere uma esfera de raio unitário centrado em $\boldsymbol{O}$ de forma que os pontos $P_1, \ldots, P_n$ sejam identificados em sua superfície. É possível considerar os seguintes sistemas de coordenadas para representar dados esféricos:

- Coordenadas Polares: Nesse sistema, um ponto P definido na superfície esférica pode ser identificado como um vetor unitário $\overrightarrow{OP}$. É possível também representar este vetor a partir dos três eixos ortogonais $(x,\ y,\ z)$, centrados na origem da esfera. A colatitude, $\theta \in [0°, 180°]$, é o ângulo entre o vetor unitário $\overrightarrow{OZ}$ e $\overrightarrow{OP}$, enquanto que a longitude, $\phi \in [0°, 360°)$, é o ângulo entre $\overrightarrow{OX}$ e $\overrightarrow{OP^*}$ (medido no sentido anti-horário). O vetor $\overrightarrow{OP^*}$ representa a projeção de $\overrightarrow{OP}$ no plano X-Y. O par (θ, ϕ) pode ser representado por meio de seus respectivos cossenos direcionais

  $$x = \text{sen}\theta\cos\phi, \quad y = \text{sen}\theta\text{sen}\phi \quad \text{e} \quad z = \cos\theta,$$

  que são as contribuições de cada componente da base para a formação de um vetor unitário ou eixo na direção de (θ, ϕ) (FISHER; LEWIS; EMBLETON, 1993). A Figura 5 ilustra a posição dos ângulos e dos eixos ortogonais $(x,\ y,\ z)$ na esfera unitária. Em relação a dados axiais, o vetor unitário $\overrightarrow{OP}$ é substituído pelo diâmetro definido por POP';

- Coordenadas Geográficas: Aqui, a longitude (ϕ') continua sendo uma medida no sentido anti-horário entre os vetores unitários $\overrightarrow{OX}$ e $\overrightarrow{OP^*}$. Nesse sistema, a latitude (θ'), é o ângulo medido entre o vetor $\overrightarrow{OP}$ e o plano X-Y, localizado no plano equatorial. Os

ângulos definidos acima do plano equatorial são convencionados a serem positivos, caso contrário, são negativos (FISHER; LEWIS; EMBLETON, 1993). A transformação do par (θ', ϕ) em coordenadas polares é dado por

$$\theta' = 90^\circ - \theta \quad \text{e} \quad \phi' = \phi;$$

- Coordenadas Geológicas: Recursos lineares com significado direcional, ou seja, linhas dirigidas, são usualmente tratados como vetores unitário. Nesse sistema, uma linha direcionada pode ser representada pelo seu ângulo de declinação (Dec), ou azimuth, e pelo seu ângulo de inclinação (Inc). Este último é o ângulo formado entre o vetor e o plano horizontal (FISHER; LEWIS; EMBLETON, 1993). A transformação em coordenadas polares é dada por

$$\theta = Inc + 90^\circ, \quad \phi = 360^\circ - Dec$$

  Existem outros tipos de ângulos usados nos sistemas de coordenadas geológicas. No entanto, para fins de aplicação nos Capítulos 3 e 4, os apresentados são suficientes.

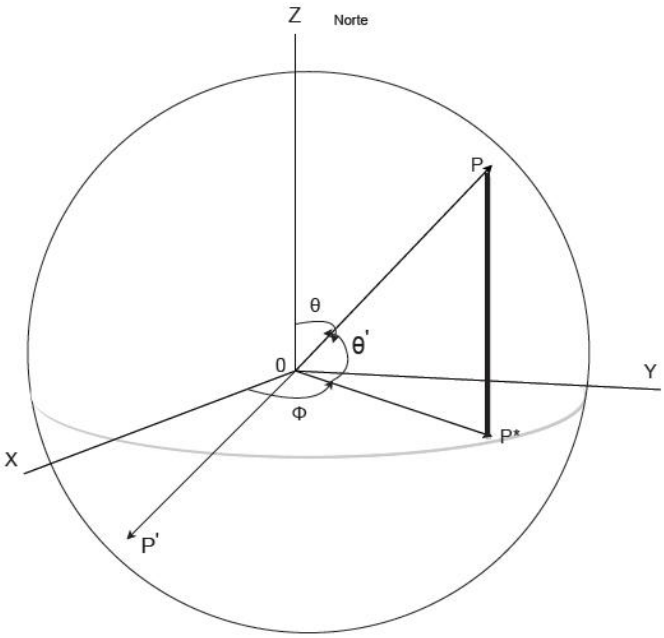
Em geral, é possível usar o vetor unitário $\boldsymbol{X}$ para representar direções na superfície da esfera unitária centrada na origem $\boldsymbol{O}$. Ou seja, os pontos são definidos no espaço $\boldsymbol{S}^{p-1} = \left\{\boldsymbol{X} \mid \boldsymbol{X}^\top \boldsymbol{X} = 1\right\}$ configurando uma hiperesfera com $p-1$ dimensões. Neste cenário, tomando $p = 3$, é possível construir um sistema de coordenadas esféricas (θ, ϕ) com cossenos direcionais

$$\boldsymbol{X} = (\cos\theta,\ \text{sen}\theta\cos\phi,\ \text{sen}\theta\text{sen}\phi)^\top,$$

colatitude (θ) e longitude (ϕ). Por sua vez, o ângulo θ se relaciona com a latitude (θ') da seguinte maneira

$$\theta' = \frac{\pi}{2} - \theta.$$

Figura 5 – Sistema de coordenadas polares: θ é a colatitude e ϕ é a longitude.



2.4 MÉTODOS DE PROJEÇÃO

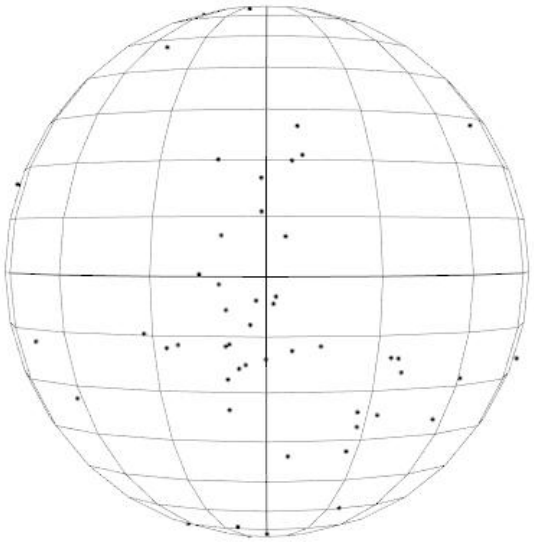
Nesta seção, são apresentadas técnicas para visualizar e projetar observações esféricas. Entende-se que esta é uma forma de visualizar e avaliar a dispersão dos dados. A Figura 6 apresenta as medições das posições dos polos, determinadas a partir de estudos paleomagnéticos de lateritas da Nova Caledônia, disponíveis em Fisher, Lewis e Embleton (1993). Observe que os pontos esféricos estão espalhados pelos hemisférios norte e sul. Contudo, as análises visuais são fortemente influenciadas pela percepção do indivíduos. Dessa forma, torna-se necessário procurar por ferramentas ou mecanismos capazes de checar o grau de concentração destes pontos.

Os pontos esféricos também podem ser visualizados a partir de projeções. Estas transformações buscam mapear a superfície esférica, definida no espaço tridimensional, em um plano tangente. Por exemplo, é possível citar a *Lambert's equal-area projection*, definida como

$$(\cos\theta,\ \text{sen}\theta\cos\phi,\ \text{sen}\theta\text{sen}\phi)^\top \mapsto 2\text{sen}\left(\frac{\theta}{2}\right)(\cos\phi,\ \text{sen}\phi)^\top,$$

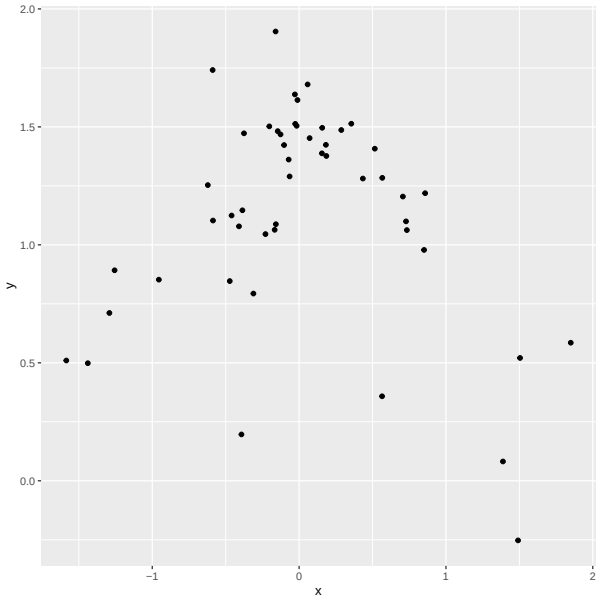
que mapeia a esfera em $\mathbb{R}^3$ para o disco de raio 2. Em geologia esta transformação é conhecida como *Schmidt projection* e possui a propriedade de preservar áreas. No entanto, ela distorce o hemisfério inferior mais do que o superior (MARDIA; JUPP, 1999). A Figura 7 mostra um exemplo de *Lambert's equal-area projection* para os dados de posições dos polos.

Figura 6 – Gráfico de dados esféricos das posições dos polos, medidos em latitude, θ', e longitude, ϕ.



Fonte: O autor (2021).

Figura 7 – *Lambert's equal-area projection* das posições dos polos, medidos em latitude, θ', e longitude, ϕ.



Fonte: O autor (2021).

Em situações em que os dados ocupam mais de um hemisfério, a projeção separada deles é recomendada. Alternativamente, é possível usar símbolos para projetar dados de diferentes hemisférios (FISHER; LEWIS; EMBLETON, 1993). Naturalmente, existe um grande número de projeções devido a necessidade das diferentes áreas da ciência. Assim, a escolha da projeção ideal dependerá das propriedades que o pesquisador deseja preservar após a projeção dos dados no plano. Por exemplo, Fisher, Lewis e Embleton (1993) citam:

- *Lambert's equal-area projection*: é apropriada para estimativas de densidade, pois preserva as densidades dos pontos após a projeção. É necessário destacar que esta projeção provoca uma variação na forma dos grupos de pontos de acordo com suas posições na superfície esférica;
- *Orthographic projection*: é amplamente utilizada para fins estéticos e retrata a visão de corpos celestes pelos astrônomos;
- *Wulff projection*: é uma projeção de ângulo igual amplamente usada por engenheiros, pois as construções geométricas usadas para resolver problemas de engenharia são mais simples e precisas;
- *Central projection*: é usada quando as observações esféricas estão concentradas em um único hemisfério.

2.5 DISTRIBUIÇÕES PARA DADOS NA ESFERA REAL

Nesta seção, são apresentadas algumas das densidades mais populares para descrever observações esféricas.

2.5.1 Distribuição Degenerada

Existem situações em que os ângulos (θ, ϕ) não variam, tomando valores fixos $\theta = \alpha$ e $\phi = \beta$. Neste cenário, o par aleatório (θ, ϕ) é tratado como um vetor degenerado com probabilidade

$$\Pr(\theta = \alpha, \phi = \beta) = 1,$$

denotado por $(\theta, \phi) \sim D(\alpha, \beta)$. Isto equivale a dizer que toda probabilidade esta concentrada em um único ponto (α, β) (FISHER; LEWIS; EMBLETON, 1993).

2.5.2 Distribuição Uniforme

Seja $\boldsymbol{X}$ uma variável aleatória uniformemente distribuída pela superfície da esfera unitária em $\boldsymbol{S}^2$. Neste cenário, todas as direções (θ, ϕ) têm igual probabilidade de ocorrerem. Logo, a variável aleatória (θ, ϕ) tem densidade definida por

$$f(\theta, \phi) = \frac{\text{sen}\theta}{4\pi}, \tag{2.2}$$

com $(0 \leq \theta \leq \pi, 0 \leq \phi < 2\pi)$, denotada por $(\theta, \phi) \sim U_S$. A densidade definida em (2.2) não depende especificamente de ϕ, sendo um análogo à distribuição uniforme na linha, dada por

$$f(x) = \frac{1}{b-a} \quad \text{para} \quad x \in [a, b]\,,$$

em que x é o valor observado da variável aleatória X e sua densidade depende dos parâmetros a e b, em contraste a U_s que é livre de parâmetros. Alguns pontos interessantes sobre U_S são: (i) não possui média direcional; (ii) o comprimento resultante é zero; (iii) a variância esférica é 1 (FISHER; LEWIS; EMBLETON, 1993).

A distribuição U_S na esfera tem um papel importante na análise de dados esféricos. Ela serve como modelo nulo e fornece evidências de características isotrópicas na população (FISHER; LEWIS; EMBLETON, 1993). Ou melhor, a hipótese de uniformidade das distribuições esféricas é uma das mais importantes (MARDIA; JUPP, 1999).

Quando a distribuição de $\boldsymbol{X}$ é uniformemente distribuída na esfera $(p = 3)$ é conhecido que $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{0}$. Logo, se o vetor de médias amostrais é muito diferente de $\boldsymbol{0}$, existem evidências de desvio na suposição de uniformidade da distribuição. Adicionalmente, neste cenário, é costume observar grandes valores para o comprimento médio resultante. O teste de Rayleigh (RAYLEIGH, 1919) para uniformidade de distribuições esféricas usa esta intuição, assumindo $p = 3$. A estatística de teste é dada por

$$p\, n\, \bar{R}^2 \sim \chi^2_p,$$

em que $\bar{R} = R/n$ é o comprimento resultante, $R = \sqrt{S_x^2 + S_y^2 + S_z^2}$ e n é a quantidade de observações de uma amostra. Ou seja, sob a suposição de uniformidade na hipótese nula, a distribuição assintótica para grandes amostras é uma qui-quadrado com p graus de liberdade (MARDIA; JUPP, 1999).

2.5.3 Distribuição von Mises-Fisher

Seja $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ um vetor aleatório definido na superfície de uma hiperesfera de dimensão p, $\boldsymbol{S}^{p-1} = \left\{\boldsymbol{X} \mid \boldsymbol{X} \in \mathbb{R}^p,\ \boldsymbol{X}^\top \boldsymbol{X} = 1\right\}$. Então, o vetor aleatório $\boldsymbol{X}$ segue distribuição von Mises-Fisher se sua densidade é dada por

$$f_p(\boldsymbol{X}; \boldsymbol{\mu}, \kappa) = \left(\frac{\kappa}{2}\right)^{p/2-1} \frac{1}{2\pi^{p/2} I_{(p/2)-1}(\kappa)} \exp(\kappa\, [\boldsymbol{\mu}^\top \boldsymbol{X}]), \tag{2.3}$$

em que $\boldsymbol{\mu}, \boldsymbol{X} \in \boldsymbol{S}^{p-1}$, $\kappa \in (0, \infty)$ e $I_p(\cdot)$ denota a função Bessel modificada de primeiro tipo e ordem p, veja o Apêndice B, (PAINDAVEINE; VERDEBOUT, 2020). O parâmetro de locação $\boldsymbol{\mu}$, denominado por média direcional, é a localização modal na esfera enquanto que o parâmetro κ é responsável por regular a concentração dos pontos ao redor de $\boldsymbol{\mu}$.

De fato, κ é um parâmetro de concentração. Quando o valor de κ aumenta, a distribuição se torna mais concentrada sobre sua locação $\boldsymbol{\mu}$. Quando κ converge para zero, a densidade (2.3) se aproxima da distribuição uniforme em $\boldsymbol{S}^{p-1}$. No outro caso extremo, tomando valores arbitrariamente grandes para κ, são produzidas distribuições degeneradas. Ou seja, suas probabilidades convergem para um ponto de massa em $\boldsymbol{\mu}$. Na prática, o problema surge sobre o que significa κ grande tal que o estado de alta concentração seja obtido. Procuramos dar um entendimento sobre isso nesta tese.

Quando $p = 3$, a distribuição $\text{vMF}_3(\boldsymbol{\mu}, \kappa)$ é conhecida como a distribuição de Fisher, denotada por $F(\boldsymbol{\mu}, \kappa)$, devido as pesquisas de (FISHER, 1953). A densidade de probabilidade de $F(\boldsymbol{\mu}, \kappa)$ é dada por

$$f(\boldsymbol{X}; \boldsymbol{\mu}, \kappa) = \frac{\kappa}{\text{senh}(\kappa)} \exp\left(\kappa\ \boldsymbol{\mu}^\top \boldsymbol{X}\right), \tag{2.4}$$

em que $\text{senh}(\cdot)$ representa a função seno hiperbólico (MARDIA; JUPP, 1999).

Sejam $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ uma amostra aleatória (independente e identicamente distribuída) de $\boldsymbol{X} \sim \text{vMF}_p(\boldsymbol{\mu}, \kappa)$. Os estimadores de máxima verossimilhança para os parâmetros $\boldsymbol{\mu}$ e κ são, respectivamente,

$$\hat{\boldsymbol{\mu}} = \frac{\bar{\boldsymbol{X}}}{||\bar{\boldsymbol{X}}||}$$

e

$$\hat{\kappa} = A_p^{-1}(\bar{R}),$$

em que $\bar{\boldsymbol{X}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i$, $\bar{R} = ||\bar{\boldsymbol{X}}||$ e

$$\begin{aligned} A_p(\kappa) &= \frac{\int_{-1}^{1} t^2 e^{\kappa t}(1-t^2)^{(p-3)/2}\,dt}{\int_{-1}^{1} t^2 e^{\kappa t}(1-t^2)^{(p-3)/2}\,dt} \\ &= \frac{I(p/2)(\kappa)}{I(p/2-1)(\kappa)}. \end{aligned}$$

Note que a estimativa de máxima verossimilhança definida acima depende do inverso de uma razão de funções Bessel modificadas. Esta dependência requer o uso de métodos computacionais para solucionar a complexidade dessas funções (LEY; VERDEBOUT, 2017).

No caso em que $p = 3$,

$$A_3(\kappa) = \coth(\kappa) - \frac{1}{\kappa},$$

com $\coth(\cdot)$ denotando a função cotangente hiperbólica.

2.5.4 Distribuição Watson

A distribuição von Mises-Fisher descrita na seção anterior é amplamente usada para descrever observações que são direções. Porém, em algumas situações onde $\pm\boldsymbol{X}$ são observados não é possível distinguir os vetores unitários $\boldsymbol{X}$ e $-\boldsymbol{X}$. Isto configura um cenário de dados axiais. Nesta situação, é necessário recorrer a densidades de probabilidade antipodalmente simétricas para $\boldsymbol{X} \in \boldsymbol{S}^{p-1}$ (MARDIA; JUPP, 1999).

A densidade é dita ser antipodalmente simétrica quando

$$f(-\boldsymbol{X}) = f(\boldsymbol{X})$$

definida a partir dos pontos $\pm\boldsymbol{X}_1, \ldots, \pm\boldsymbol{X}_n$ opostos na esfera $\boldsymbol{S}^{p-1}$.

A densidade Watson, conhecida como distribuição Dimroth–Scheidegger (DIMROTH, 1962; SCHEIDEGGER, 1965), é um dos modelos mais simples para dados axiais. Ela é distribuída com simetria rotacional e tem densidade girdle ou bipolar. A distribuição Watson, denotada por $\mathrm{W}_p(\boldsymbol{\mu}, \kappa)$, tem densidade

$$f_p(\boldsymbol{X}; \boldsymbol{\mu}, \kappa) = \frac{\Gamma(p/2)}{2\pi^{p/2}} M\left(\frac{1}{2}, \frac{p}{2}, \kappa\right)^{-1} \exp\left\{\kappa\left[\boldsymbol{\mu}^{\top}\boldsymbol{X}\right]^2\right\}, \tag{2.5}$$

com $\boldsymbol{X}, \boldsymbol{\mu} \in \boldsymbol{S}^{p-1}$, $\kappa \in \mathbb{R}$ e $M\left(\frac{1}{2}, \frac{p}{2}, \kappa\right)$ denotando a função de Kummer (KUMMER, 1837), ou seja,

$$M\left(\frac{1}{2}, \frac{p}{2}, \kappa\right) = B\left(\frac{p-1}{2}, \frac{1}{2}\right)^{-1} \int_{-1}^{1} \exp(\kappa t^2)(1-t^2)^{(p-3)/2}\,dt,$$

em que $t = \boldsymbol{X}^\top \boldsymbol{\mu}$ e $B(\cdot, \cdot)$ é a função beta (MARDIA; JUPP, 1999).

A densidade (2.5) apresenta forma bipolar (eixo principal) para $\kappa > 0$ com máximo em $\pm\boldsymbol{\mu}$. Para $\kappa < 0$, a densidade Watson tem forma girdle (eixo polar), concentrando a distribuição ao redor do grande círculo ortogonal a $\boldsymbol{\mu}$. Para $\kappa = 0$ é obtida a distribuição uniforme. O parâmetro de concentração κ é responsável por regular a concentração das observações esféricas. Portanto, à medida que o valor de κ aumenta, a distribuição torna-se mais concentrada em $\pm\boldsymbol{\mu}$ (LEY; VERDEBOUT, 2017).

Sejam $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ uma amostra aleatória de $\boldsymbol{X} \sim \mathrm{W}_p(\boldsymbol{\mu}, \kappa)$. A função de verossimilhança com parâmetros $\boldsymbol{\mu}$ e κ é dada por

$$L(\boldsymbol{\mu}, \kappa; \boldsymbol{X}) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{p}{2}\right)}{2\pi^{p/2} M\left(\frac{1}{2}, \frac{p}{2}, \kappa\right)} \exp\left\{\kappa \left[\boldsymbol{\mu}^\top \boldsymbol{X}\right]^2\right\}. \tag{2.6}$$

Aplicando a função logarítmica na Equação (2.6) é possível obter a seguinte função de log-verossimilhança

$$l(\boldsymbol{\mu}, \kappa; \boldsymbol{X}) = n\left(\kappa \boldsymbol{\mu}^\top \bar{\boldsymbol{T}} \boldsymbol{\mu} - \log M\left(\frac{1}{2}, \frac{p}{2}, \kappa\right) + \gamma\right), \tag{2.7}$$

em que $\bar{\boldsymbol{T}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top$ é a matriz de dispersão da amostra e

$$\gamma = \log\left[\frac{\Gamma\left(\frac{p}{2}\right)}{2\pi^{p/2}}\right]$$

é um termo constante (NASCIMENTO; SILVA; AMARAL, 2018).

Sra e Karp (2013) mostram que a estimativa de máxima verossimilhança do parâmetro $\boldsymbol{\mu}$ pode ser obtida maximizando (2.7). Este processo leva a estimativa do vetor de médias para dois casos

$$\hat{\boldsymbol{\mu}} = \boldsymbol{te}_1, \quad \text{se} \quad \hat{\kappa} > 0$$

e

$$\hat{\boldsymbol{\mu}} = \boldsymbol{te}_p, \quad \text{se} \quad \hat{\kappa} < 0,$$

em que $\boldsymbol{te}_1, \ldots, \boldsymbol{te}_p$ são os autovetores normalizados da matriz de dispersão $\bar{\boldsymbol{T}}$. Os autovalores associados são $\bar{t}_1 \geq \bar{t}_2 \geq \ldots, \geq \bar{t}_p$. Observe que primeiro precisamos computar o valor estimado de κ.

O estimador de máxima verossimilhança para κ pode ser definido usando a solução

$$g(a, c; \kappa) \equiv \frac{M'(a, c, \kappa)}{M(a, c, \kappa)} = r, \quad c > a > 0, \ 0 \leq r \leq 1 \tag{2.8}$$

em que $M'(a, c, \kappa)$ é a derivada da função Kummer em relação a κ e r é o menor ou maior autovalor de $\bar{\boldsymbol{T}}$. O estimador de máxima verossimilhança de $\boldsymbol{\mu}$ é simples de ser obtido. Basicamente, é o menor ou maior autovetor de $\bar{\boldsymbol{T}}$ correspondendo aos casos bipolar ($\hat{\kappa} > 0$) e girdle ($\hat{\kappa} < 0$). Entretanto, o estimador $\hat{\kappa}$ é difícil de ser calculado utilizando a Equação (2.8).

A abordagem assintótica para $\hat{\kappa}$, apresentada a seguir, foi desenvolvida por Sra e Karp (2013). Seja $c > a > 0$, $r \in (0, 1)$ e $\kappa(r)$ uma solução de $g(a, c, \kappa) = r$ definida em (2.8). A aproximação para $\hat{\kappa}$ é fornecida pela expressão

$$\kappa(r) = \frac{-a}{r} + (c - a - 1) + \frac{(c - a - 1)(1 + a)}{a} r + O(r^2), \qquad r \to 0,$$

$$\kappa(r) = \left(r - \frac{a}{c}\right)\left\{\frac{c^2(1 + c)}{a(c - a)} + \frac{c^3(1 + c)^2(2a - c)}{a^2(c - a)^2(c + 2)}\left(r - \frac{a}{c}\right) + O\left(\left(r - \frac{a}{c}\right)^2\right)\right\}, \qquad r \to \frac{a}{c}$$

e

$$\kappa(r) = \frac{c - a}{1 - r} + 1 - a + \frac{(a - 1)(a - c - 1)}{c - a}(1 - r) + O((1 - r)^2), r \to 1.$$

A ideia é resolver ambas as equações $g(a, c; \kappa) = \bar{t}_1$ e $g(a, c; \kappa) = \bar{t}_p$ e então escolher o $\hat{\kappa}$ que fornece a maior probabilidade de log-verossimilhança.

2.6 DISTRIBUIÇÃO PARA O ESPAÇO DE PRÉ-FORMAS (ESFERA COMPLEXA)

Um dos modelos mais relevantes para a análise de pré-formas é a distribuição Bingham complexa. Ela é comumente utilizada para lidar com *landmarks* em duas dimensões (KENT; CONSTABLE; ER, 2004). A esfera complexa unitária em $\mathbb{C}^k$ é denotada por

$$\mathbb{C}\boldsymbol{S}^{k-1} = \left\{\boldsymbol{z} = (z_1, z_2, \ldots, z_k)^\top : \sum |\boldsymbol{z}_i|^2 = 1\right\} \subset \mathbb{C}^k.$$

Um vetor complexo aleatório, $\boldsymbol{z}$, tem distribuição Bingham complexa, denotada por $\mathbb{C}B_{k-1}(\boldsymbol{A})$, se sua densidade é dada por

$$f(\boldsymbol{z}) = c(\boldsymbol{A})^{-1} \exp(\boldsymbol{z}^* \boldsymbol{A} \boldsymbol{z}),$$

com $\boldsymbol{z} \in \mathbb{C}\boldsymbol{S}^{k-1}$, em que $\boldsymbol{z}^*$ é o transposto conjugado complexo de $\boldsymbol{z}$. Aqui, $c(\boldsymbol{A})$ é uma constante de normalização dada por

$$c(\boldsymbol{A}) = 2\pi^k \sum_{j=1}^{k} a_j \exp(\lambda_j),$$

em que

$$a_j^{-1} = \prod_{i \neq j} (\lambda_j - \lambda_i)$$

e $\lambda_1 < \lambda_2 < \ldots < \lambda_j = 0$ denotam os autovalores da matriz Hermitiana $\boldsymbol{A}$, com ordem $p \times p$. A distribuição Bingham complexa é invariante sobre rotações escalares. Ou seja, $\boldsymbol{z}$ e $\exp(i\theta)\boldsymbol{z}$ são iguais em distribuição para todo θ. Esta propriedade permite sua aplicação para análise de forma de *landmarks* em duas dimensões (KENT, 1994).

Seja $n \geq k$ e $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n$ uma amostra aleatória de vetores complexos provenientes de uma distribuição Bingham complexa. Suponha que os autovalores de $\boldsymbol{S} = \sum_{j=1}^{n} \boldsymbol{z}_j \boldsymbol{z}_j^*$ sejam distintos e positivos. Ou seja, tome os autovalores $0 < l_1 < \ldots < l_k$ e seus respectivos autovetores $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k$. Então, a função de log-verossimilhança para os dados é dada por

$$\ell = \sum l_j \lambda_j - n \log c(\Lambda),$$

em que $\Lambda = \operatorname{diag}(\lambda_j)$.

A estimativa de máxima verossimilhança dos autovalores são encontrados por solucionar

$$\frac{\partial \{\log c(\Lambda)\}}{\partial \lambda_j} = \frac{1}{n} l_j, \tag{2.9}$$

com $j = 1, \ldots, k-1$. A Equação (2.9) deve ser resolvida numericamente. Contudo, sob a suposição de alta concentração, tem se que

$$\hat{\lambda_j} \cong -\frac{n}{l_j},$$

com $j = 1, \ldots, k-1$.

2.7 ESTIMAÇÃO DE DENSIDADE KERNEL PARA DADOS NA ESFERA REAL

Os conceitos de estimação de densidade disponíveis na literatura podem ser estendidos ou adaptados para observações esféricas. Ou melhor, para dados cujas medidas são ângulos como: latitude, longitude ou colatitude. Aqui, o espaço amostral dessas observações é a superfície de uma esfera unitária $\boldsymbol{S}^{p-1}$ centrada na origem (DIGGLE; FISHER, 1985).

Mardia e Jupp (1999) discutem métodos de estimação de densidade na esfera para destacar as principais características da distribuição. O mais simples deles usa a função Kernel e é baseado na distribuição von Mises-Fisher, muito comum para dados direcionais. O método consiste em substituir os pontos $\boldsymbol{X}_i, \ldots, \boldsymbol{X}_n$ pela distribuição $\text{vMF}_p(\boldsymbol{X}_i, \kappa)$ definindo a

seguinte estimativa de densidade

$$\hat{f}_F(\boldsymbol{X};\kappa) = n^{-1}a_p(\kappa)\sum_{i=1}^{n} e^{\kappa \boldsymbol{X}^\top \boldsymbol{X}_i},$$

em que

$$a_p^{-1}(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}$$

é a constante de normalização para a distribuição von Mises-Fisher e κ desempenha o papel de parâmetro de suavização. Diggle e Fisher (1985) apresentam procedimentos computacionais para estimação de densidade quando $p = 3$. Em sua análise gráfica, os autores procuram associar a altura da densidade com a intensidade da escala de cor, além de mostrar a forma da densidade.

2.8 SIMETRIA ROTACIONAL NA ESFERA REAL

Na literatura, existe uma ampla quantidade de modelos usados para descrever o comportamento de dados esféricos. Alguns deles, tais como as distribuições von Mises-Fisher e Watson, são discutidos em Mardia e Jupp (1999). Essas duas densidades, em particular, pertencem a uma família de distribuições com propriedade de simetria rotacional sobre sua localização (SAU; RODRIGUEZ, 2017).

As densidades pertencentes a esta classe de distribuições podem ser escritas como

$$f(\boldsymbol{X}) = c_{fa,p} f_a(\boldsymbol{X}^\top \boldsymbol{\mu}), \tag{2.10}$$

em que $f_a : [-1, 1] \to \mathbb{R}^+$, denotada como função angular, é absolutamente contínua, $c_{fa,p}$ refere-se à constante de normalização e $\boldsymbol{X}, \boldsymbol{\mu} \in \boldsymbol{S}^{p-1}$ (LEY; VERDEBOUT, 2017). O uso dessas distribuições é comum em estudos envolvendo rotações tridimensionais, uma vez que as densidades dessas distribuições permanecem invariantes às rotações simétricas provocadas sobre um centro de rotação (BINGHAM; SCRAY, 2017).

Distribuições com propriedade de simetria rotacional sobre sua direção modal $\boldsymbol{\mu}$ tem

$$\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[T]\,\boldsymbol{\mu}$$

e

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{Var}(T)\boldsymbol{\mu}\boldsymbol{\mu}^\top + \frac{1 - \mathbb{E}[T^2]}{p-1}(\boldsymbol{I}_p - \boldsymbol{\mu}\boldsymbol{\mu}^\top),$$

em que $T = \boldsymbol{X}^\top \boldsymbol{\mu}$ assume valores em $-1 \leq t \leq 1$ (MARDIA; JUPP, 1999) e $\boldsymbol{I}_p$ é a matriz identidade de ordem p.

Esta classe de densidade generaliza muitas das distribuições clássicas para dados esféricos por meio de diferentes escolhas da função angular; algumas delas são discutidos em Ley e Verdebout (2017). Por exemplo, tomando $T = \boldsymbol{X}^\top \boldsymbol{\mu} \in [-1, 1]$, as distribuições von Mises-Fisher e Watson são definidas pelas funções angulares $\exp(\kappa t)$ e $\exp(\kappa t^2)$, respectivamente. Ou seja, a importância desta família se deve à capacidade de agrupar diferentes distribuições esféricas. Além disso, a suposição de simetria rotacional é necessária para diferentes procedimentos estatísticos (LEY; VERDEBOUT, 2017).

A propriedade de simetria rotacional pode ser explorada por meio da decomposição estrutural do vetor aleatório $\boldsymbol{X} \in \boldsymbol{S}^{p-1}$ na direção $\boldsymbol{\mu} \in \boldsymbol{S}^{p-1}$, dada por

$$\boldsymbol{X} = (\boldsymbol{X}^\top \boldsymbol{\mu})\boldsymbol{\mu} + (1 - (\boldsymbol{X}^\top \boldsymbol{\mu})^2)^{1/2} S_{\boldsymbol{\mu}}(\boldsymbol{X}), \tag{2.11}$$

chamada *tangent-normal decomposition*, em que

$$S_{\boldsymbol{\mu}}(\boldsymbol{X}) := \frac{\boldsymbol{X} - (\boldsymbol{X}^\top \boldsymbol{\mu})\boldsymbol{\mu}}{||\boldsymbol{X} - (\boldsymbol{X}^\top \boldsymbol{\mu})\boldsymbol{\mu}||}$$

é o vetor de sinais definido no espaço tangente $\boldsymbol{S}^{p-1}(\boldsymbol{\mu}) := \left\{\boldsymbol{v} \in \mathbb{R}^p |\ ||\boldsymbol{v}|| = 1,\ \boldsymbol{v}^\top \boldsymbol{\mu} = 0\right\}$, $||\boldsymbol{A}|| = \sqrt{\langle \boldsymbol{A}, \boldsymbol{A} \rangle} = \sqrt{\boldsymbol{A}^\top \boldsymbol{A}}$ e $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \boldsymbol{A}^\top \boldsymbol{B}$ é o produto interno considerado nesta tese. Ley e Verdebout (2017) mostram a decomposição estrutural (2.11) de $\boldsymbol{X}$ em sua projeção no eixo $\boldsymbol{\mu}$, $(\boldsymbol{\mu}^\top \boldsymbol{X})\,\boldsymbol{\mu}$, e no eixo $\boldsymbol{S}_{\boldsymbol{\mu}}(\boldsymbol{X})$. A Figura 8 ilustra esses termos na esfera unitária. Esses autores enfatizam a projeção em $\boldsymbol{\mu}$ cujo coeficiente associado é $T_p = T_p(\boldsymbol{X}, \boldsymbol{\mu}) := \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle = \boldsymbol{X}^\top \boldsymbol{\mu}$.

A *tangent-normal decomposition* permite definir a seguinte mudança de variável

$$d\sigma_{p-1}(\boldsymbol{x}) = (1 - t^2)^{(p-3)/2} dt d\sigma_{p-2}(\boldsymbol{v}), \tag{2.12}$$

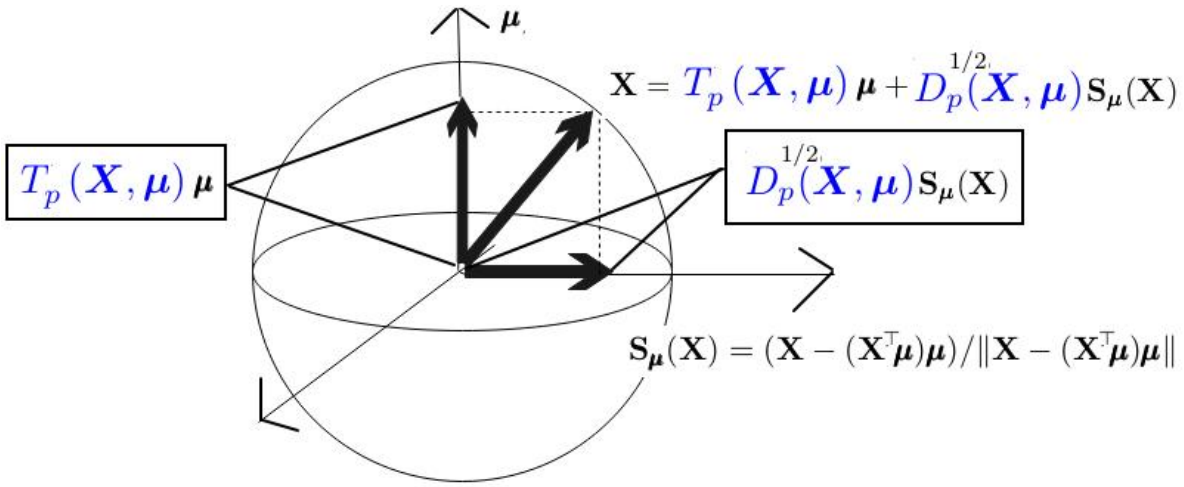
com $\boldsymbol{v} \in \boldsymbol{S}^{p-2}$ e $t \in [-1, 1]$.

Lema 1 (Ley e Verdebout (2017)) *Seja $\boldsymbol{X} \in \boldsymbol{S}^{p-1}$ rotacionalmente simétrico sobre a locação $\boldsymbol{\mu}$. Então, a densidade da projeção de $T = \boldsymbol{X}^\top \boldsymbol{\mu}$ com $t \in [-1, 1]$ é definida por*

$$f_T(t) = \frac{w_p c_{fa,p}}{B\left(\frac{1}{2}, \frac{p-1}{2}\right)} f_a(t)(1 - t^2)^{(p-3)/2}, \tag{2.13}$$

em que $B(\cdot, \cdot)$ é a função beta, $w_p = w_{p-1} B\left(\frac{1}{2}, \frac{p-1}{2}\right)$ e $c_{fa,p}$ é a constante de normalização.

Figura 8 – Geometria para $T_p(\boldsymbol{X}, \boldsymbol{\mu})$ e $D_p(\boldsymbol{X}, \boldsymbol{\mu})$.



Fonte: O autor (2021).

Uma maneira de visualizar a densidade de $T = \boldsymbol{X}^\top\boldsymbol{\mu}$ é demonstrada em Ley e Verdebout (2017) por integrar a densidade (2.10) e considerar a mudança de variável em (2.12). Segue que,

$$\begin{aligned}
\int_{\boldsymbol{S}^{p-1}} c_{fa,p} f_a(\boldsymbol{x}^\top\boldsymbol{\mu})\, d\sigma_{p-1}(\boldsymbol{x}) &= \int_{\boldsymbol{S}^{p-2}} \int_{-1}^{1} c_{fa,p} f_a(t)(1-t^2)^{(p-3)/2}\, dt d\sigma_{p-2}(\boldsymbol{v}) \\
&= \int_{-1}^{1} w_{p-1} c_{fa,p} f_a(t)(1-t^2)^{(p-3)/2}\, dt \\
&= \int_{-1}^{1} \frac{w_p c_{fa,p}}{B\left(\frac{1}{2}, \frac{p-1}{2}\right)} f_a(t)(1-t^2)^{(p-3)/2}\, dt.
\end{aligned}$$

Ou seja, para os casos em que $\boldsymbol{X}$ segue a lei de von Mises-Fisher ou Watson, é possível obter a densidade de $T = \boldsymbol{X}^\top\boldsymbol{\mu}$ a partir Equação (2.13). Este resultado é válido para outras densidades esféricas com propriedades de simetria rotacional, tais como:

- Distribuição de Arnold (ARNOLD, 1941): é usada para modelar dados axiais e tem densidade girdle. Sua função angular é definida por

  $$f_a(t) = e^{-\kappa|t|},$$

  e seu parâmetro de forma κ é capaz de medir a dispersão das observações nos grandes círculos (região equatorial);

- Distribuição de Selby (SELBY, 1964): ela é uma distribuição com densidade girdle. Sua função angular é dada por

$$f_a(t) = e^{\pm\kappa(1-t^2)^{1/2}}$$

e o parâmetro de concentração κ mede a concentração ao redor do grande círculo;

- Distribuição de Purkayastha (PURKAYASTHA, 1991): é um modelo para dados direcionais, tal como a von Mises-Fisher. Seu uso para valores grandes de κ foi considerado por Cabrera e Watson (1990). Sua função angular é dada por

$$f_a(t) = e^{-\kappa[\arccos(t)]},$$

com $\kappa \geq 0$.

Por fim, essa propriedade de simetria rotacional será de fundamental importância para determinar as distribuições baseadas em medidas de distâncias propostas nesta tese.

3 NOVA ABORDAGEM ESTATÍSTICA PARA DETECTAR ALTA CONCENTRAÇÃO EM DADOS DIRECIONAIS

Neste Capítulo, propomos uma nova distribuição, a qual denotamos por TD_1. Ela tem um único parâmetro κ cujo suporte representa a raiz quadrada do coeficiente da projeção de dados direcionais, $\boldsymbol{X}$, sobre o eixo $\boldsymbol{S}_{\boldsymbol{\mu}}(\boldsymbol{X})$ ou uma nova medida de distância entre $\boldsymbol{X}$ e $\boldsymbol{\mu}$. Algumas propriedades dessa nova distribuição são derivadas e discutidas: função geradora de momentos, curtose e assimetria. Vale destacar que o objetivo para esta primeira parte da tese não é propor uma nova distribuição de probabilidade, mas utilizar essa ferramenta para estudar o fenômeno da concentração em dados esféricos. Em seguida, são realizados estudos de simulação envolvendo o parâmetro κ de TD_1. Por fim, é proposta uma estatística para dados direcionais, baseada nos resultados assintóticos de Mardia e Jupp (1999), e a abordagem discutida no texto é aplicada a dados paleomagnéticos.

3.1 PROPOSTA DE UMA NOVA DISTRIBUIÇÃO BASEADA EM DISTÂNCIA A PARTIR DE UM VETOR ALEATÓRIO VON MISES-FISHER

Em análise de dados na hiperesfera unitária $\boldsymbol{S}^{p-1} = \{\boldsymbol{X} \in \mathbb{R}^p : \boldsymbol{X}^\top \boldsymbol{X} = 1\}$, direções ou vetores unitários assumem valores no espaço não Euclidino de dimensão p, produzidos quando a magnitude das observações é irrelevante ou desconhecida (MARDIA; JUPP, 1999). A distribuição vMF é amplamente utilizada em estudos envolvendo dados direcionais. Ela possui dois parâmetros: $\boldsymbol{\mu}$ é a locação e $\kappa \in [0, \infty)$ é a concentração.

Um vetor unitário $\boldsymbol{X}$ tem distribuição von Mises-Fisher, denotada por $\boldsymbol{X} \sim \text{vMF}_p(\boldsymbol{\mu}, \kappa)$, se sua densidade é dada por (para $\boldsymbol{\mu}, \boldsymbol{X} \in \boldsymbol{S}^{p-1}$)

$$f_{\text{vMF}}(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) \;=\; a_p^{-1}(\kappa) \exp(\kappa\, [\boldsymbol{\mu}^\top \boldsymbol{x}]), \tag{3.1}$$

em que $a_p^{-1}(\kappa) = \kappa^{p/2-1} / \left[(2\pi)^{p/2} I_{p/2-1}(\kappa)\right]$ e $I_p(\cdot)$ é a função Bessel modificada de primeiro tipo na ordem p (KO, 1992).

Como discutido em Ko (1992), se $\boldsymbol{X} \sim \text{vMF}_p(\boldsymbol{\mu}, \kappa)$, então o produto interno estocástico $T := T(\boldsymbol{\mu}, \boldsymbol{X}) := \langle \boldsymbol{\mu}, \boldsymbol{X} \rangle = \boldsymbol{\mu}^\top \boldsymbol{X}$ tem densidade (para $t \in [-1, 1]$) dada por

$$f_T(t) \;=\; a_p^{*-1}(\kappa)\; \exp(\kappa t)\, (1 - t^2)^{\frac{p-3}{2}}, \tag{3.2}$$

em que $a_p^{*-1}(\kappa) = w_{p-1}\, a_p^{-1}(\kappa)$, $w_p = 2(\pi)^{p/2} / \Gamma(p/2)$ e $\Gamma(\cdot)$ é a função gama. Para maiores detalhes sobre o produto interno, veja o Apêndice A. Note que T descreve a associação

entre um possível resultado da vMF e o parâmetro de locação, indicando quão concentrado é o resultado. Sabe-se (MARDIA; JUPP, 1999) que a concentração é um dos fenômenos mais importantes na teoria de dados direcionais. No entanto, até onde se conhece, não existem estatísticas de teste para lidar com ela diretamente. Ou seja, um mecanismo capaz de checar alta ou baixa concentração.

3.1.1 Introduzindo a distribuição TD_1

Nesta tese, é entendida que a proposta da distribuição de uma medida de distância em termos de T (chamada de lei $\mathrm{TD}_1(\kappa)$) pode ser uma boa maneira de estudar a concentração. A partir da Equação (3.2), é possível calcular a função de distribuição acumulada (fda) da distância estocástica

$$D := D_p(\boldsymbol{X}, \boldsymbol{\mu}) := 1 - T^2, \tag{3.3}$$

que satisfaz as condições:

1. $D_p(\boldsymbol{X}, \boldsymbol{\mu}) \geq 0$ (Não negatividade).
2. $D_p(\boldsymbol{X}, \boldsymbol{\mu}) = D_p(\boldsymbol{\mu}, \boldsymbol{X})$ (Simetria).
3. $D_p(\boldsymbol{X}, \boldsymbol{\mu}) = 0 \Leftrightarrow \boldsymbol{X} = \boldsymbol{\mu}$ (Definitividade).
4. $D_p(\boldsymbol{X} + \boldsymbol{Y}, \boldsymbol{\mu}) \leq D_p(\boldsymbol{X}, \boldsymbol{\mu}) + D_p(\boldsymbol{Y}, \boldsymbol{\mu})$ (Desigualdade triangular).

Basedo em $D_p(\cdot, \cdot)$, é introduzida uma nova distribuição.

Lema 2 *Seja $\boldsymbol{X} \sim vMF_3(\boldsymbol{\mu}, \kappa)$, a função de distribuição acumulada (fda) de $D_p(\boldsymbol{X}, \boldsymbol{\mu})$ é dada por*

$$F_D(d) = 1 - \frac{\sqrt{2\pi\kappa}}{I_{1/2}(\kappa)} \left[\frac{e^{\kappa\sqrt{1-d}}}{\kappa} - \frac{e^{-\kappa\sqrt{1-d}}}{\kappa} \right], \tag{3.4}$$

para $0 \leq d \leq 1$ e sua densidade tem a seguinte forma:

$$f_D(d) = \frac{a_3^{*-1}(\kappa)}{2\sqrt{1-d}} \left[e^{(\kappa\sqrt{1-d})} + e^{(-\kappa\sqrt{1-d})} \right]. \tag{3.5}$$

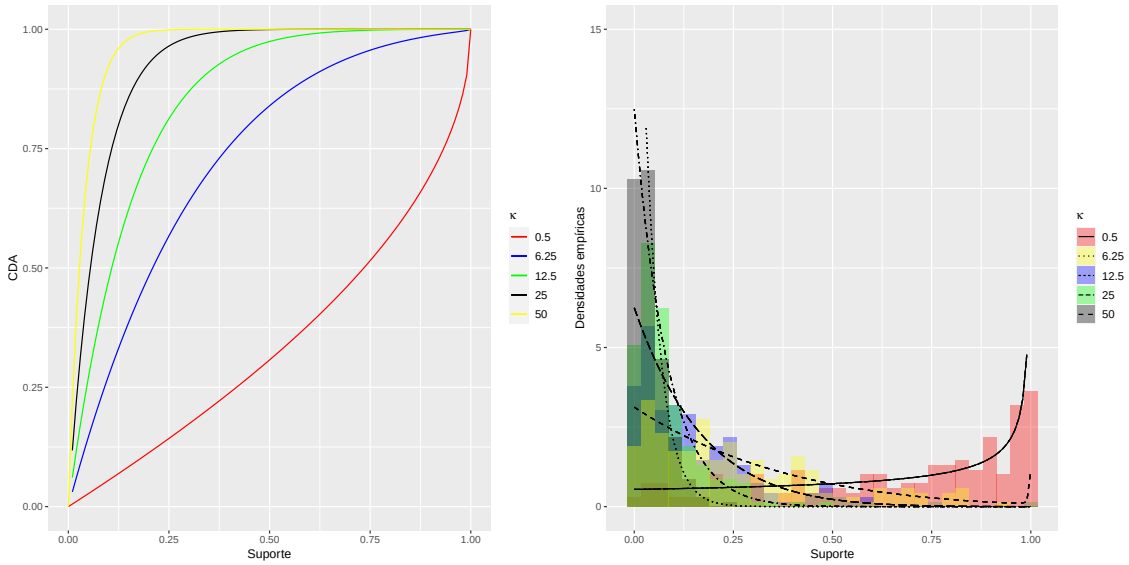
Para maiores detalhes ver Apêndice C.

Na prática, cada resultado gerado a partir deste novo modelo pode ser entendido como uma medida da distância $D(\cdot,\cdot)$ entre uma observação esférica distribuída segundo a lei von Mises-Fisher e seu parâmetro de localização.

Baseado no método de inversão da fda para simulação (RIZZO, 2007), a Figura 9 mostra gráficos de densidade empírica e teórica para alguns valores de κ, além das curvas de distribuição acumulada. Note que quando κ cresce, os dados tendem a se acumular próximo a zero sugerindo que essa distribuição tende a uma variável degenerada, o que indica um cenário de dados concentrados. Ou seja, neste cenário não existe diferença entre os possíveis pontos esféricos gerados (da lei vMF) e um eixo médio pré-especificado. Por outro lado, se κ diminui, a densidade de TD_1 se desloca para a direita e a probabilidade do evento "grandes distâncias" aumenta.

Figura 9 – Curvas da distribuição acumulada e densidade de probabilidade de $D \sim \text{TD}_1(\kappa)$ (em dados esféricos).

(a) Curvas da fda. (b) Curvas da fdp.



Fonte: O autor (2021).

3.1.2 Algumas propriedades matemáticas de TD_1

Entre as propiedades da função geradora de momento (fgm) de uma variável aleatória X, diga-se $M_X(t) = \mathbb{E}(\mathrm{e}^{tx})$, é possível obter as expressões dos momentos a partir de suas

sucessivas derivadas com respeito a t (ROSS, 2010). Assim, a fgm de TD_1 é dada por

$$M_D(t) = \frac{a_p^{*-1}(\kappa)}{2}\left[\int_0^1 \frac{e^{tx}e^{(\kappa\sqrt{1-x})}x^{(p-3)/2}}{\sqrt{1-x}}\,\mathrm{d}x + \int_0^1 \frac{e^{tx}e^{(-\kappa\sqrt{1-x})}x^{(p-3)/2}}{\sqrt{1-x}}\,\mathrm{d}x\right].$$

O Apêndice D apresenta mais informações sobre a fgm de TD_1. A partir de agora, os resultados são derivados assumindo $p = 3$ (para dados esféricos). Essa restrição permitirá estudar o fenômeno de alta concentração na esfera. Depois de algumas manipulações analíticas, o próximo resultado segue.

Corolário 1 *Seja $\boldsymbol{X} \sim vMF_3(\boldsymbol{\mu}, \kappa)$ e $T = \boldsymbol{\mu}^\top \boldsymbol{X}$. Então, a fgm de $D = 1 - T^2$ é*

$$M_D(t) = \frac{a_3^{*-1}(\kappa)\sqrt{\pi}}{2\sqrt{t}}\left[e^{t+\frac{\kappa^2}{4t}}\,\mathrm{erf}\left(\frac{2t+\kappa}{2\sqrt{t}}\right) + e^{t+\frac{\kappa^2}{4t}}\,\mathrm{erf}\left(\frac{2t-\kappa}{2\sqrt{t}}\right)\right],$$

em que

$$erf(d) = \Phi(d) = \frac{2}{\sqrt{\pi}}\int_0^d e^{t^2}\,\mathrm{d}t$$

é a função de erro e $\Phi(\cdot)$ é a fda da distribuição normal padrão. Para maiores detalhes sobre a função de erro ver Gradshteyn e Ryzhik (2000).

Na proposição a seguir, são apresentadas as expressões de momentos que são necessárias para derivar as fórmulas de assimetria e curtose da distribuição TD_1.

Proposição 1 *Sejam $D \sim TD_1(\kappa)$ e $p = 3$. Os quatro primeiros momentos são:*

$$\begin{aligned}
\mathbb{E}(D) &= \frac{a_3^{*-1}(\kappa)e^{-\kappa}\left[(4\kappa-4)\ e^{2\kappa}+4\kappa+4\right]}{2\kappa^3},\\
\mathbb{E}(D^2) &= \frac{a_3^{*-1}(\kappa)e^{-\kappa}\ \left[(16\kappa^2-48\kappa+48)\ e^{2\kappa}-16\kappa^2-48\kappa-48\right]}{2\kappa^5},\\
\mathbb{E}(D^3) &= \frac{a_3^{*-1}(\kappa)e^{-\kappa}\left[(96\kappa^3-576\kappa^2+1440\kappa-1440)\ e^{2\kappa}+96\kappa^3+576\kappa^2\right]}{2\kappa^7}\\
&\quad + \frac{a_3^{*-1}(\kappa)e^{-\kappa}(1440\kappa+1440)}{2\kappa^7}
\end{aligned}$$

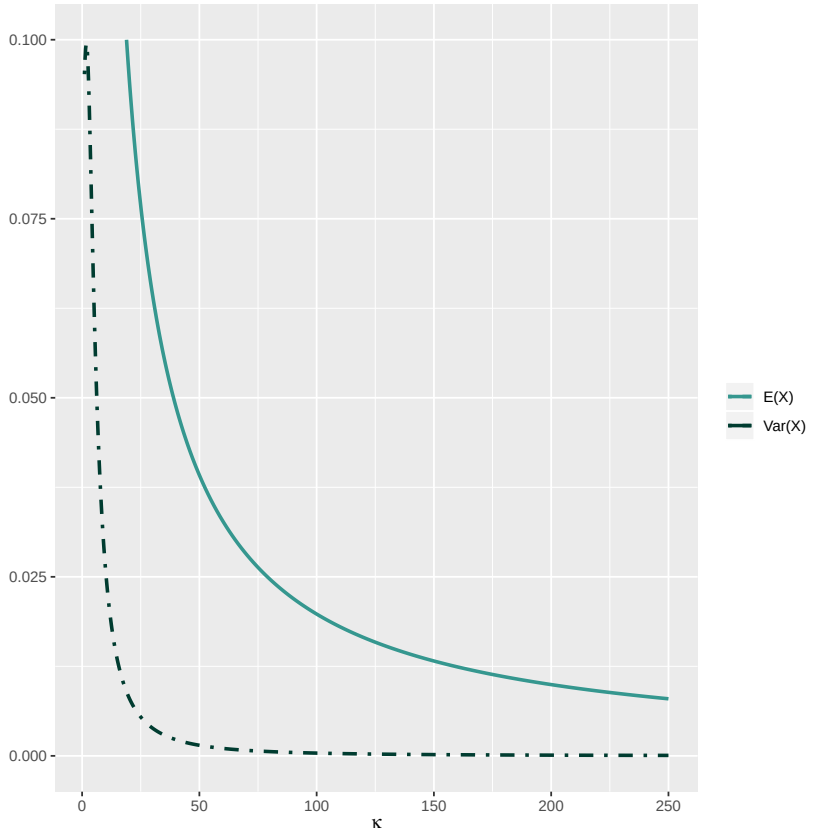
e

$$\begin{aligned}
\mathbb{E}(D^4) &= \frac{a_3^{*-1}(\kappa)e^{-\kappa}[(768\kappa^4-7680\kappa^3+34560\kappa^2-80640\kappa+80640)e^{2\kappa}]}{2\kappa^9}\\
&\quad + \frac{a_3^{*-1}(\kappa)e^{-\kappa}(-768\kappa^4-7680\kappa^3-34560\kappa^2-80640\kappa-80640)}{2\kappa^9}.
\end{aligned}$$

A partir da Proposição 1, é possível obter o k-ésimo momento central, $\mu_k = \mathbb{E}\left[(X - \mathbb{E}(X))^k\right]$. Além disso, as medidas de assimetria, $\mu_3/\mu_3^{3/2}$, e curtose, μ_4/μ_2^2, são deduzidas também. A primeira medida consiste em uma ferramenta para quantificar a falta de simetria; enquanto a segunda pode ser entendida para identificar caudas pesadas e leves das distribuições empíricas.

As curvas de variância e esperança em termos de κ são apresentadas na Figura 10. Note que à medida que κ aumenta a distância média tende a zero com variância zero. Ou seja, a distribuição TD_1 tende a $\Pr(X = 0) = 1$.

Figura 10 – Gráficos da esperança e variância em termos do coeficiente de concentração κ.



Fonte: O autor (2021).

3.1.3 Inferência estatística para o parâmetro da TD_1

Primeiro, considere o método dos momentos (MM) para estimar κ em (3.5). Seja $D_1, \ldots, D_n$ uma amostragem aleatória simples cuja a população tem distribuição $D \sim \text{TD}_1(\kappa)$, $\mu = \mathbb{E}(D)$ e $\bar{D} = n^{-1}\sum_{i=1} D_i$ são os momentos populacionais e amostrais, respectivamente. Tomando a

igualdade entre μ e $\bar{D}$, o estimador MM é obtido a partir da solução da seguinte equação não linear:

$$\frac{\sqrt{\kappa}}{2\sqrt{2\pi}I_{1/2}(\kappa)}\left\{\frac{e^{-\kappa}\ [(4\kappa-4)\ e^{2\kappa}+4\kappa+4]}{\kappa^3}\right\}-\bar{D}\ =\ 0.$$

Agora considere o processo de estimação por máxima verossimilhança para κ. A densidade de TD_1, $f_D(d)$, e a log-verossimilhança, $\ell(\kappa)$, são dadas, respectivamente, por

$$f_D(d)=\frac{a_3^{*-1}(\kappa)}{2\sqrt{1-d}}\left[e^{(\kappa\sqrt{1-d})}+e^{(-\kappa\sqrt{1-d})}\right]=\frac{a_3^{*-1}(\kappa)}{2\sqrt{1-d}}\left[\frac{e^{2\kappa\sqrt{1-d}}+1}{e^{\kappa\sqrt{1-d}}}\right],$$

em que $a_3^{*-1}(\kappa)=\sqrt{\kappa}/[\sqrt{2\pi}I_{1/2}(\kappa)]$, e, para $\boldsymbol{d}=[d_1,\ldots,d_n]^\top$ como uma amostra observada de $D_1,\ldots,D_n$, tem-se que

$$\begin{aligned}\ell(\kappa)=\ell(\kappa;\boldsymbol{d})&=n\log\left(\frac{\sqrt{\kappa}}{\sqrt{2\pi}I_{1/2}(\kappa)}\right)-n\log 2-\frac{1}{2}\sum_{i=1}^{n}\log(1-d_i)\\&\quad+\sum_{i=1}^{n}\log(e^{2\kappa\sqrt{1-d_i}}+1)-\sum_{i=1}^{n}\kappa\sqrt{1-d_i}\\&=n\log(\sqrt{\kappa})-n\log(\sqrt{2\pi})-n\log(I_{(1/2)}(\kappa))-n\log 2\\&\quad-\frac{1}{2}\sum_{i=1}^{n}\log(1-d_i)+\sum_{i=1}^{n}\log(e^{2\kappa\sqrt{1-d_i}}+1)-\sum_{i=1}^{n}\kappa\sqrt{1-d_i},\end{aligned}\tag{3.6}$$

em que $I_p(\cdot)$ é a função de Bessel modificada do primeiro tipo e ordem p. A partir de (3.6), é obtida a função escore, $U(\kappa)=U(\kappa;\boldsymbol{d})=\mathrm{d}\,\ell(\kappa)/\mathrm{d}\,\kappa$, e a matrix Hessiana, $H(\kappa)=H(\kappa;\boldsymbol{d})=\mathrm{d}^2\,\ell(\kappa)/\mathrm{d}\,\kappa^2$. Elas são dadas, respectivamente, por

$$U(\kappa)=\frac{n}{2\kappa}-\frac{\left(I_{3/2}(\kappa)+I_{-1/2}(\kappa)\right)n}{2I_{1/2}(\kappa)}+2\sum_{i=1}^{n}\frac{\sqrt{1-d_i}\,e^{2\kappa\sqrt{1-d_i}}}{e^{2\kappa\sqrt{1-d_i}}+1}-\sum_{i=1}^{n}\sqrt{1-d_i}$$

e

$$\begin{aligned}H(\kappa)&=\frac{-n}{2\kappa^2}+\frac{\left(I_{3/2}(\kappa)+I_{-1/2}(\kappa)\right)^2 n}{4I_{1/2}(\kappa)^2}-n\left(\frac{I_{5/2}(\kappa)+2I_{1/2}(\kappa)+I_{-3/2}(\kappa)}{4I_{1/2}(\kappa)}\right)\\&\quad+4\sum_{i=1}^{n}\left(\frac{(1-d_i)\ e^{2\kappa\sqrt{1-d_i}}}{e^{2\kappa\sqrt{1-d_i}}+1}-\frac{(1-d_i)\ e^{4\kappa\sqrt{1-d_i}}}{\left(e^{2\kappa\sqrt{1-d_i}}+1\right)^2}\right).\end{aligned}$$

Assim, o estimador de máxima verossimilhança (EMV) para κ é dado por

$$\widehat{\kappa}=\arg\max_{\kappa\in\Theta}\left[\ell(\kappa;D_1,\ldots,D_n)\right],$$

em que Θ indica o espaço paramétrico ou, equivalentemente, pela solução da equação não linear $U(\kappa)|_{\kappa=\widehat{\kappa}}=0$. É perceptível que o estimador $\widehat{k}$ não tem forma fechada e, portanto,

seu cálculo requer o uso de processos iterativos, tais como os algoritmos Broyden-Fletcher-Goldfarb-Shanno (BFGS) ou Newton-Raphson. Neste caso, foi usado o BFGS combinado com as estimativas de MM como ponto inicial.

Para a estimação intervalar do parâmetro de TD_1, pode-se adotar a matriz de informações de Fisher (MIF), $K(\kappa) = \mathbb{E}[(\mathrm{d}\,\ell(\kappa)/\mathrm{d}\,\kappa)^2]$ ou, sob as condições de regularidade (BOLFARINE; SANDOVAL, 2010), $K(\kappa) = \mathbb{E}[-H(\kappa; D_1)]$. Sob algumas condições de regularidade (BOLFARINE; SANDOVAL, 2010), $\sqrt{n}\,(\widehat{\kappa} - \kappa) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, K^{-1}(\kappa))$, em que $N(\mu, \sigma^2)$ representa a distribuição normal com média e variância μ and σ^2, respectivamente.

3.2 ESTATÍSTICA DE TESTE EM FUNÇÃO DE UMA VARIÁVEL TD_1 PARA CHECAR ALTA CONCENTRAÇÃO

Na análise esférica de dados, conforme o parâmetro de concentração aumenta, os pontos se sobrepõem ao eixo médio em $\boldsymbol{S}^{p-1}$. Desse modo, uma questão importante sobre o grau de concentração da amostra é: existem evidências de que os dados em estudo são altamente ou ligeiramente concentrados? A resposta para esta pergunta pode ser formulada a partir da construção de um teste de hipóteses, utilizando alguns resultados assintóticos disponíveis em Mardia e Jupp (1999), discutidos a seguir.

Suponha que $\boldsymbol{X} \in \boldsymbol{\Omega}_q$ é distribuído segundo uma von Mises-Fisher com parâmetros de locação $\boldsymbol{\mu}$ e concentração κ. Mardia e Jupp (1999) mostram que

$$2\,\kappa\,(1 - \boldsymbol{X}^\top \boldsymbol{\mu}) \;\sim\; \chi^2_{p-1}, \quad \kappa \to \infty, \tag{3.7}$$

em que χ^2_{p-1} é a distribuição qui-quadrado com $p-1$ graus de liberdade. É possível reescrever este resultado na forma da seguinte estatística: Como $\boldsymbol{\mu}^\top \boldsymbol{X} > 0$ para $\boldsymbol{X} \sim \mathrm{vMF}_3$,

$$S_{\mathrm{DT}} = 2\kappa \left[1 - \sqrt{1 - D_p(\boldsymbol{X}, \boldsymbol{\mu})}\right] \xrightarrow[\kappa\to\infty]{\mathcal{D}} \chi^2_{p-1}, \tag{3.8}$$

em que "$\xrightarrow{\mathcal{D}}$" denota a convergência em distribuição. Com base na Definição (3.5) a distribuição exata de S_{DT} tem função densidade dada por

$$\begin{aligned} f_{\mathsf{S}_{DT}}(s) =& \frac{1}{\kappa}\left(1 - \frac{s}{2\kappa}\right) f_D\left(1 - \left[1 - \frac{s}{2\kappa}\right]^2\right) \\ =& \frac{1}{\kappa}\left(1 - \frac{s}{2\kappa}\right) \frac{a_3^{*-1}(\kappa)}{2\sqrt{\left[1 - \frac{s}{2\kappa}\right]^2}} \frac{\exp\left\{2\kappa\sqrt{\left[1 - \frac{s}{2\kappa}\right]^2}\right\} + 1}{\exp\left\{\kappa\sqrt{\left[1 - \frac{s}{2\kappa}\right]^2}\right\}}, \end{aligned} \tag{3.9}$$

para $0 \leq s \leq 2\kappa$, em que f_D é a densidade de $D(\boldsymbol{X}, \boldsymbol{\mu})$, $a_p^{*-1}(\kappa) = w_{p-1}\, a_p^{-1}(\kappa)$, $w_p = 2(\pi)^{p/2}/\Gamma(p/2)$ e $\Gamma(\cdot)$ é a função gama. Essa nova distribuição denotada por SD_1 apresenta um único parâmetro κ. A partir da discussão anterior, a seguinte proposição pode ser verificada.

Proposição 2 *Sejam $\boldsymbol{X} \sim$ vMF$(\boldsymbol{\mu}, \kappa)$, a distribuição da estatística S_{DT} tem densidade exata dada por* (3.9) *e é assintoticamente distribuída como uma qui-quadrado quando $\kappa \to \infty$.*

Seja $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ uma amostra aleatória de pontos na esfera real. Na prática, é desejado utilizar uma amostra transformada $D_i = 1 - \langle \boldsymbol{X}_i^\top \widehat{\boldsymbol{\mu}}_n \rangle^2$ para $i = 1, \ldots, n$, quando $\widehat{\boldsymbol{\mu}}_n$ é um estimador consistente para $\boldsymbol{\mu}$. A distribuição exata de $1 - \langle \boldsymbol{X}^\top \boldsymbol{\mu} \rangle^2$ difere da distribuição de $1 - \langle \boldsymbol{X}^\top \hat{\boldsymbol{\mu}} \rangle^2$. Contudo, essas distâncias são assintoticamente equivalentes como destacado na proposição a seguir. O resultado assintótico (3.7) é válido ao substituir $\boldsymbol{\mu}$ por $\widehat{\boldsymbol{\mu}}_n$.

Proposição 3 *Seja $\boldsymbol{X} \sim$ vMF$(\boldsymbol{\mu}, \kappa)$ e $\widehat{\boldsymbol{\mu}}_n$ o estimador de máxima verossimilhança (ou outro que satisfaça a propriedade de consistência) para $\boldsymbol{\mu}$, baseado em uma amostra de tamanho n. Então,*

$$D_p(\boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n) \xrightarrow[n\to\infty]{\mathcal{D}} D_p(\boldsymbol{X}, \boldsymbol{\mu}).$$

Prova: Observe que as seguintes igualdades são válidas:

$$\begin{aligned}
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} + \boldsymbol{\mu} \rangle^2 \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - [\langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \rangle + \langle \boldsymbol{X}, \mu \rangle]^2 \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \left[\langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \rangle^2 + \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2 + 2\langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \rangle \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle \right] \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2 - \langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle^2 - 2\langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2 - R_n,
\end{aligned}$$

em que

$$R_n = \langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle^2 + 2\langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle.$$

Por hipótese, $\widehat{\boldsymbol{\mu}}_n$ é consistente (ou seja, $\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \xrightarrow[n\to\infty]{\mathcal{P}} \mathbf{0}$, em que $\mathbf{0}$ é o vetor nulo e "$\xrightarrow{\mathcal{P}}$" significa uma convergência em probabilidade). A partir dos resultados de convergência no produto interno apresentados em Brockwell e Davis (1991, Capítulo 2), $\langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle \xrightarrow[n\to\infty]{\mathcal{P}} 0$, e, portanto, $R_n \xrightarrow[n\to\infty]{\mathcal{P}} 0$. Finalmente, como a convergência em probabilidade implica a em distribuição,

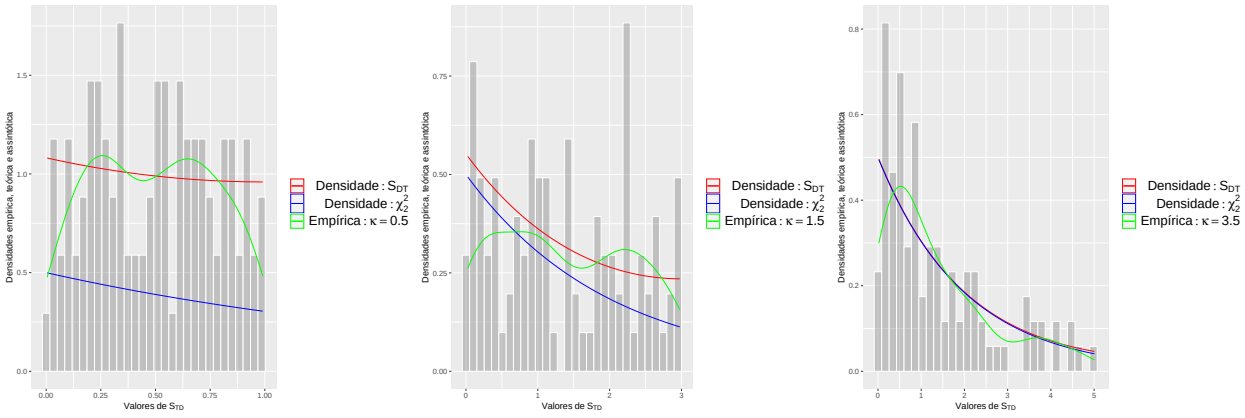
$$1 - \langle \boldsymbol{X}, \hat{\boldsymbol{\mu}} \rangle^2 \xrightarrow{\mathcal{D}} 1 - \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2. \quad \blacksquare$$

A partir do uso de um estimador consistente para κ e usando a Proposição 3, o resultado (3.8) pode ser verificado. É possível utilizar o teste de Kolmogorov-Smirnov para examinar, sob alta concentração, as hipóteses $\mathcal{H}_0\colon S_{\mathsf{DT}} \sim \chi^2_{p-1}$ *versus* $\mathcal{H}_1\colon S_{\mathsf{DT}} \sim \mathsf{SD}_1$.

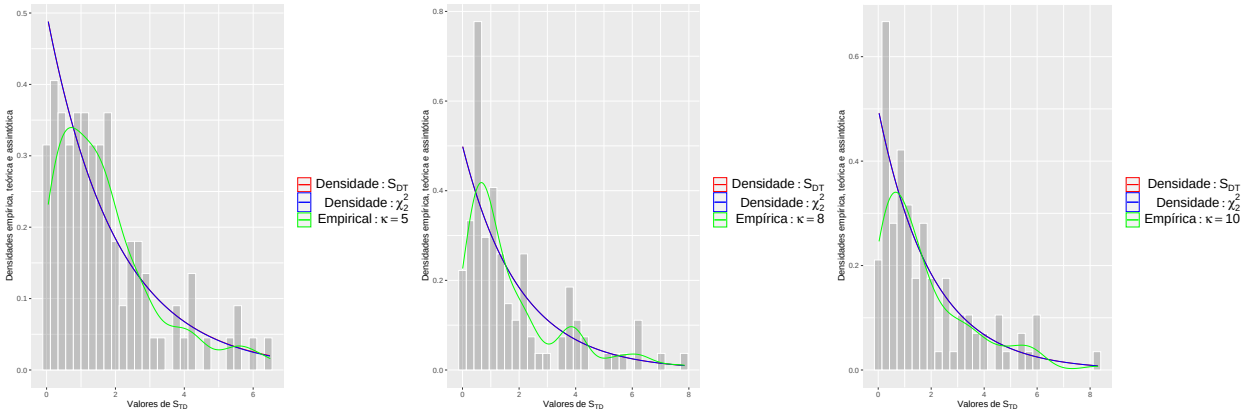
Com base nas discussões anteriores, é possível ilustrar o comportamento da estatística S_{DT}. Neste experimento, cem observações provenientes da distribuição von Mises-Fisher foram geradas para cada cenário, usando a função `rvmf` do pacote `Directional` no *software* `R`. A Figura 11 mostra o comportamento de S_{DT}, avaliado a partir de histogramas, quando κ cresce. Como esperado, as curvas de densidade das distribuições exatas estão mais próximas das curvas de densidade empírica. Contudo, ela se aproxima da qui-quadrado quando κ cresce. Em seguida, o teste de Kolmogorov-Smirnov foi utilizado para verificar a hipótese nula de que a distribuição empírica da estatística é aproximadamente χ^2_2. Os valores-p são fornecidos na Figura 11. Logo, rejeita-se a hipótese nula quando o $p\text{-valor for} \leq 5\%$ (nível de significância adotado). Note que a densidade empírica se aproxima da χ^2_2 para $\kappa \geq 5$.

Figura 11 – Curva de densidade empírica de 100 observações de $2\,\kappa\,\left[1-\sqrt{1-d}\right]$ para diferentes valores de κ.

(a) κ=0.5 e valor-p <0.001. (b) κ=1.5 e valor-p <0.001. (c) κ=3.5 e valor-p=0.0102.



(d) κ=5 e valor-p=0.3046. (e) $\kappa = 8$ e valor-p=0.1957. (f) $\kappa = 10$ e valor-p=0.4067.



Fonte: O autor (2021).

3.3 RESULTADOS DE SIMULAÇÃO

Esta seção tem dois objetivos principais: (i) avaliar o desempenho de dois métodos de estimação desenvolvidos para o parâmetro κ de TD_1; (ii) analisar o comportamento de três testes de hipóteses para diferentes graus de concentração. O estudo por simulação para avaliar o comportamento das estimativas de máxima verossimilhança e método dos momentos considerou o viés médio (VM) e o erro quadrático médio (EQM) como critérios de comparação. Estas medidas são dadas, respectivamente, por

$$\mathrm{B}_{\kappa}(\widehat{\kappa}) = \mathbb{E}(\widehat{\kappa} - g(\kappa))$$

e

$$\mathrm{EQM}(\widehat{\kappa}) = \mathrm{Var}(\widehat{\kappa}) + B^2_{\kappa}(\widehat{\kappa}).$$

Referente aos testes de hipóteses, considere testar a hipótese nula $\mathcal{H}_0$: $\kappa \leq \kappa^{(0)}$ *versus* a hipótese alternativa $\mathcal{H}_1$: $\kappa > \kappa^{(0)}$, em que $\kappa^{(0)}$ é uma concentração pré-especificada. O parâmetro κ de TD_1 aproxima a concentração de pontos sobre a esfera; em particular, grandes valores de κ indicam eventos esféricos concentrados. Assim, é razoável estudar o comportamento dos testes de hipóteses para diferentes configurações κ. Três maneiras clássicas de testar essa hipótese nula é através das seguintes estatísticas:

1. Estatística da Razão de Verossimilhanças (RV) (NEYMAN; PEARSON, 1928)

$$w \;=\; 2\,\{\ell(\hat{\kappa}) - \ell(\tilde{\kappa})\}\,,$$

   em que $\hat{\kappa}$ e $\tilde{\kappa}$ representam os estimadores irrestrito e restrito (ou seja, sob a hipótese nula).

2. Estatística Rao escore (RS) (RAO, 1948)

$$S_R = U(\kappa^{(0)})^\top\, K^{-1}(\kappa^{(0)})\, U(\kappa^{(0)}).$$

3. Estatística de Wald (W) (WALD, 1943)

$$W = (\hat{\kappa} - \kappa^{(0)})^\top K(\hat{\kappa})(\hat{\kappa} - \kappa^{(0)}).$$

Assumindo que as condições de regularidade em Bolfarine e Sandoval (2010) são válidas, as estatísticas de teste acima são equivalentes até primeira ordem. Adicionalmente, elas são distribuídas assintoticamente como uma qui-quadrado com 1 grau de liberdade, dita χ^2_1, sob a hipótese nula. Assim, como regra de decisão, a hipótese nula é rejeitada para $\{w,\ S_R,\ W\} \geq \chi^2_{1,1-\alpha}$, em que $\chi^2_{1,1-\alpha}$ é o quantil $(1-\alpha)\%$ da qui-quadrado. Nas próximas seções, são quantificados o tamanho e o poder dos testes acima para sugerir qual deles é o mais recomendado na prática.

3.3.1 Estimação pontual

Nesta seção, é apresentado um estudo de simulação para avaliar o comportamento das estimativas de MM e EMV. Para tanto, foram consideradas 5000 réplicas de Monte Carlo assim como $\kappa \in \{0.5, 1, 3, 5, 10\}$ e $n \in \{20, 50, 100\}$. Como critérios de comparação, foi utilizado o viés médio (VM) e o erro quadrático médio (EQM).

A Tabela 1 apresenta os resultados obtidos no estudo de simulação. Ao comparar as duas abordagens verificou-se que as estimativas de MM tiveram menor VM para a maioria dos cenários. Além disso, conforme o tamanho da amostra aumenta, as medidas de erro nas duas abordagens tendem a diminuir, como esperado dos resultados assintóticos. Ou seja, a medida que o tamanho da amostra aumenta o VM e o EQM diminuem. Note que valores pequenos de κ tendem a produzir estimativas mais precisas, reduzindo as medidas de erro.

Tabela 1 – Resultados de simulação, referentes ao modelo TD_1, para $\widehat{\kappa}$, $B(\hat{\kappa})$ e $EQM(\hat{\kappa})$ dos EMVs e EMMs.

		EMV			MM		
κ	n	$\hat{\kappa}$	VM	EQM	$\hat{\kappa}$	VM	EQM
0.5	20	1.129	0.629	0.745	0.641	0.141	0.530
	50	0.889	0.389	0.333	0.531	0.031	0.292
	100	0.749	0.249	0.188	0.494	-0.006	0.209
1	20	1.320	0.32	0.511	0.961	-0.039	0.647
	50	1.103	0.103	0.230	0.917	-0.083	0.354
	100	1.028	0.028	0.142	0.938	-0.062	0.212
3	20	3.107	0.107	0.821	3.069	0.069	0.818
	50	3.043	0.043	0.273	3.047	0.047	0.281
	100	3.013	0.013	0.137	3.007	0.007	0.139
5	20	5.264	0.264	1.654	5.181	0.181	1.643
	50	5.093	0.093	0.562	5.082	0.082	0.610
	100	5.046	0.046	0.280	5.044	0.044	0.273
10	20	10.560	0.560	6.721	10.513	0.513	6.712
	50	10.225	0.225	2.176	10.177	0.177	2.247
	100	10.085	0.085	1.059	10.104	0.104	1.056

Fonte: O autor (2021).

3.3.2 Testes de hipóteses

Este é um segundo estudo de simulação para quantificar o desempenho dos testes de hipóteses discutidos anteriormente. Aqui, foram consideradas 5000 réplicas de Monte Carlo, $n \in \{20, 50, 100\}$ e $\kappa \in \{0.5, 1, 3, 5, 10\}$. Em cada cenário foram considerados os testes da razão de verossimilhanças, Wald e escore com objetivo de computar o poder e o tamanho empírico dos testes.

Tabela 2 – Tamanho empírico estimado para os testes da razão de verossimilhanças, escore e Wald, referente ao modelo TD_1. Cenário: $n = \{20,\ 50,\ 100\}$ e $\alpha = 0.05$.

		RV	Escore	Wald
κ	n	$\hat{\alpha}$	$\hat{\alpha}$	$\hat{\alpha}$
0.5	20	0.0466	0.0466	0.0888
	50	0.0432	0.0428	0.0924
	100	0.0350	0.0394	0.0888
1	20	0.0382	0.0380	0.0386
	50	0.0326	0.0302	0.0456
	100	0.0284	0.0260	0.0396
3	20	0.0530	0.0486	0.0356
	50	0.0472	0.0460	0.0444
	100	0.0494	0.0544	0.0478
5	20	0.0506	0.0468	0.0444
	50	0.0572	0.0456	0.0498
	100	0.0480	0.0482	0.0464
10	20	0.0476	0.0468	0.0518
	50	0.0532	0.0502	0.0496
	100	0.0502	0.0488	0.0544

Fonte: O autor (2021).

A Tabela 2 apresenta o tamanho empírico do teste com respeito a $RV,\ RS$, e W. No geral, os testes de hipóteses tendem a ser mais precisos para grandes valores de κ, ou seja, sob alta concentração. Este fato também é observado por Nascimento, Silva e Amaral (2018). Os autores verificaram que os testes de hipóteses em dados direcionais tendem a trabalhar bem apenas sob alta concentração. De modo geral, para $\kappa = 1$, a estimativa do nível nominal foi bem menor do que o valor especificado. Ou seja, os testes são conservadores. O teste de Wald é o mais liberal para pequenas concentrações, especificamente, para $\kappa = 0.5$. Por outro lado, o teste escore é o mais conservador, apresentando um $\hat{\alpha}$ menor do que 5%. Por fim, na maioria dos cenários, o tamanho empírico do teste da razão de verossimilhanças oscilou próximo ao valor adotado de 5%.

As Figuras 12, 13 e 14 mostram o poder empírico dos testes da razão de verossimilhanças, Wald e escore para $\kappa^{(0)} \in \{13,\ 50\}$. A construção desses gráficos considera

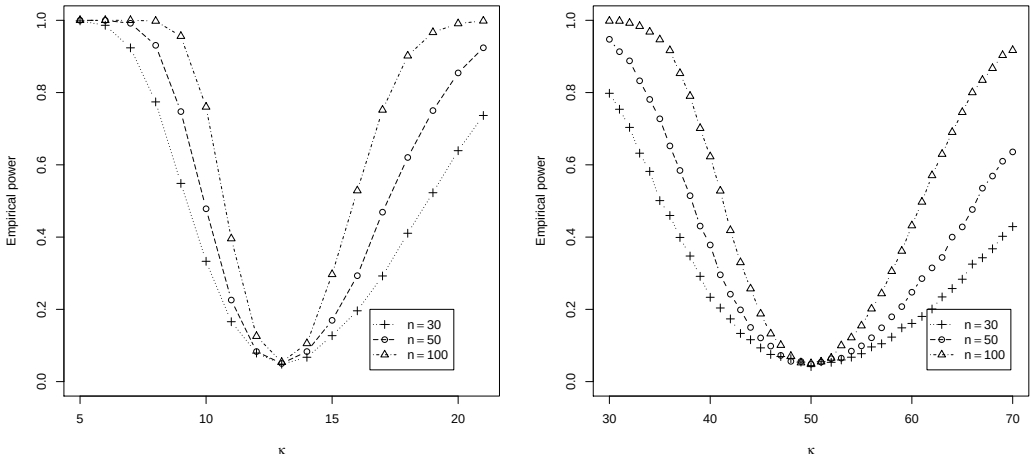
$$\pi(\kappa) \ = \ \Pr(\text{Rejeitar } \mathcal{H}_0 \,|\, \kappa \in \Theta)$$

em duas situações: (i) $\mathcal{H}_0$: $\kappa_0 \ = \ 13$ *versus* $\mathcal{H}_1$: $\kappa_1 \ \neq \ \kappa$ em que $\kappa \in \{5,\ 6,\ 7, \ldots,\ 19,\ 20,\ 21\}$

versus e (ii) $\mathcal{H}_0$: $\kappa_0 = 50$ *versus* $\mathcal{H}_1$: $\kappa_1 \neq \kappa$ em que $\kappa \in \{30,\ 31,\ 32\ldots,\ 68,\ 69,\ 70\}$. Como esperado, o poder empírico ficou próximo ao nível nominal adotado (de 5%) quando κ está próximo de $\kappa^{(0)} \in \{13,\ 50\}$ e cresce à medida que κ se distância de $\kappa^{(0)}$, aproximando-se de um. Em todos os casos, o poder empírico do teste cresce à medida que o tamanho da amostra aumenta.

Figura 12 – Poder empírico do teste da razão de verossimilhanças: $\boldsymbol{X} \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$.

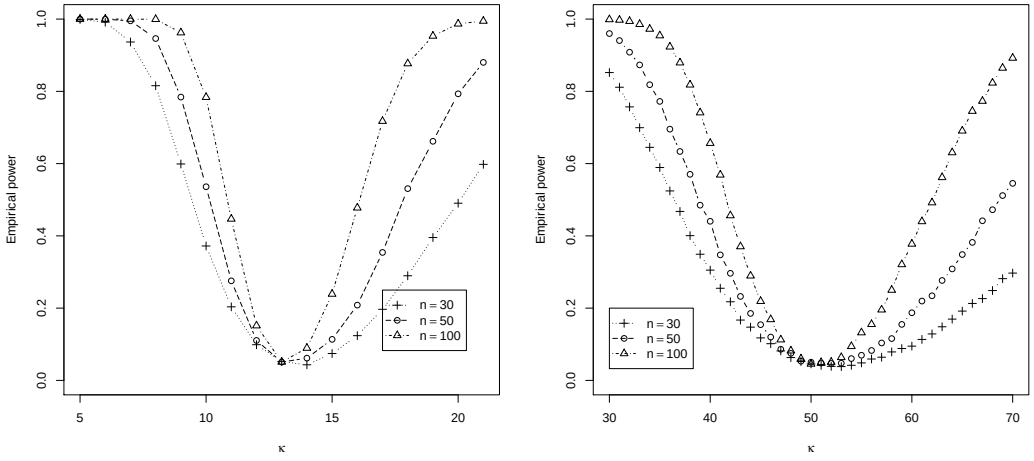
(a) $\kappa_0 = 13$ e $\kappa_1 = \{5, \ldots, 21\}$. (b) $\kappa_0 = 50$ e $\kappa_1 = \{30, \ldots, 70\}$.



Fonte: O autor (2021).

Figura 13 – Poder empírico do teste escore: $\boldsymbol{X} \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$.

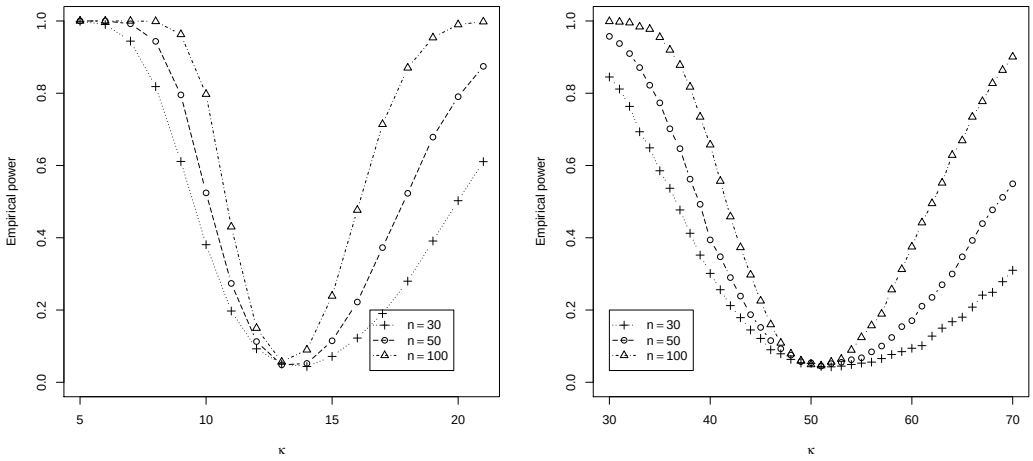
(a) $\kappa_0 = 13$ e $\kappa_1 = \{5, \ldots, 21\}$. (b) $\kappa_0 = 50$ e $\kappa_1 = \{30, \ldots, 70\}$.



Fonte: O autor (2021).

Figura 14 – Poder empírico do teste de Wald: $\boldsymbol{X} \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$.

(a) $\kappa_0 = 13$ e $\kappa_1 = \{5, \ldots, 21\}$. (b) $\kappa_0 = 50$ e $\kappa_1 = \{30, \ldots, 70\}$.



Fonte: O autor (2021).

3.4 APLICAÇÃO A DADOS REAIS: MEDIDAS DE REMANÊNCIA MAGNÉTICA

Nesta seção, é utilizada uma base de dados com 26 observações correspondendo ao sistema de coordenadas de declinação (Dec) e inclinação (Inc), avaliadas em graus e coletadas em *red-beds* (Argentina). Os dados estão disponíveis em Fisher, Lewis e Embleton (1993) e as análises estatísticas foram realizadas por meio do software estatístico R (R Core Team, 2013). As orientações dos vetores unitários podem ser representadas pelas coordenadas polares (θ, ϕ), dadas a seguir

$$\theta = \text{Inc} + 90^\circ \quad \text{e} \quad \phi = 360^\circ - \text{Dec},$$

que determinam um sistema de coordenadas esféricas (FISHER; LEWIS; EMBLETON, 1993).

Fisher, Lewis e Embleton (1993) discutiram esta base de dados e concluíram, a partir da análise dos autovalores da matriz de dispersão, que: a distribuição é unimodal e notaram fortes indícios de simetria rotacional sobre uma direção preferida. Dito isto, a distribuição von Mises-Fisher parece ser uma excelente candidata para modelar as medidas de remanência magnética.

A Figura 15a apresenta uma análise gráfica das medidas de remanência magnética definidas a partir dos ângulos de colatitude (θ) e longitude (ϕ). Na Figura 15b, as estimativas das curvas

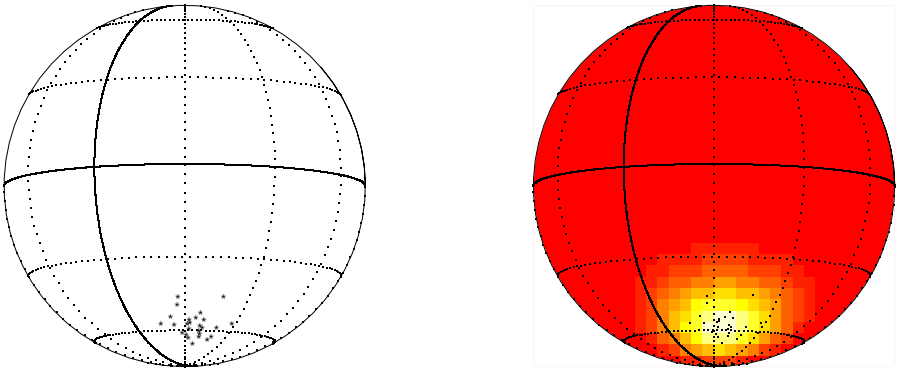
de contorno são traçadas utilizando a função kernel (DIGGLE; FISHER, 1985)

$$\hat{f}(t_1, t_2, t_3) \propto \sum_{i=1}^{n} \exp\Big[h\,(\,t_1\,x_i\,+\,t_2\,y_i\,+\,t_3\,z_i\,)\Big],$$

em que h refere-se ao parâmetro de suavização e $\{(x_i, y_i, z_i); i = 1, \ldots, n\}$ são as observações esféricas avaliadas nas três coordenadas tridimensionais. Esta análise permite relacionar o aumento da densidade com a intensidade da escala de cor (BOWMAN; AZZALINI, 1997). Por fim, é possível observar a proximidade das observações esféricas.

Figura 15 – Gráfico de dados esféricos dos dados de remanência magnética.

(a) Gráfico esférico dos dados. (b) Estimativa de densidade esférica.



Fonte: O autor (2021).

A Figura 16 exibe o histograma e o *boxplot* das medidas de distância $d_1, \ldots, d_n$ de acordo com (2.1), aplicada aos cossenos direcionais dos dados paleomagnéticos. Observe que a distribuição empírica é assimétrica à direita com valores muito próximos de zero. Ou seja, existem evidências da proximidade dos vetores unitários, indicando uma alta concentração dos dados. A partir da análise visual, percebe-se uma proximidade entre as curvas de densidade estimada e empírica. A estimativa do parâmetro κ de TD_1 foi obtida maximizando numericamente a função log-verossimilhança. Para o modelo analisado, o valor estimado de κ foi 113,19. Na sequência, o teste de Kolmogorov-Smirnov foi utilizado para verificar se as distâncias amostrais são provenientes da distribuição de TD_1. O valor-p de 0,9301 sugere, a um nível de significância de 5%, não rejeitar a hipótese nula de que os dados seguem a distribuição proposta.

A identificação de observações discrepantes pode ser feita utilizando os quartos amostrais,

F_L e F_U, para construir o intervalo

$$(F_L - 1.5(F_U - F_L),\ F_U + 1.5(F_U - F_L)),$$

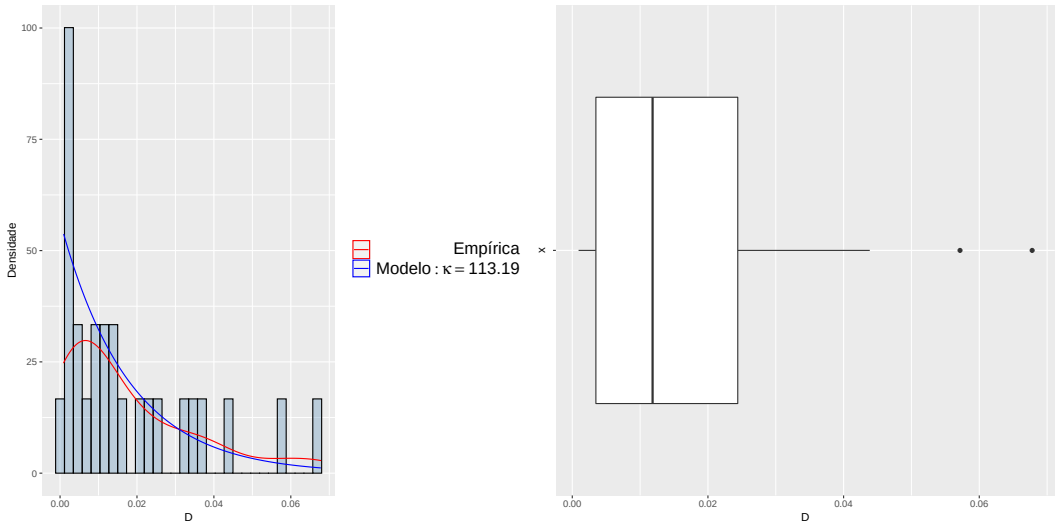
que rotula as observações externas como *outliers*. Similar ao apresentado em Hoaglin e Iglewicz (1987), a primeira etapa é definir a estatística de ordem na amostra $d_1, \ldots, d_n$, dita $d_{(1)}, \ldots, d_{(n)}$, e obter os valores subsequentes de

$$F_L = d_{(f)} \quad \text{e} \quad F_U = d_{(n+1-f)},$$

em que $f = \frac{1}{2}\left[(n+3)/2\right]$ e $[\cdot]$ significa o maior inteiro menor do que o argumento. Para esses dados amostrais, as quantidades $F_L = 0.0033$ e $F_U = 0.0249$ produzem o intervalo $(-0.0293,\ 0.0575)$ indicando, assim, a observação $d_{26} = 0,0678$ como possível *outlier*. O boxplot, Figura 16b, por sua vez, indica duas observações como extremas. A Tabela 6 apresenta algumas estatísticas descritivas como: mínimo (Min.), primeiro quartil ($Q_{1/4}$), mediana, média, terceiro quartil ($Q_{3/4}$), máximo (Max.) e desvio padrão (DP). Ao avaliar a amplitude dos dados, com mínimo de 0.0009 e máximo de 0.0678, é esperado um cenário de alta concentração.

Figura 16 – Histograma e boxplot de D: dados de remanência magnética.

(a) Histograma. (b) Boxplot.



Fonte: O autor (2021).

O principal objetivo é avaliar a concentração dos dados de remanência magnética. Para este propósito, foi utilizada a estatística S_{DT} na Proposição 2 para testar se o fenômeno de alta concentração é obervado nos dados. O p-valor$=$ 0,9304, obtido a partir dos dados amostrais, sugere, ao nível de significância de 5%, que não existem evidências suficientes para

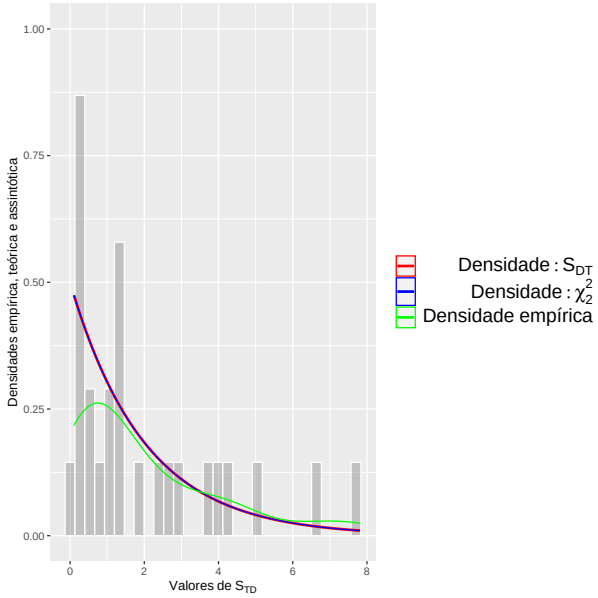
Tabela 3 – Estatísticas descritivas da medida de distância d_i: dados de remanência magnética.

Min.	$Q_{1/4}$	Mediana	Média	$Q_{3/4}$	Max.	DP
0.0009	0.0035	0.0118	0.0175	0.0244	0.0678	0.0181

Fonte: O autor (2021).

rejeitar a hipótese nula. Ou seja, o modelo proposto fornece indícios de um cenário altamente concentrado. A Figura 17 mostra o histograma e a densidade empírica dos valores de S_{DT} e compara com as curvas teórica e assintótica (χ^2_2). Note que a densidade assintótica é próxima das densidades empírica e exata, indicando a presença de alta concentração para os dados na Figura 15.

Figura 17 – Histograma e densidades empírica, teórica e assintótica para valores de S_{DT}.



Fonte: O autor (2021).

4 NOVA ABORDAGEM ESTATÍSTICA PARA DETECTAR ALTA CONCENTRAÇÃO EM DADOS AXIAIS

Neste Capítulo, é explorada a propriedade de simetria rotacional da distribuição Watson que permite definir a distribuição TD_2. O suporte de TD_2 representa a raíz quadrada do coeficiente da projeção de dados axiais, $\boldsymbol{X}$, sobre o eixo $\boldsymbol{S}_{\boldsymbol{\mu}}(\boldsymbol{X})$ ou uma nova medida de distância entre $\boldsymbol{X}$ e $\boldsymbol{\mu}$. Para este modelo, foi derivada a matriz de informação de Fisher e algumas expressões de momentos. Vários estudos de simulação ilustram e quantificam o desempenho de diferentes estimadores e testes assintóticos para a nova distribuição. Por fim, é proposta uma estatística para dados axiais para checagem de alta concentração e sua aplicação é ilustrada em dados reais e simulados.

4.1 PROPOSTA DE UMA NOVA DISTRIBUIÇÃO BASEADA EM DISTÂNCIA A PARTIR DE UM VETOR ALEATÓRIO WATSON REAL

No processo de estimação de dados esféricos, é comum o uso de muitas distribuições, entre as quais é possível citar: a von Mises-Fisher e a Watson. Estas distribuições, por exemplo, pertencem a uma classe de distribuições com a propriedade de simetria rotacional. A propriedade de simetria rotacional, definida para a distribuição Watson, pode ser explorada através da *tangent-normal decomposition* do vetor aleatório $\boldsymbol{X} \in \boldsymbol{S}^{p-1}$ na direção de sua locação $\boldsymbol{\mu} \in \boldsymbol{S}^{p-1}$ (LEY; VERDEBOUT, 2017). Esta propriedade permite obter a densidade da projeção de $T = \boldsymbol{X}^\top \boldsymbol{\mu}$ com $\boldsymbol{X} \sim \text{W}(\boldsymbol{\mu}, \kappa)$ por decompor o vetor aleatório $\boldsymbol{X}$ em suas componentes na esfera, como discutido na Seção 2.8. Com a densidade de T em mãos é possível determinar a distribuição da medida de distância proposta para dados axiais.

4.1.1 Introduzindo a distribuição TD_2

Uma função densidade de probabilidade (fdp) mais geral para T_p, com $\boldsymbol{X}$ rotacionalmente simétrico sobre $\boldsymbol{\mu}$, é dada por

$$f_T(t) = \frac{w_p c_{fa,p}}{B\left(\frac{1}{2}, \frac{p-1}{2}\right)} f_a(t)(1-t^2)^{(p-3)/2},$$

em que $B(\cdot,\cdot)$ é a função beta e $w_p = w_{p-1}B\left(\frac{1}{2}, \frac{p-1}{2}\right)$. Então, considere a seguinte relação

$$\frac{w_p}{B\left(\frac{1}{2}, \frac{p-1}{2}\right)} = \frac{w_p w_{p-1}}{w_p} = w_{p-1}$$

e tome a função angular $f_a(t) = \exp(\kappa t^2)$ que caracteriza a distribuição Watson nesta classe de modelos. A pdf de $T_p(\boldsymbol{X}, \boldsymbol{\mu}) \in [-1, 1]$ para $\boldsymbol{X} \sim \mathsf{W}(\boldsymbol{\mu}, \kappa)$ é definida por

$$f_T(t) = w_{p-1} c_{fa,p} \exp(\kappa t^2)(1-t^2)^{(p-3)/2}, \tag{4.1}$$

em que $w_p = 2\pi^{p/2}/\Gamma(p/2)$ e $c_{fa,p}$ se refere à constante de normalização da Watson. A partir da Equação (4.1), é possível calcular a fda da distância estocástica $D := D_p(\boldsymbol{x}, \boldsymbol{\mu}) := 1 - T_p^2(\boldsymbol{X}, \boldsymbol{\mu})$, que satisfaz as condições de não-negatividade, simetria, definitividade e desigualdade triangular.

Com base em $D_p(\cdot,\cdot)$ é introduzida a próxima distribuição para dados axiais.

Lema 3 *Seja $\boldsymbol{X} \sim W(\boldsymbol{\mu}, \kappa)$. A fda de $D_p(\boldsymbol{X}, \boldsymbol{\mu})$ é dada por*

$$F_{D_p}(d) = \frac{\sqrt{\pi}}{2\sqrt{-\kappa}\; M\left(\frac{1}{2}, \frac{3}{2}, \kappa\right)} \left[\operatorname{erf}(\sqrt{-\kappa}) \; - \; \operatorname{erf}(\sqrt{-\kappa(1-d)}) \right] \tag{4.2}$$

para $0 \leq d \leq 1$ e $p = 3$ (caso esférico), em que erf$(\cdot)$ é a função de erro e $F_{T_p}(\cdot)$ é a fda de T_p. Por derivar a Equação (4.2) com respeito a d, a densidade de $D_p(\boldsymbol{X}, \boldsymbol{\mu})$ tem a seguinte forma

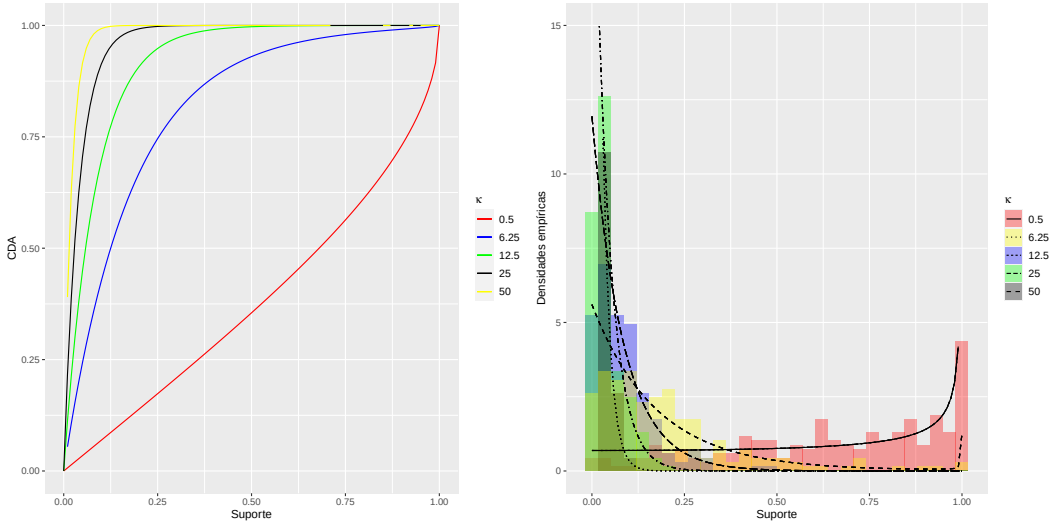
$$f_{D_p}(d) = \frac{\Gamma\left(\frac{p}{2}\right) \exp(\kappa)}{\sqrt{\pi}\,\Gamma\left(\frac{p-1}{2}\right) M\left(\frac{1}{2}, \frac{p}{2}, \kappa\right)} \frac{d^{\frac{p-3}{2}} \exp(-\kappa\, d)}{\sqrt{1-d}}. \tag{4.3}$$

A Equação (4.3) é a densidade proposta que possui distribuição projetada a partir da distância entre um vetor aleatório da Watson real e sua localização $\boldsymbol{\mu}$. Nesta tese, é procurado estudar a concentração de dados axiais na esfera unitária a partir da distribuição dessa medida de distância em termos de T_p. Tomando $p = 3$ na Definição 3, esta distribuição será denotada por $D \sim \mathsf{TD}_2(\kappa)$. Note, a partir das Figuras 18a e 18b, que se $\kappa > 0$ (positivo) diminui, a probabilidade do evento "grandes distâncias" aumenta; à medida que o parâmetro de concentração aumenta, a curva da fdp se aproxima do lado esquerdo, fazendo com que a variável tenda a uma função de probabilidade em $\Pr(D = 0) = 1$. Por outro lado, se $\kappa < 0$ (negativo) diminui, a curva fdp tende a uma função de probabilidade em $\Pr(D = 1) = 1$ indicando que $\boldsymbol{\mu}$ é totalmente ortogonal aos dados localizados na esfera do equador (veja Figura 19).

Figura 18 – Funções densidade de probabilidade e acumulada para diferentes configurações de κ. Assumindo $\boldsymbol{X} \sim W_3(\boldsymbol{\mu}, \kappa)$.

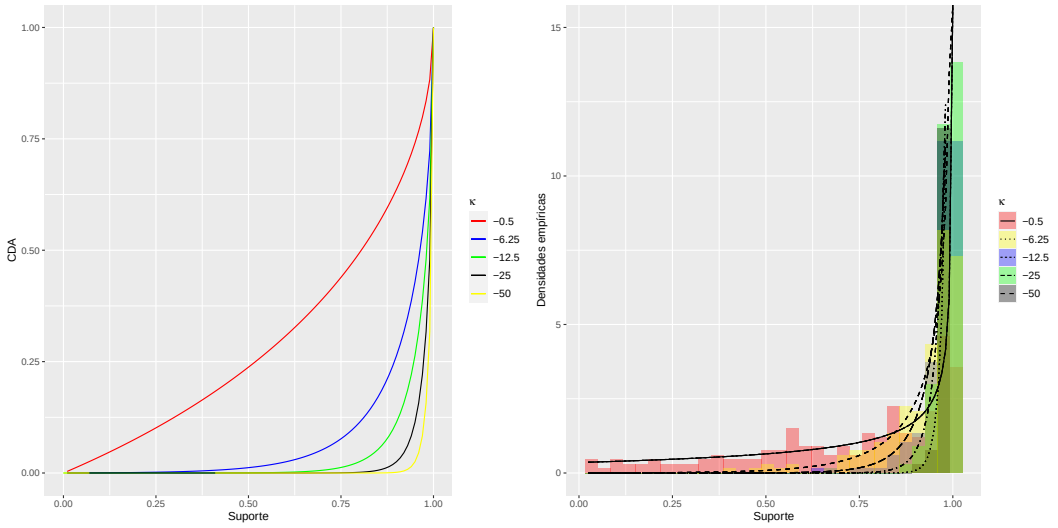
(a) Função de distribuição acumulada (κ positivo).

(b) Função densidade de probabilidade (κ positivo).



(c) Função de distribuição acumulada (κ negativo).

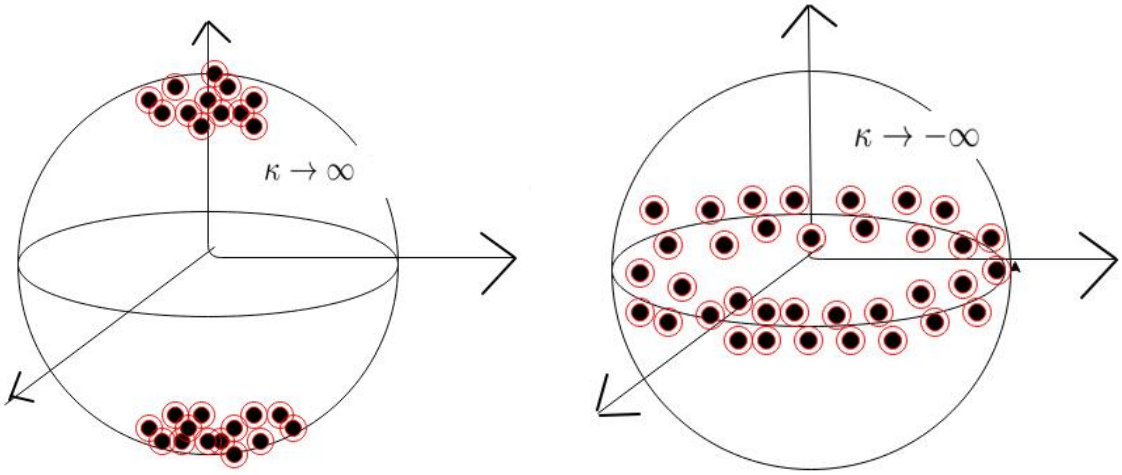
(d) Função densidade de probabilidade (κ negativo).



Fonte: O autor (2021).

Figura 19 – Geometria da Watson real com densidade bipolar ($\kappa > 0$) e densidade girdle ($\kappa < 0$).

(a) Caso bipolar. (b) Caso girdle.



Fonte: O autor (2021).

Na próxima proposição, são apresentadas expressões de momento necessárias para derivar as fórmulas de curtose e assimetria de TD_2. Por meio dessas expressões, pode-se ter mais percepção sobre o uso da distribuição proposta.

Proposição 4 *Sejam* $D \sim TD_2(\kappa)$ *e* $\mu'_k = \int_0^1 t^k f_{D_p}(t)\,\mathrm{d}t$, *seu* k*-ésimo momento, para* $k = 1, 2, 3, 4$ *e* $p = 3$,

$$\mu'_1 = -\frac{c_p(\kappa)\left\{2\,\kappa\,\mathrm{e}^{\kappa} + \sqrt{\pi}\,\sqrt{-\kappa}\,\left[2\,\mathrm{erf}\left(\sqrt{-\kappa}\right)\kappa + \mathrm{erf}\left(\sqrt{-\kappa}\right)\right]\right\}}{2\,\kappa^2},$$

$$\mu'_2 = -\frac{c_p(\kappa)\,\sqrt{\pi}\,(-4\,\kappa^2\,\mathrm{erf}(\sqrt{-\kappa}) - 4\,\kappa\,\mathrm{erf}(\sqrt{-\kappa}) - 3\,\mathrm{erf}(\sqrt{-\kappa}))}{4\kappa^2\sqrt{-\kappa}} - \frac{c_p(\kappa)\sqrt{-\kappa}(4\kappa+6)e^{\kappa}}{4\kappa^2\sqrt{-\kappa}},$$

$$\mu'_3 = -\frac{c_p(\kappa)\sqrt{\pi}\sqrt{-\kappa}(8\,\kappa^3\mathrm{erf}(\sqrt{-\kappa}) + 12\,\kappa^2\,\mathrm{erf}(\sqrt{-\kappa}) + 18\kappa\,\mathrm{erf}(\sqrt{-\kappa}))}{8\kappa^4} - \frac{c_p(\kappa)\sqrt{\pi}\sqrt{-\kappa}(15\,\mathrm{erf}(\sqrt{-\kappa}))}{8\kappa^4} - \frac{(c_p(\kappa)(8\kappa^3+16\kappa^2+30\kappa)e^{\kappa})}{8\kappa^4}$$

e

$$\begin{aligned}\mu_4' = &- \frac{c_p(\kappa)\sqrt{\pi}(-16\kappa^4\mathrm{erf}(\sqrt{-\kappa}) - 32\kappa^3\mathrm{erf}(\sqrt{-\kappa}) - 72\kappa^2\mathrm{erf}(\sqrt{-\kappa}))}{16\kappa^4\sqrt{-\kappa}} \\ &- \frac{c_p(\kappa)\sqrt{\pi}[-120\,\kappa\,\mathrm{erf}(\sqrt{-\kappa}) - 105\,\mathrm{erf}(\sqrt{-\kappa})]}{16\kappa^4\sqrt{-\kappa}} \\ &- \frac{c_p(\kappa)\sqrt{-\kappa}(100\,\kappa\,+\,200)\mathrm{e}^{\kappa}}{16\,\kappa^4\sqrt{-\kappa}} - \frac{c_p(\kappa)\sqrt{-\kappa}(16\,\kappa^3\,+\,40\,\kappa^2)\mathrm{e}^{\kappa}}{16\,\kappa^4\sqrt{-\kappa}}.\end{aligned}$$

O s-ésimo momento central (μ_s) e cumulantes (κ_s) de D são

$$\mu_s = E(D-\mu_1')^s = \sum_{i=0}^{s}(-1)^i \binom{s}{i} (\mu_1')^s\,\mu_{s-i}'$$

e

$$\kappa_s = \mu_s' - \sum_{i=0}^{s-1}\binom{s-1}{i-1}\kappa_r\,\mu_{s-r}', \tag{4.4}$$

respectivamente, em que $\kappa_1 = \mu_1'$. A assimetria e a curtose de D podem ser calculados a partir da Proposição 4. Como discutido por Cordeiro, Silva e Nascimento (2020), a partir da Equação (4.4), as seguintes identidades são obtidas

$$\mu_2 \,=\, \mu_2' - \mu^2,\ \mu_3 \,=\, \mu_3' \,-\, 3\mu_2'\mu \,+\, 2\mu^3$$

e

$$\mu_4 \,=\, \mu_4' \,-\, 4\,\mu_3'\,\mu \,+\, 6\,\mu_2'\,\mu^2 \,-\, 3\,\mu^4.$$

É perceptível que (i) $\mu_1 = 0$, (ii) $\mu_2 = \mathrm{Var}(D) = \mathrm{E}(D-\mu_1')^2 = \sigma^2$ é a variância de D (em que podem ser definidas outras quantidades importantes, tais como o *coeficiente de variação* $\mathrm{CV}(D) = \sigma/\mu_1'$), (iii) $\mathrm{Assimetria}(D) = \mathrm{E}\left(\frac{D-\mu_1'}{\sigma}\right)^3 = \mu_3/\sigma^3$ é a medida de assimetria e (iv) $\mathrm{Curtose}(D) = \mathrm{E}\left(\frac{D-\mu_1'}{\sigma}\right)^4 - 3 = \mu_4/\sigma^4 - 3$ é a medida de curtose.

4.1.2 Inferência estatística para os parâmetros da TD_2

Esta seção aborda os procedimentos de inferência estatística para o parâmetro κ de TD_2 para $p = 3$. Sejam $D_1, \ldots, D_n$ uma amostra aleatória (independente e identicamente distribuída) de $D \sim \mathrm{TD}_2(\kappa)$ com fdp definida por (4.3) e $\boldsymbol{X} \sim W_3(\boldsymbol{\mu}, \kappa)$.

Sejam $d_1, \ldots, d_n$ observações amostrais provenientes de $D_1, \ldots, D_n$. Então, a função de verossimilhança é definida por

$$L(\kappa) := L(\kappa; d) = \prod_{i=1}^{n} \frac{c_3^*(\kappa) e^{(\kappa(1-d_i))}}{\sqrt{1-d_i}},$$

em que

$$c_3^*(\kappa) = w_2 c_{fa,3} = \frac{2\pi\Gamma(3/2)}{2\pi^{3/2} M(1/2, 3/2, \kappa)}.$$

Assim, a função log-verossimilhança, $\ell(\kappa)$, é dada por:

$$\begin{aligned}\ell(\kappa) &= n \log c_3^*(\kappa) + \sum_{i=1}^{n} \kappa(1-d_i) - \frac{1}{2}\sum_{i=1}^{n}(1-d_i) \\ &= -n\log 2 - n \log M(1/2, 3/2, \kappa) + \sum_{i=1}^{n} \kappa(1-d_i) - \frac{1}{2}\sum_{i=1}^{n}\log(1-d_i)\end{aligned}$$

Uma importante quantidade para obter a estimativa de máxima verossimilhança (EMV) para κ, diga-se $\hat{\kappa}$, é a função escore,

$$U(\kappa) = \frac{\partial \ell(\kappa; d)}{\partial \kappa}.$$

A função escore de TD_2 é dada por

$$U(\kappa) = -\frac{nM(3/2,\ 5/2,\ \kappa)}{3M(1/2,\ 3/2,\ \kappa)} + \sum_{i=1}^{n}(1-d_i).$$

Portanto, o estimador de máxima verossimilhança de κ é dado pela solução da equação não linear $U(\kappa)|_{\kappa=\widehat{\kappa}} = 0$ ou, equivalentemente, por

$$\widehat{\kappa} = \arg\max_{\kappa\in\Theta} [\ell(\kappa; D_1, \ldots, D_n)],$$

para o espaço paramétrico Θ.

Outra quantidade importante é a matriz de informação de Fisher (MIF) que pode ser obtida de duas maneiras, $K(\kappa) = \mathbb{E}[(\mathrm{d}\,\ell(\kappa)/\mathrm{d}\,\kappa)^2]$ ou, satisfazendo as condições de regularidade (BOLFARINE; SANDOVAL, 2010), $K(\kappa) = \mathbb{E}[-H(\kappa; D_1)]$. O próximo corolário fornece uma expressão para o TD_2 (MIF).

Corolário 2 *Seja D uma variável aleatória com densidade* (4.3)*. Então sua matriz de informação de Fisher é dada por*

$$K(\kappa) = n\left(\frac{M(5/2,\ 7/2,\ \kappa)}{5M(1/2,\ 3/2,\ \kappa)} - \frac{M(3/2,\ 5/2,\ \kappa)^2}{9M(1/2,\ 3/2,\ \kappa)^2}\right).$$

Observe que a matriz de informação de Fisher depende apenas do parâmetro desconhecido κ, do tamanho amostral e envolve a função especial de Kummer (KUMMER, 1837).

4.2 ESTATÍSTICA DE TESTE EM FUNÇÃO DE UMA VARIÁVEL TD_2 PARA CHECAR ALTA CONCENTRAÇÃO

O parâmetro κ é responsável por regular a concentração dos pontos na superfície esférica. No entanto, até onde se conhece, não existem critérios ou mecanismos capazes de caracterizar os dados axiais como altamente concentrados ou pouco concentrados. Mardia e Jupp (1999) propuseram as seguintes propriedades da distribuição Watson para observações esféricas altamente concentradas. Seja $\boldsymbol{X} \sim \mathrm{W}(\boldsymbol{\mu}, \kappa)$, os autores mostram que: para $\kappa \to \infty$ (caso bipolar),

$$2\kappa \left\{1 - (\boldsymbol{X}^\top \boldsymbol{\mu})^2\right\} \sim \chi^2_{p-1}, \tag{4.5}$$

em que χ^2_{p-1} é a distribuição qui-quadrado com $p-1$ graus de liberdade. Este resultado foi reformulado na forma da estatística:

$$S_{\mathsf{wDT1}} \;=\; 2\,\kappa\, D_p(\boldsymbol{X}, \boldsymbol{\mu}) \xrightarrow[\kappa\to\infty]{\mathcal{D}} \chi^2_{p-1}, \tag{4.6}$$

em que "$\xrightarrow{\mathcal{D}}$" denota a convergência em distribuição. A partir da Definição 3, S_{wDT1} tem distribuição exata com densidade dada por:

$$\begin{aligned} f_{S_{\mathsf{wDT1}}}(s) \;&= \frac{1}{2\,\kappa}\, f_{D_p}\left(\frac{s}{2\,\kappa}\right) \\ &= \frac{\Gamma\left(\frac{p}{2}\right)\,\exp(\kappa)}{(2\,\kappa)^{\frac{p-1}{2}}\sqrt{\pi}\,\Gamma\left(\frac{p-1}{2}\right)\,M\left(\frac{1}{2},\frac{p}{2},\kappa\right)}\,\frac{s^{\frac{p-3}{2}}\,\exp(-\,s/2)}{\sqrt{1-\frac{s}{2\,\kappa}}}, \end{aligned} \tag{4.7}$$

para $s < 2\,\kappa$. Esta distribuição, denotada por SDw_1, apresenta um único parâmetro κ.

Por outro lado, para $\kappa \to -\infty$ (caso girdle), Mardia e Jupp (1999) mostraram que

$$2|\kappa|(\boldsymbol{X}^\top \boldsymbol{\mu})^2 \sim \chi^2_1. \tag{4.8}$$

Então, a seguinte estatística é proposta para descrever alta concentração para dados no equador da esfera:

$$S_{\mathsf{wDT2}} \;=\; 2\,|\kappa|\,[\,1 \;-\; D_p(\boldsymbol{X}, \boldsymbol{\mu})\,] \xrightarrow[|\kappa|\to\infty]{\mathcal{D}} \chi^2_1. \tag{4.9}$$

A partir da Definição 3, S_{wDT2} tem distribuição exata com densidade dada por

$$\begin{aligned} f_{S_{\mathsf{wDT2}}}(s) \;&= \frac{1}{2\,|\kappa|}\, f_{D_p}\left(1 - \frac{s}{2\,|\kappa|}\right) \\ &= \frac{\Gamma\left(\frac{p}{2}\right)}{\sqrt{2\,|\kappa|\,\pi}\;\Gamma\left(\frac{p-1}{2}\right)\,M\left(\frac{1}{2},\frac{p}{2},\kappa\right)}\,\frac{\left(1-\frac{s}{2\,|\kappa|}\right)^{\frac{p-3}{2}}\,\exp(-s/2)}{\sqrt{s}}, \end{aligned} \tag{4.10}$$

para $s < 2|\kappa|$. Esta distribuição apresenta um único parâmetro κ e passa a ser denotada por SDw_2.

A partir da discussão anterior, a próxima proposição é verificada.

Proposição 5 *Seja $\boldsymbol{X} \sim W(\boldsymbol{\mu}, \kappa)$. As estatísticas S_{wDT1} e S_{wDT2} têm distribuições exatas com densidades dadas pelas Equações (4.7) e (4.10) e são distribuídas assintoticamente como uma qui-quadrado quando $|\kappa| \to \infty$.*

Na prática, dada uma amostra de pontos na esfera, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, é desejado trabalhar com uma amostra transformada $D_i = 1 - \langle \boldsymbol{x}_i^\top \widehat{\boldsymbol{\mu}}_n \rangle^2$ para $i = 1, \ldots, n$, quando $\widehat{\boldsymbol{\mu}}_n$ é um estimador consistente para $\boldsymbol{\mu}$, de maneira a capturar a concentração dos dados na amostra. É importante notar que a distribuição exata de $1 - \langle \boldsymbol{X}^\top \boldsymbol{\mu} \rangle^2$ difere daquela devido a $1 - \langle \boldsymbol{X}^\top \widehat{\boldsymbol{\mu}} \rangle^2$, mas essas distâncias são assintoticamente equivalentes como apresentadas na próxima proposição. Assim, os resultados assintóticos (4.5) e (4.8) são válidos ao substituir $\boldsymbol{\mu}$ por $\widehat{\boldsymbol{\mu}}_n$.

Proposição 6 *Sejam $\boldsymbol{X} \sim W(\boldsymbol{\mu}, \kappa)$ e $\widehat{\boldsymbol{\mu}}_n$ o estimador de máxima verossimilhança (ou outro que satisfaça a propriedade de consistência) para $\boldsymbol{\mu}$ baseado em uma amostra de tamanho n. Então,*

$$D_p(\boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n) \xrightarrow[n\to\infty]{\mathcal{D}} D_p(\boldsymbol{X}, \boldsymbol{\mu}).$$

Prova: Observe que as seguintes igualdades são válidas:

$$\begin{aligned}
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} + \boldsymbol{\mu} \rangle^2 \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \left[\langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \rangle + \langle \boldsymbol{X}, \mu \rangle \right]^2 \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \left[\langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \rangle^2 + \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2 + 2 \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \rangle \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle \right] \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2 - \langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle^2 - 2 \langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle \leftrightarrow \\
&1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}}_n \rangle^2 = 1 - \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2 - R_n,
\end{aligned}$$

em que $R_n = \langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle^2 + 2 \langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle$. Por hipótese, $\widehat{\boldsymbol{\mu}}_n$ é consistente (ou seja, $\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu} \xrightarrow[n\to\infty]{\mathcal{P}} \mathbf{0}$, em que $\mathbf{0}$ é o vetor nulo e "$\xrightarrow{\mathcal{P}}$" significa uma convergência em probabilidade). A partir dos resultados de convergência no produto interno apresentados em Brockwell e Davis (1991, Capítulo 2), $\langle \boldsymbol{X}, \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_n \rangle \xrightarrow[n\to\infty]{\mathcal{P}} 0$, e, portanto, $R_n \xrightarrow[n\to\infty]{\mathcal{P}} 0$. Finalmente, como a convergência em probabilidade implica a convergência em distribuição,

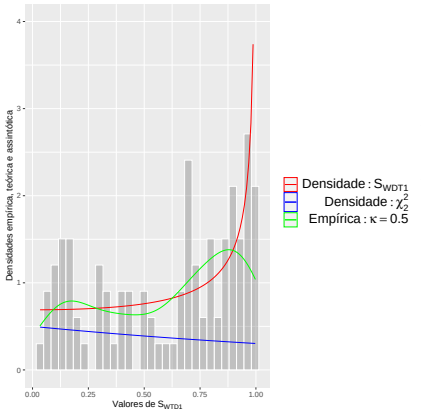
$$1 - \langle \boldsymbol{X}, \widehat{\boldsymbol{\mu}} \rangle^2 \xrightarrow{\mathcal{D}} 1 - \langle \boldsymbol{X}, \boldsymbol{\mu} \rangle^2. \quad \blacksquare$$

A partir da Proposição 6 e do uso de um estimador consistente para κ, os resultados (4.6) e (4.9) podem ser verificados ao substituir $\widehat{\boldsymbol{\mu}}$ e $\widehat{\kappa}$ por $\boldsymbol{\mu}$ e κ, respectivamente. É possível, sob alta concentração, usar o teste de Kolmogorov-Smirnov para verificar as hipóteses: (i) $\mathcal{H}_0\colon S_{\text{wDT1}} \sim \chi^2_{p-1}$ *versus* $\mathcal{H}_1\colon S_{\text{wDT1}} \sim \text{SDw}_1$ ou (ii) $\mathcal{H}_0\colon S_{\text{wDT2}} \sim \chi^2_1$ *versus* $\mathcal{H}_1\colon S_{\text{wDT2}} \sim \text{SDw}_2$.

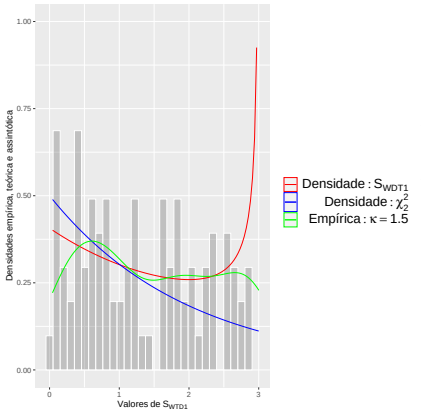
Agora, consideremos um estudo de simulação para avaliar o comportamento das estatísticas S_{wDT1} e S_{wDT2}. Para esse fim, foram geradas cem observações provenientes da distribuição Watson para as situações girdle $(\kappa < 0)$ e bipolar $(\kappa > 0)$. Para isso, foi utilizado um gerador de números aleatórios discutido em (BARROS et al., 2016; LI; WONG, 1993). As Figuras 20 e 21 exibem o comportamento das estatísticas por meio de histogramas para diferentes valores de κ. Como esperado, pode-se notar que as curvas de densidade das distribuições exatas estão mais próximas das densidades empíricas do que das respectivas leis assintóticas da qui-quadrado. O teste de Kolmogorov-Smirnov foi usado para verificar, sob a hipótese nula, se as distribuições empíricas das estatísticas eram aproximadamente χ^2_2 (caso esférico bipolar) ou χ^2_1 (caso esférico girdle). Dadas as estimativas do valor-p, rejeita-se a hipótese nula de que a distribuição dos dados é qui-quadrado para valores menores do que o nível de significância 5%. Observe que para o caso girdle a distribuição empírica da estatística converge mais rápido para uma qui-quadrado do que no caso bipolar. Na Figura 20, a densidade empírica começa a se aproximar χ^2_2 para $\kappa > 8$. Por outro lado, Na Figura 21, a densidade empírica se aproxima da χ^2_1 para $\kappa < -3.5$.

Figura 20 – Histograma das estatísticas axiais para média direcional $(0, 0, 1)$ (caso bipolar).

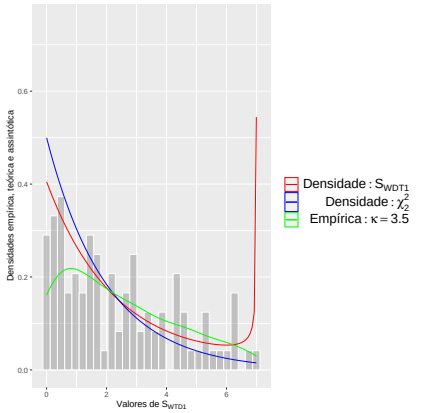
(a) valor-p <0.001 (K-S test).



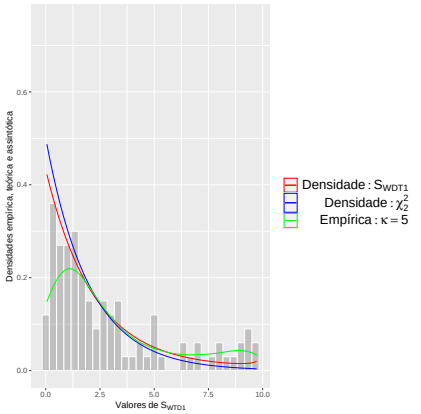
(b) valor-p <0.001 (K-S test).



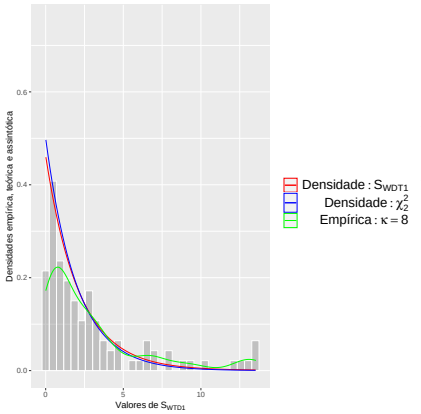
(c) valor-p =0.0253 (K-S test).



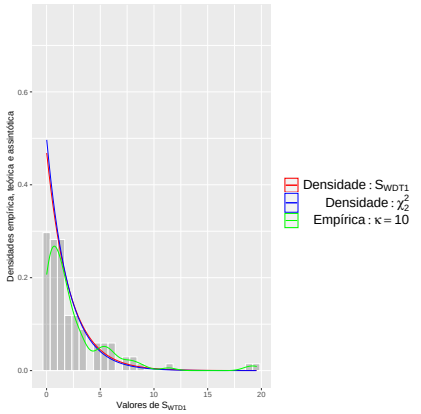
(d) valor-p =0.0357 (K-S test).



(e) valor-p =0.0559 (K-S test).



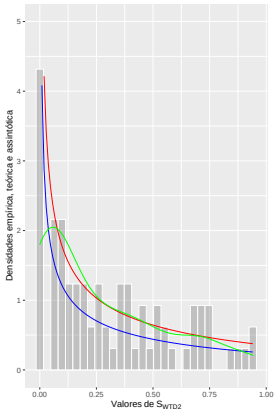
(f) valor-p =0.2611 (K-S test).



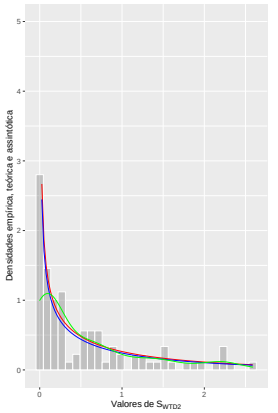
Fonte: O autor (2021).

Figura 21 – Histograma das estatísticas axiais para média direcional $(0, 0, 1)$ (caso girdle).

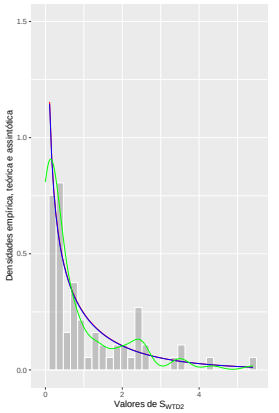
(a) valor-p <0.001 (K-S test).



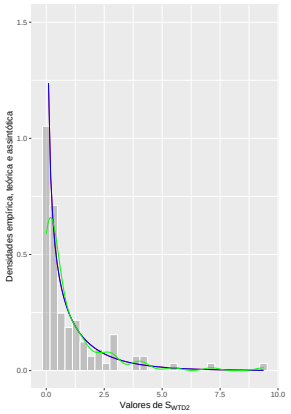
(b) valor-p <0.0170 (K-S test).



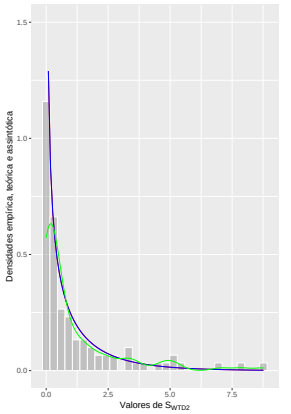
(c) valor-p =0.0767 (K-S test).



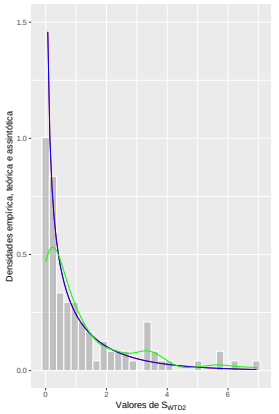
(d) valor-p =0.8172 (K-S test).



(e) valor-p =0.6687 (K-S test).



(f) valor-p = 0.5972 (K-S test).



Fonte: O autor (2021).

4.3 RESULTADOS DE SIMULAÇÃO

Nesta seção, são realizados estudos de simulação para avaliar o comportamento das estimativas pontuais e testes de hipótese, referentes ao parâmetro κ do modelo TD_2. As ferramentas metodológicas são definidas na Seção 3.3.

4.3.1 Estimação Pontual

Nesta seção, é apresentado um estudo de simulação, envolvendo 5000 réplicas de Monte Carlo, a fim de estudar o comportamento das estimativas de máxima verossimilhança para o parâmetro κ de TD_2. Para isso, são avaliados: o viés Monte Carlo (VM) e erro quadrático médio (EQM). Basicamente, dois cenários foram considerados, referentes a Watson bipolar ($\kappa > 0$) e a Watson girdle ($\kappa < 0$), com tamanhos amostrais $n \in \{20,\ 50,\ 100\}$. Para a densidade bipolar foi adotado $\kappa \in \{0.5,\ 1,\ 3,\ 5\}$, enquanto que para densidade girdle foi definido $\kappa \in \{-0.5,\ -1,\ -3,\ -5\}$.

Os resultados avaliados na Tabela 4 mostram que as estimativas de máxima verossimilhança produzem menores medidas de dispersão para $\kappa > 0$. Além disso, para as diferentes concentrações, o erro quadrático médio tende a diminuir com o aumenta do tamanho amostral, mostrando a consistência das estimativas. Por fim, o viés e o erro quadrático médio tendem a aumentar com o aumento da concentração κ.

4.3.2 Testes de hipóteses

Este é um segundo estudo de simulação para avaliar o comportamento dos testes de hipóteses discutidos na Seção 3.3. O experimento considerou 5000 réplicas de Monte Carlo, com tamanhos amostrais $n \in \{20,\ 50,\ 100\}$ e concentrações $\kappa \in \{-5,\ -3,\ -1,\ 1,\ 3,\ 5\}$, para quantificar o desempenho dos testes a partir de seu tamanho empírico. Ou seja, para cada cenário os dados foram gerados sob a suposição da hipótese nula e então estimou-se a

$$\Pr(\text{Rejeitar } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ é verdadeira}).$$

Na prática, espera-se que a probabilidade estimada do erro tipo I seja próxima do nível de significância de 5%. Os resultados são apresentados na Tabela 5. De modo geral, observou-se que a probabilidade estimada do erro tipo I oscila em torno do nível nominal adotado. Ou seja,

Tabela 4 – Resultados de simulação, referente ao modelo TD_2, para $\widehat{\kappa}$, $B(\widehat{\kappa})$ e $EQM(\widehat{\kappa})$.

		EMV			EMV		
κ	n	$\hat{\kappa}$	VM	EQM	$-\hat{\kappa}$	VM	EQM
0.5	20	0.485	-0.015	0.54	-0.569	-0.069	0.718
	50	0.508	0.008	0.21	-0.521	-0.021	0.256
	100	0.505	0.005	0.102	-0.495	0.005	0.127
1	20	0.993	-0.007	0.538	-1.088	-0.088	0.849
	50	0.998	-0.002	0.205	-1.012	-0.012	0.300
	100	1.002	0.002	0.101	-0.993	0.007	0.145
3	20	3.076	0.076	0.637	-3.222	-0.222	1.883
	50	3.035	0.035	0.244	-3.034	-0.034	0.600
	100	3.012	0.012	0.114	-3.01	-0.01	0.277
5	20	5.182	0.182	1.266	-5.471	-0.471	4.368
	50	5.052	0.052	0.417	-5.128	-0.128	1.290
	100	5.034	0.034	0.206	-5.019	-0.019	0.573

Fonte: O autor (2021).

os testes apresentam bons desempenhos para as diferentes concentrações estudadas. O testes de Wald e escore tendem a ser mais conservadores para valores positivos de κ. Ou seja, seu tamanho empírico é menor que o nível nominal adotado (de 5%). Por outro lado, o teste da razão de verossimilhanças tende a ser mais liberal para $\kappa \geq 1$.

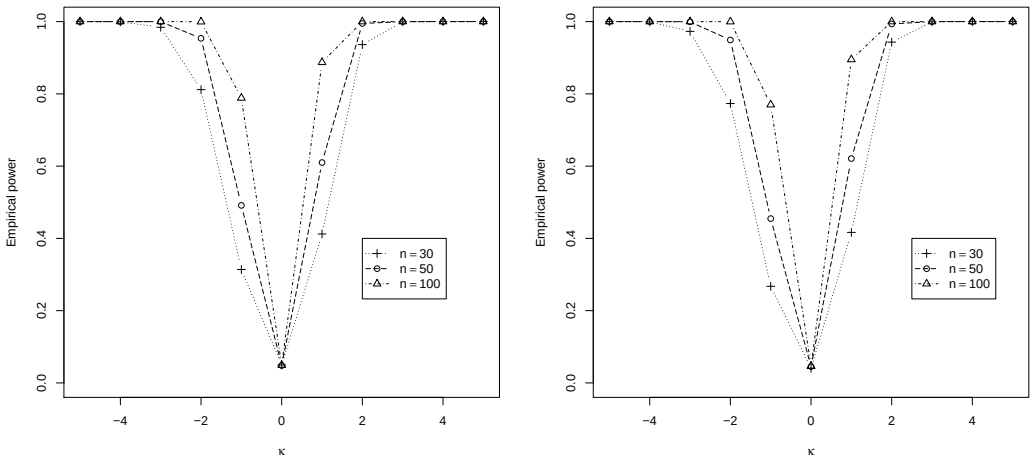
A Figura 22 apresenta o gráfico do poder empírico do teste, dado por

$$\pi(\kappa) = \Pr(\text{Rejeitar } \mathcal{H}_0 \,|\, \kappa \in \Theta),$$

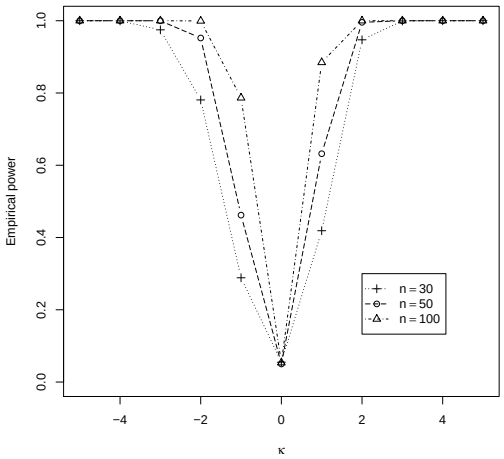
ou seja, foram consideradas as hipóteses $\mathcal{H}_0 : \kappa_0 = 0$ *versus* $\mathcal{H}_1 : \kappa_1 \neq \kappa$ em que $\kappa \in \{-5,\ -4,\ -3\ldots,\ 3,\ 4,\ 5\}$. Para o teste da razão de verossimilhanças, verificou-se que quando κ é próximo de $\kappa_0 = 0$ o poder empírico é próximo de 5%. Por outro lado, quando κ se distância de κ_0 o poder empírico cresce e converge para 1. Adicionalmente, $\pi(\kappa)$ cresce em conjunto com o tamanho da amostra, sendo esse crescimento ligeiramente mais rápido para $\kappa > 0$. Resultados similares são observados para os demais testes, não sendo possível identificar grandes diferenças para os tamanhos amostrais e valores de κ considerados.

Figura 22 – Poder empírico do teste para $n = \{30,\ 50,\ 100\}$ e $\kappa \in [-5, 5]$.

(a) Teste da razão de verossimilhanças. (b) Teste de Wald.



(c) Teste escore.



Fonte: O autor (2021).

Tabela 5 – Tamanho empírico dos testes da razão de verossimilhanças (RV), escore e Wald, referente ao modelo TD_2.

		RV	Escore	Wald
κ	n	$\hat{\alpha}$	$\hat{\alpha}$	$\hat{\alpha}$
	20	0.0528	0.0460	0.0446
-5	50	0.0486	0.0522	0.044
	100	0.0516	0.0506	0.0578
	20	0.0478	0.0448	0.0416
-3	50	0.0472	0.0512	0.0498
	100	0.0456	0.0496	0.0502
	20	0.0452	0.0496	0.0368
-1	50	0.0496	0.0454	0.0522
	100	0.0528	0.0520	0.0476
	20	0.0568	0.0538	0.0428
1	50	0.0514	0.0498	0.0442
	100	0.0554	0.047	0.0498
	20	0.0488	0.0454	0.0474
3	50	0.0538	0.0498	0.0482
	100	0.0502	0.0498	0.0482
	20	0.0528	0.0512	0.043
5	50	0.0512	0.0432	0.0492
	100	0.0468	0.0474	0.0472

Fonte: O autor (2021).

4.4 APLICAÇÃO A DADOS REAIS: MEDIÇÕES DE ORIENTAÇÃO DE CAMPO DENDRÍTICO

Nesta seção, são aplicadas as abordagens de dados axiais discutidas no texto. A base de dados utilizada está disponível em Fisher, Lewis e Embleton (1993) e as análises estatísticas foram feitas no *software* R (R Core Team, 2013). A base de dados se refere a 94 medições das orientações de campo dendrítico em diferentes locais da retina de 6 gatos. As orientações foram provocadas por diferentes formas de luz polarizada (KEILSON et al., 1983). O sistema de coordenadas utilizado foram os ângulos de colatitude (θ) e longitude (ϕ), medidos em graus. Ao analisar os dados, Fisher, Lewis e Embleton (1993) observaram desvios de uniformidade das observações esféricas a partir de evidências fornecidas por uma estatística para dados axiais. Logo, a distribuição Watson pode ser um possível modelo para descrever estes dados.

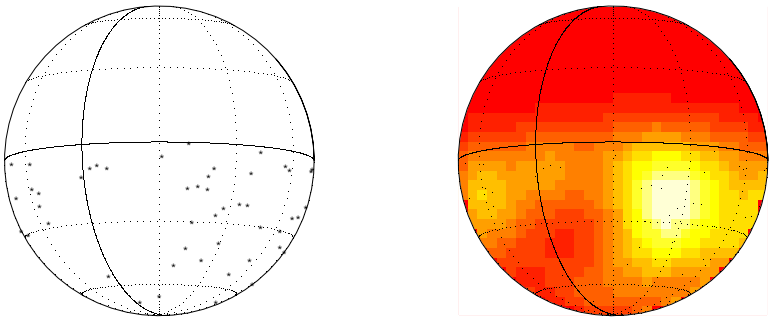
Para fornecer uma análise descritiva prévia, a Figura 23 mostra a proximidade das medições dos campos dendríticos na esfera unitária e uma estimativa de sua curva de densidade empírica por meio da função Kernel (DIGGLE; FISHER, 1985):

$$\hat{f}(t_1, t_2, t_3) \propto \sum_{i=1}^{n} \exp\Big[h\,(\,t_1\,x_i\,+\,t_2\,y_i\,+\,t_3\,z_i\,)\Big],$$

em que h refere-se ao parâmetro de suavização e $\{(x_i, y_i, z_i); i = 1, \ldots, n\}$ são as observações esféricas avaliadas nas três coordenadas tridimensionais (BOWMAN; AZZALINI, 1997). Note uma dispersão das observações próxima à região equatorial, indicando que a distribuição Watson com densidade girdle é um possível modelo para descrever esses dados axiais.

Figura 23 – Gráfico de dados esféricos das medições de orientação do campo dendrítico.

(a) Gráfico esférico de dados. (b) Estimativa da densidade esférica.



Fonte: O autor (2021).

A Tabela 6 apresenta algumas estatísticas descritivas para as medidas de distância $d_1, \ldots, d_n$ provenientes dos dados de campo dendrítico : mínimo (Min.), primeiro quartil ($Q_{1/4}$), mediana, média, terceiro quartil ($Q_{3/4}$), máximo (Máx.) e desvio padrão (DP). Observando o mínimo de 0,0117 e máximo de 0,9999 é verificado que a amplitude dos dados é alta ao considerar o suporte de possíveis valores das distâncias. Além disso, a média e o desvio padrão apresentam grandes valores, indicando que as distâncias são bem heterogêneas. Ou seja, existe um grande número de pontos relativamente distantes entre si, gerando indícios de um cenário pouco concentrado.

A Figura 24a mostra o histograma e as densidades empírica e teórica dos dados. As distâncias empíricas evidenciam uma certa assimetria à esquerda, destacando um grande número de observações afastadas da direção média. Adicionalmente, é verificada uma proximidade entre

Tabela 6 – Estatística descritiva de D: Medidas de orientação de campo dendrítico.

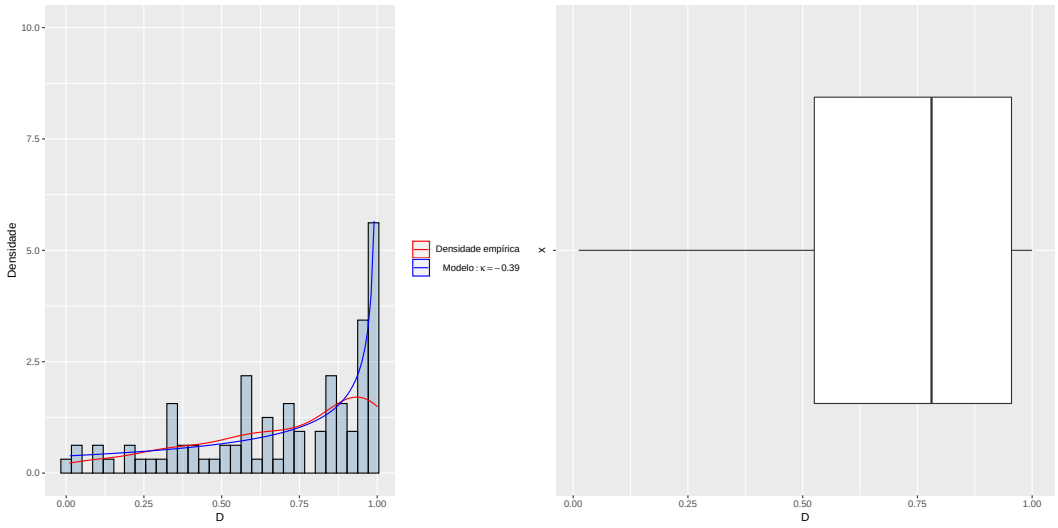
Min.	$Q_{1/4}$	Mediana	Média	$Q_{3/4}$	Max.	DP
0.0117	0.5253	0.7808	0.6996	0.9550	0.9999	0.2867

Fonte: O autor (2021).

as duas curvas de densidade, sugerindo o modelo proposto como uma boa alternativa para descrever os dados. O teste de Kolmogorov-Smirnov forneceu o valor-p de 0,9209, indicando a não rejeição da hipótese nula de que os dados seguem a distribuição de probabilidade proposta. Aqui, o valor estimado do parâmetro κ, via máxima verossimilhança, foi de $-0,39$. No boxplot, Figura 24b, não é verificada a presença de *outliers*.

Figura 24 – Histograma e boxplot das medidas de distância.

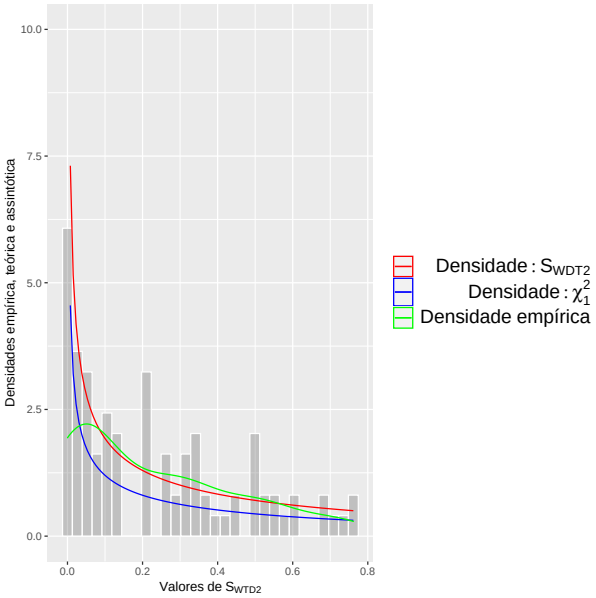
(a) Histograma. (b) Boxplot.



Fonte: O autor (2021).

A partir da Proposição 5 e tomando a estatística S_{wDT2} é possível testar se o fenômeno de alta concentração é encontrado. Para estes dados amostrais, o valor-p associado foi < 0.001, sugerindo em um nível de significância de 5% a rejeição da hipótese nula. Ou seja, existem evidências de um cenário com pouca concentração. A Figura 25 mostra o histograma e a densidade empírica dos valores de S_{wDT2} e as curvas teórica e assintótica (χ^2_1). É perceptível que a densidade assintótica mostrou-se distinta das densidades empírica e exata, indicando a presença de baixa concentração para os dados da Figura 23.

Figura 25 – Histograma e densidades empírica, exata e assintótica para valores de S_{wDT2}.



Fonte: O autor (2021).

5 PROPOSTA DE UMA MEDIANA EXTRÍNSECA DE FRÉCHET PARA DADOS DE PRÉ-FORMAS

Neste Capítulo, é proposta uma alternativa robusta e analiticamente simples para a média extrínseca de Fréchet, nomeada de *mediana extrínseca projetada de Fréchet*. Esta medida pode ser útil para uma variedade de espaços não euclidianos de interesse, a exemplo da esfera complexa. Inicialmente, são apresentados conceitos básicos para introduzir a média extrínseca e posteriormente é apresentado um algoritmo para computar a mediana extrínseca de Fréchet. O segundo objetivo é desenvolver um procedimento de detecção de *outliers* baseado na mediana extrínseca projetada. Por fim, são feitos estudos de simulação para comparar as duas medidas de posição.

5.1 CONCEITOS BÁSICOS PARA DEFINIR A MEDIANA EXTRÍNSECA DE FRÉCHET

Inicialmente, é apresentada a média extrínseca de Fréchet na variedade $\mathcal{M}$, tal como discutido em Dryden e Mardia (2016). Seja $d(\cdot,\cdot)$ uma distância apropriada a uma variedade $\mathcal{M}$, então é possível definir a seguinte média populacional:

$$\boldsymbol{\mu}_{\mathcal{M}} = \underset{\boldsymbol{\mu}}{\arg\inf} \quad \mathbb{E}_X \left[d^2(\boldsymbol{X}, \boldsymbol{\mu}) \right], \tag{5.1}$$

em que $\mathbb{E}_X[\cdot]$ é o operador esperança para $\boldsymbol{X} \in \mathcal{M}$. Se (5.1) é um mínimo global, então $\boldsymbol{\mu}_{\mathcal{M}}$ é conhecido como a média de Fréchet. Similarmente, a variância de Fréchet é denotada por

$$\sigma^2_{\mathcal{M}} = \underset{\boldsymbol{\mu}}{\inf} \quad \mathbb{E}_X \left[d^2(\boldsymbol{X}, \boldsymbol{\mu}) \right], \tag{5.2}$$

em que (5.2) é o mínimo global.

Seja $j : \mathcal{M} \to \mathbb{R}^p$ uma aplicação de $\mathcal{M}$ no espaço Euclidiano $\mathbb{R}^p$ e Q uma medida de probabilidade em $\mathcal{M}$. A média populacional extrínseca de Fréchet, para $\boldsymbol{X} \in \mathcal{M}$, é dada por

$$\underset{\boldsymbol{\mu}}{\arg\inf} \quad \mathbb{E}_X \left[d^2(\boldsymbol{X}, \boldsymbol{\mu}) \right] = \underset{\boldsymbol{\mu}}{\arg\inf} \quad \mathbb{E}_X \left[||j(\boldsymbol{X}) - \boldsymbol{\mu}||^2 \right],$$

ou equivalentemente, como discutido em Bhattacharya e Bhattacharya (2012), a média populacional extrínseca de Fréchet pode ser escrita como

$$\underset{\boldsymbol{x} \in M}{\arg\min} \quad \left[\int_{\boldsymbol{y} \in M} ||j(\boldsymbol{x}) - j(\boldsymbol{y})||^2 \, dQ(\boldsymbol{y}) \right], \tag{5.3}$$

em que $\boldsymbol{x}$ é uma possível realização de $\boldsymbol{X}$, $||\cdot||$ é a norma euclidiana utilizada no $\mathbb{R}^p$. Quando o $\arg\min$ não é único, a média de Fréchet consiste em um conjunto de pontos em $\mathcal{M}$.

Para uma série de variedades mais simples, contudo não menos importante (tais como esferas, espaços reais e complexos projetados e certos espaços matriciais), se a média extrínseca de Fréchet for definida de maneira única, então ela pode ser calculada de forma equivalente a (5.3) a partir de duas etapas para dois métodos diferentes, conforme a discussão subsequente.

Método 1:

Etapa 1: Calcule a média:

$$\boldsymbol{\tau}_{mean} = \int_{\boldsymbol{y}\in M} j(\boldsymbol{y})\, dQ(\boldsymbol{y}). \tag{5.4}$$

Etapa 2: Projete $\boldsymbol{\tau}_{mean} \in \mathbb{R}^p$ no espaço $j(\mathcal{M}) \subset \mathbb{R}^p$.

Esta tese tem o objetivo de desenvolver uma alternativa robusta e analiticamente simples para a média extrínseca definida em (5.3), que seja útil em uma variedade de espaços não euclidianos. Naturalmente, um estimador a ser considerado é a mediana extrínseca de Fréchet, definida por

$$\arg\min_{\boldsymbol{x}\in M} \left[\int_{\boldsymbol{y}\in M} ||j(\boldsymbol{x}) - j(\boldsymbol{y})||\, dQ(\boldsymbol{y}) \right]. \tag{5.5}$$

Esta última definição decorre de uma analogia à mediana espacial tal como discutido por Small (1990). Entretanto, o processo de minimização envolvido em (5.5) pode ser um desafio, pelo menos em alguns cenários. Como alternativa a minimização de (5.5), é considerado um procedimento análogo ao Método 1, como segue.

Método 2

Etapa 1: Calcule a mediana espacial de $j(Y)$, em que $Y \sim Q$,:

$$\boldsymbol{\tau}_{med} = \arg\min_{\boldsymbol{\tau}\in\mathbb{R}^p} \left[\int_{\boldsymbol{y}\in M} ||j(\boldsymbol{y}) - \boldsymbol{\tau}||\, dQ(y) \right].$$

Etapa 2: Projete $\boldsymbol{\tau}_{med}$ em $j(\mathcal{M}) \subset \mathbb{R}^p$ para obter $\tilde{\mu}$.

Observe que a mediana espacial é única, exceto em circunstâncias especiais. Isso também se aplica a versão amostral. Para as variedades consideradas nesta tese, as projeções (Euclidianas) na Etapa 2 acima também são definidas de maneira única.

5.2 PROPOSTA TEÓRICA DE UMA EXPRESSÃO PARA A MEDIANA

Nesta seção, é definida a mediana extrínseca projetada de Fréchet em uma série de casos de interesse.

5.2.1 Noções preliminares

Sejam $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{r \times p}$ matrizes reais $r \times p$. A distância euclidiana entre $\boldsymbol{A}$ e $\boldsymbol{B}$ é definida como

$$d(\boldsymbol{A}, \boldsymbol{B}) = \left[\mathrm{tr}\left\{ (\boldsymbol{A} - \boldsymbol{B})^\top (\boldsymbol{A} - \boldsymbol{B}) \right\}\right]^{1/2},$$

em que $\mathrm{tr}(\cdot)$ define o traço matricial. No caso em que $\boldsymbol{A}$ e $\boldsymbol{B}$ são matrizes $r \times p$ com elementos complexos, define-se

$$d(\boldsymbol{A}, \boldsymbol{B}) = \left[\mathrm{tr}\left\{(\boldsymbol{A} - \boldsymbol{B})^*(\boldsymbol{A} - \boldsymbol{B})\right\}\right]^{1/2},$$

em que * denota o transposto conjugado.

5.2.2 Esfera real

Sejam $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \boldsymbol{S}^{p-1}$ uma amostra observada de vetores unitários na hiperesfera $\boldsymbol{S}^{p-1} = \{\boldsymbol{X} \in \mathbb{R} : \boldsymbol{X}^\top \boldsymbol{X} = 1\}$. Para calcular a mediana espacial da amostra, veja Etapa 1 do Método 2, encontra-se

$$\hat{\boldsymbol{\tau}}_{med} = \arg\min_{\boldsymbol{\tau} \in \mathbb{R}} \sum_{i=1}^{n} d(\boldsymbol{X}_i, \boldsymbol{\tau}).$$

Considerando a Etapa 2 do Método 2, é definido que

$$\tilde{\boldsymbol{\mu}}_{med} = \hat{\boldsymbol{\tau}}_{med} / ||\hat{\boldsymbol{\tau}}_{med}||.$$

Claro que a projeção só é bem definida se $\hat{\boldsymbol{\tau}} \neq \mathbf{0}_p$, em que $\mathbf{0}_p$ é o vetor contendo p zeros.

5.2.3 Espaço real projetivo

O espaço projetivo $\mathbb{R}\mathbf{P}^p$ é uma variedade formada por todas as linhas que passam pela origem em $\mathbb{R}^{p+1}$. Aqui, a situação é similar a esfera $\mathbf{S}^p$. Contudo, identificam-se os pontos antipodais $\pm\boldsymbol{X}_i$, configurando um espaço axial para $p = 1$ ou 2 (BHATTACHARYA; BHATTACHARYA, 2012). Aqui, é definido que

$$\tilde{\boldsymbol{\tau}}_{med} = \arg\min_{\boldsymbol{\tau} \in \mathbb{R}^{p \times p}} \sum_{i=1}^{n} d(\boldsymbol{X}_i \boldsymbol{X}_i^\top, \boldsymbol{\tau}).$$

Observe que isso equivale a encontrar a mediana espacial no espaço Euclidiano de dimensão p^2. É possível mostrar que $\hat{\boldsymbol{\tau}}_{med}$ é simétrico e não negativo definido. Logo, possui autovalores não negativos. Na etapa de projeção é encontrada a matriz de autovalores correspondente ao maior autovalor de $\hat{\boldsymbol{\tau}}_{med}$, se $\hat{\boldsymbol{\tau}}_{med} = \sum_{k=1}^{p} \lambda_k \boldsymbol{q}_k \boldsymbol{q}_k^\top$, em que $0 \leq \lambda_1, \ldots, \leq \lambda_{k-1} < \lambda_k$, então a projeção de $\hat{\boldsymbol{\tau}}_{med}$ no espaço $\{\boldsymbol{X}\boldsymbol{X}^\top : \boldsymbol{X} \in \boldsymbol{S}^{p-1}\}$ é dada por

$$\tilde{\boldsymbol{\mu}}_{med} = \boldsymbol{q}_k \boldsymbol{q}_k^\top,$$

em que $\boldsymbol{q}_k$ é um autovetor correspondente ao maior autovalor λ_k.

5.2.4 Espaço complexo projetivo

O espaço complexo projetivo $\mathbb{C}\boldsymbol{P}^{k-2}$ é o espaço de todas as linhas complexas que passam pela origem em $\mathbb{C}^{k-1}$. Ele acaba sendo o espaço de forma para formas de objetos em duas dimensões. Ou seja, os k *landmarks* são denotados por k número complexos, $(z_j = x_j + iy_j : 1 \leq j \leq k)$, e passam a ser representados no plano complexo (BHATTACHARYA; BHATTACHARYA, 2012). Tecnicamente, este caso é semelhante ao anterior, exceto por definir

$$\hat{\boldsymbol{\tau}}_{med} = \underset{\boldsymbol{\tau} \in \mathbb{C}^{p \times p}}{\arg\min} \sum_{i=1}^{n} d(\boldsymbol{X}_i \boldsymbol{X}_i^\star, \boldsymbol{\tau}).$$

Este procedimento é equivalente a encontrar a mediana espacial no espaço real de dimensão $2p^2$. A matriz resultante $\hat{\boldsymbol{\tau}}_{med}$ é uma matriz Hermitiana de ordem $p \times p$, não negativa definida. Além disso, se o maior autovalor, dito λ_k, é distinto, então a projeção de $\hat{\boldsymbol{\tau}}_{med}$ no espaço $\{\boldsymbol{X}\boldsymbol{X}^\top : \boldsymbol{X} \in C^p, \boldsymbol{X}^\star\boldsymbol{X} = 1\}$ é dada por

$$\tilde{\boldsymbol{\mu}}_{med} = \boldsymbol{q}_k \boldsymbol{q}_k^\star,$$

em que $\boldsymbol{q}_k$ é um autovetor correspondente ao maior autovalor λ_k.

5.3 COMO COMPUTAR A MEDIANA EXTRÍNSECA PROJETADA DE FRÉCHET

O primeiro passo para construir a mediana extrínseca para o caso Hermitiano é definir a matriz $\boldsymbol{T}$ de ordem $j \times k$ com $j, k = 1, \ldots, p$. Logo, os elementos matriciais t_{jk} podem assumir o seguintes valores

$$t_{jk} = \begin{cases} 0, & \text{se} \quad j > k, \\ 1, & \text{se} \quad j \leq k \end{cases}. \tag{5.6}$$

Ou seja, t_{jk} pode assumir o valor 0 ou 1 dependendo dos índices j, k. Uma operação útil para a construção da mediana extrínseca é o produto Hadamard de duas matrizes, definido a seguir.

Seja $\boldsymbol{A} = (a_{jk})$ e $\boldsymbol{B} = (b_{jk})$ duas matrizes, então, o produto Hadamard é dado por

$$\boldsymbol{C} = \boldsymbol{A} \circ \boldsymbol{B} = (c_{jk}),$$

em que $c_{jk} = a_{jk}b_{jk}$. Ou seja, basta multiplicar os elementos de $\boldsymbol{A}$ e $\boldsymbol{B}$ que possuam os mesmos índices.

Considere uma amostra de vetores unitários complexos $\boldsymbol{z}_j = u_j + iv_j$, $j = 1, \ldots, n$ e escreva

$$\boldsymbol{z}_j\boldsymbol{z}_j^* = (u_j + iv_j)(u_j^\top - iv_j^\top) = u_ju_j^\top + v_jv_j^\top + i(v_ju_j^\top - u_jv_j^\top) = \boldsymbol{U}_j + i\boldsymbol{V}_j,$$

em que

$$\boldsymbol{U}_j = u_ju_j^\top + v_jv_j^\top$$

e

$$\boldsymbol{V}_j = v_ju_j - u_jv_j$$

são matrizes reais $p \times p$ e $i = \sqrt{-1}$. Aqui, o transposto conjugado de $\boldsymbol{z}$ é denotado por $\boldsymbol{z}^*$.

Uma matriz complexa $p \times p$, $\boldsymbol{Z} = \boldsymbol{U} + i\boldsymbol{V}$, é dita ser Hermitiana se $\boldsymbol{Z}^\star = \boldsymbol{Z}$. Isto ocorre se e somente se $\boldsymbol{U}$ é simétrica ($\boldsymbol{U}^\top = \boldsymbol{U}$) e $\boldsymbol{V}$ é anti-simétrica ($\boldsymbol{V}^\top = -\boldsymbol{V}$). Dada uma matriz real $p \times p$, $\boldsymbol{\Theta} = (\theta_{jk})$, defina a matriz Hermitiana

$$\boldsymbol{Z}(\boldsymbol{\Theta}) = \boldsymbol{U}(\boldsymbol{\Theta}) + i\boldsymbol{V}(\boldsymbol{\Theta}),$$

em que

$$\boldsymbol{U}(\boldsymbol{\Theta}) = \boldsymbol{\Theta} - \boldsymbol{\Theta} \circ (\boldsymbol{T}^\top) + (\boldsymbol{\Theta} \circ \boldsymbol{T})^\top,$$

$$\boldsymbol{V}(\boldsymbol{\Theta}) = \boldsymbol{\Theta} \circ (\boldsymbol{T}^\top) - (\boldsymbol{\Theta} \circ (\boldsymbol{T})^\top)^\top,$$

em que $\boldsymbol{T}$ é definido em (5.6). Aqui, o triângulo inferior e diagonal da matriz $\boldsymbol{\Theta}$ coincidem respectivamente com o triângulo inferior e diagonal da matriz $\boldsymbol{U}(\boldsymbol{\Theta})$. É observado que o triângulo superior de $\boldsymbol{\Theta}$ é o triângulo superior de $\boldsymbol{V}(\boldsymbol{\Theta})$.

A construção da mediana extrínseca para o caso Hermitiano considera a seguinte função objetiva

$$F(\boldsymbol{\Theta}) = \sum_{j=1}^{n} d_H\left(\boldsymbol{Z}_j, \boldsymbol{Z}(\boldsymbol{\Theta})\right),$$

em que $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n, \boldsymbol{Z}(\boldsymbol{\Theta})$ são definidos acima e

$$d_H\left(\boldsymbol{Z}_j, \boldsymbol{Z}(\boldsymbol{\Theta})\right) = \left[tr\left\{\left(\boldsymbol{U}_j - \boldsymbol{U}(\boldsymbol{\Theta})\right)^2\right\} + tr\left\{\left(\boldsymbol{V}_j - \boldsymbol{V}(\boldsymbol{\Theta})\right)^2\right\}\right]^{1/2},$$

em que $tr(\cdot)$ denota o traço matricial. A função objetiva permite encontrar o $\hat{\theta}$ que minimiza $F(\boldsymbol{\Theta})$ a partir de um chute inicial, determinado pelos dados amostrais, obtendo assim $\hat{\boldsymbol{\Theta}}$. Então a mediana extrínseca projetada é o vetor unitário complexo correspondente ao maior autovalor de $\boldsymbol{Z}(\hat{\boldsymbol{\Theta}})$.

Uma ideia para um chute inicial para $\boldsymbol{\Theta}$ é considerar $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, em que $\boldsymbol{\Theta}_0$ é definido como

$$\boldsymbol{Z}(\boldsymbol{\Theta}_0) = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{z}_j \boldsymbol{z}_j^*.$$

No procedimento de otimização implementado no R é necessário expressar $\boldsymbol{\Theta}$ como um vetor, mais convenientemente por definir $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta})$ em que $\boldsymbol{\theta}$ é um vetor $p^2 \times 1$.

Definição 2 *(SEBER, 2008) Seja $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2 \cdots, \boldsymbol{a}_n)$ uma matriz de ordem $m \times n$.Então, $vec(\boldsymbol{A})$ é um vetor obtido ao empilhar as colunas de $\boldsymbol{A}$, dado por*

$$\mathrm{vec}(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix},$$

formando um vetor $mn \times 1$. Ou seja, $\mathrm{vec}(\boldsymbol{A})$ significa um vetor de colunas de $\boldsymbol{A}$. Por exemplo, se

$$\boldsymbol{A}_{2\times 3} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, \quad \textit{então}, \quad \mathrm{vec}(\boldsymbol{A}) = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \\ a_{13} \\ a_{23} \end{pmatrix}.$$

5.4 PROCEDIMENTO PARA DETECÇÃO DE *OUTLIER*

Uma segunda contribuição para dados de forma é um método de detecção de *outliers*, que usa a mediana extrínseca como referência. Este método é baseado nos procedimentos de Nascimento, Amaral G. J. A. e Cruz (2016) e Hoaglin e Iglewicz (1987) e permite a detecção de *outliers* em dados axiais e de forma.

A proposta desta seção é substituir a média extrínseca pela mediana extrínseca projetada de Fréchet, definida no algoritmo 1. Aqui, a distância utilizada foi a Riemanniana ρ.

Algoritmo 1: Algoritmo para detectar *outliers* com base na mediana extrínseca de Fréchet

---

Etapa 1: Considere uma amostra aleatória de eixos ou pré-formas dada por $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

Etapa 2: As distâncias são calculadas por

$$d_i = d(\tilde{\boldsymbol{\mu}}_{med}, \boldsymbol{X}_i) \quad \text{para} \quad i = 1, \ldots, n.$$

Etapa 3: Calcule as estatísticas de ordem: $d(1) \leq \ldots \leq d(n)$.

Etapa 4: Calcule as quantidades $F_L = d(f)$ e $F_U = d(n + 1 - f)$, em que $f = \frac{1}{2}[(n + 3)/2]$ e o operador $[.]$ significa o maior inteiro menor que o argumento.

Etapa 5: A i-ésima observação será rotulada como um *outlier* se

$$d_i > F_U + 2.5(F_U - F_L),$$

O experimento a seguir considerou 5000 réplicas de Monte Carlo com tamanhos amostrais 50 e 100, gerados para alta concentração. A inclusão dos *outliers* é feita ao substituir 5 observações extremas (geradas para baixa concentração) provenientes da distribuição Bingham complexa. Para cada amostra gerada, tanto a mediana extrínseca projetada como a forma média foram calculadas. Em seguida, calcula-se a distância dessas medidas para cada $\mathbf{X}_i$. Então, os procedimentos de Hoaglin e Iglewicz (1987) são computados em cima das distâncias para cada abordagem. Finalmente, os *outliers* são identificados e as taxas são calculadas, tomando o número de amostras detectadas com *outlier* sobre o número de réplicas.

Os resultados na Tabela 7, para $n = 50$, mostram que a mediana extrínseca projetada de Fréchet tem uma taxa de detecção de *outlier* mais alta. Este resultado se inverte quando tamanho da amostra cresce, uma vez que a média apresenta uma boa propriedade de robustez para grandes amostras. A mediana extrínseca projetada de Fréchet tem um desempenho melhor do que a forma média quando a concentração dos dados é aumentada consideravelmente. Ou seja, neste cenário os *outliers* ficam mais evidentes e são facilmente capturados pela nova abordagem.

Tabela 7 – Taxa de detecção de *outliers* para mediana extrínseca projetada de Fréchet e forma média. Admitindo 5 *outlier* nos dados.

		Taxa de detecção de *outlier*			
Alta concentração	Baixa concentração	$n = 50$		$n = 100$	
$\lambda = (\lambda_1, \lambda_2, \lambda_3)$	$\lambda = (\lambda_1, \lambda_2, \lambda_3)$	Mediana	Média	Mediana	Média
(20,30,300)	(30,40,50)	0.8506	0.8478	0.8484	0.8454
(10,15,400)	(15,25,40)	0.8488	0.8444	0.8516	0.8542
(50,55,400)	(25,35,40)	0.9160	0.9126	0.9272	0.9262
(2,4,500)	(1,6,8)	0.9340	0.9310	0.9452	0.9456
(2,4,800)	(1,3,5)	0.9390	0.9422	0.9492	0.9504
(1,3,800)	(2,5,10)	0.9338	0.9318	0.9484	0.9504
(5,8,900)	(10,15,35)	0.9086	0.9052	0.9224	0.9192
(20,30,900)	(30,40,50)	0.9264	0.9260	0.9376	0.9358

Fonte: O autor (2021).

5.5 RESULTADOS NUMÉRICOS

Nesta seção, é realizado um estudo de simulação, utilizando 5000 réplicas de Monte Carlo, para avaliar o desempenho da mediana extrínseca projetada de Fréchet. Para tanto, foram geradas amostras de tamanho 50 e 100 provenientes da distribuição Bingham complexa. Aqui, os *outliers* gerados para baixa concentração são substituídos na amostra (matendo o total de observações). Por exemplo, considere um cenário com 1 *outlier* e seja n o tamanho da amostra. Então, $n - 1$ observações são geradas a partir da distribuição Bingham complexa altamente concentrada e uma observação é proveniente da distribuição Bingham complexa pouco concentrada. Em seguida, são avaliadas a forma média e a mediana extrínseca projetada.

A distância Riemanniana para a forma média é calculada como

$$d_{\hat{\mu}} = \arccos |\hat{\boldsymbol{\mu}}^{\star}_{nout} \hat{\boldsymbol{\mu}}_{nwout}|, \tag{5.7}$$

em que $\hat{\boldsymbol{\mu}}_{nout}$ é a forma média de uma amostra de tamanho n sem *outlier*, $\hat{\boldsymbol{\mu}}_{nwout}$ é a forma média de uma amostra de tamanho n com pelo menos um *outlier* e $|.|$ é o módulo do número complexo.

A distância Riemanniana para a mediana extrínseca projetada é calculada por

$$d_{\tilde{\mu}} = \arccos |\tilde{\boldsymbol{\mu}}^{\star}_{nout} \tilde{\boldsymbol{\mu}}_{nwout}|, \tag{5.8}$$

em que $*$ denota o transposto conjugado, $\tilde{\boldsymbol{\mu}}_{nout}$ é a mediana extrínseca projetada de uma amostra de tamanho n sem *outliers* e $\tilde{\boldsymbol{\mu}}_{nwout}$ é a mediana extrínseca projetada de uma amostra de tamanho n com pelo menos um *outlier*.

As Tabelas 8 e 9 apresentam os resultados do experimento descrito acima. Em cada réplica de Monte Carlo, as amostras provenientes da distribuição Bingham complexa são contaminadas por um e cinco *outliers*. A mediana extrínseca projetada de Fréchet e a forma média são computadas para as amostras com e sem *outliers* e, subsequentemente, suas distâncias são obtidas a partir das Equações (5.8) e (5.7), respectivamente. Os resultados numéricos mostram que as distâncias são menores quando a mediana extrínseca projetada é utilizada. Em outras palavras, ela é menos influenciada por *outliers*. Além disso, o impacto sobre ela é reduzido com o aumento do tamanho amostral. A Tabela 10 apresenta o mesmo experimento para $k = 11$ *landmarks*. Os resultados mostram que mesmo para um grande número de *landmarks* a mediana proposta é mais robusta que a média extrínseca.

Tabela 8 – As distâncias (5.7) e (5.8) são calculadas para cada amostra com $k = 3$ *landmarks* e 1 *outlier*.

Alta concentração	Baixa concentração	Distância $\rho(\boldsymbol{p}, \boldsymbol{w})$			
		$n = 50$		$n = 100$	
$\lambda = (\lambda_1, \lambda_2, \lambda_3)$	$\lambda = (\lambda_1, \lambda_2, \lambda_3)$	Mediana	Média	Mediana	Média
(2, 4, 800)	(1, 3, 5)	0.00128	0.00866	0.00064	0.00432
(2, 4, 700)	(1, 3, 5)	0.00138	0.00866	0.00068	0.00428
(2, 4, 600)	(1, 4, 6)	0.00150	0.00867	0.00074	0.00429
(2, 4, 500)	(1, 6, 8)	0.00166	0.00862	0.00081	0.00427

Fonte: O autor (2021).

Tabela 9 – As distâncias (5.7) e (5.8) são calculadas para cada amostra com $k = 3$ *landmarks* e 5 *outliers*.

Alta concentração	Baixa concentração	Distância $\rho(\boldsymbol{p}, \boldsymbol{w})$			
		$n = 50$		$n = 100$	
$\lambda = (\lambda_1, \lambda_2, \lambda_3)$	$\lambda = (\lambda_1, \lambda_2, \lambda_3)$	Mediana	Média	Mediana	Média
(2, 4, 800)	(1, 3, 5)	0.00143	0.00955	0.00143	0.00960
(2, 4, 700)	(1, 3, 5)	0.00320	0.01999	0.00153	0.00958
(2, 4, 600)	(1, 4, 6)	0.00346	0.01987	0.00167	0.00960
(2, 4, 500)	(1, 6, 8)	0.00382	0.01978	0.00186	0.00952

Fonte: O autor (2021).

Tabela 10 – As distâncias (5.7) e (5.8) são calculadas para cada amostra com $k = 11$ *landmarks* e 5 *outliers*.

Alta concentração	Baixa concentração	Distância $\rho(\boldsymbol{p}, \boldsymbol{w})$			
		$n = 50$		$n = 100$	
$\lambda = (\lambda_1, \cdots, \lambda_{11})$	$\lambda = (\lambda_1, \cdots, \lambda_{11})$	Mediana	Média	Mediana	Média
$(1, 3, \cdots, 11, 14, 700)$	$(1, 3, \cdots, 11, 12, 13)$	0.0059	0.0178	0.0029	0.0085
$(18, 20, \cdots, 50, 51, 600)$	$(13, 15, \cdots, 23, 25, 26)$	0.0067	0.0187	0.0032	0.0089
$(30, 33, \cdots, 53, 54, 500)$	$(30, 33, \cdots, 51, 55, 56)$	0.0076	0.0204	0.0036	0.0098

Fonte: O autor (2021).

5.5.1 Aplicação: dados de microfósseis

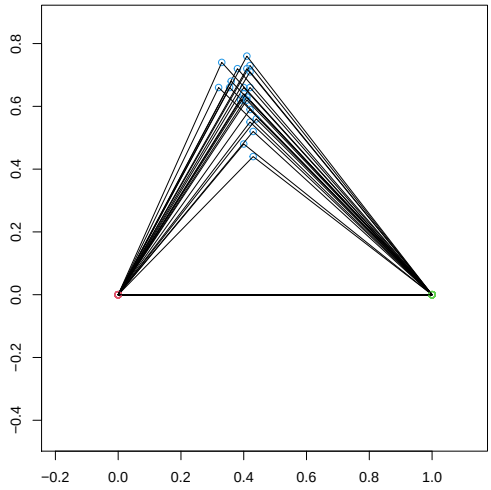
A base de dados a seguir se refere a uma amostra de 21 contornos médios de microfósseis, obtidos em diferentes latitudes no sul do Oceano Índico, publicados por Lohmann (1983) e disponível em Dryden e Mardia (2016). De modo geral, os dados disponíveis em Dryden e Mardia (2016) são essencialmente concentrados. Logo, é esperado que os valores da forma média e mediana sejam bem próximos. Utilizando os procedimentos discutidos no texto os valores da forma média ($\hat{\boldsymbol{\mu}}$) e mediana extrínseca projetada de Fréchet ($\tilde{\boldsymbol{\mu}}$) são dados, respectivamente, por

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} -0.8003549 + 0i \\ 0.0923014 - 0.5923787i \end{pmatrix} \quad \text{e} \quad \tilde{\boldsymbol{\mu}} = \begin{pmatrix} -0.8041153 + 0i \\ 0.0933511 - 0.5870981i \end{pmatrix},$$

em que a distância Riemanniana entre esses dois vetores complexos é 0.00648. A Figura 26 ilustra os 3 *landmarks*, identificados por círculos em cada um dos contornos dos dados de

microfósseis. Aqui, é possível visualizar a proximidade dos objetos, avaliados em 3 *landmarks*. Utilizando a abordagem de detecção para *outliers*, discutido no texto, é possível destacar a observação 26 (máximo), $d_{\tilde{\mu}21} = 0.1678 \notin (-0.0427;\ 0.1389)$, como *outlier*. Logo, a mediana extrínseca projetada de Fréchet pode ser uma medida mais apropriada e robusta para lidar com estes dados.

Figura 26 – *Landmarks* da amostra de microfósseis.



Fonte: O autor (2021).

6 CONCLUSÃO

Nesta tese, duas distribuições de probabilidade, desenvolvidas a partir de uma transformação baseada em distância denotadas pelas leis $\mathrm{TD}_1(\kappa)$ e $\mathrm{TD}_2(\kappa)$, foram propostas para lidar com o fenômeno da concentração em dados esféricos. Os novos modelos foram utilizados como mecanismo para checar alta ou baixa concentração, que é uma condição frequentemente assumida por resultados assintóticos na área de análise de dados esféricos. A abordagem proposta lida com o problema de concentração em dados axiais e direcionais a partir da análise univariada de uma medida de distância compacta em $[0, 1]$, ao invés de lidar com pontos na esfera real como usualmente feito. O primeiro modelo, denotado por $\mathrm{TD}_1(\kappa)$, foi construído a partir da lei von Mises-Fisher para dados direcionais. Algumas de suas propriedades matemáticas foram derivadas como: média, variância, curtose, assimetria e função geradora de momentos. Foram apresentadas expressões para computar as estimativas pontuais e realizar testes de hipóteses envolvendo estatísticas clássicas. Os experimentos de Monte Carlo indicaram que: (i) em geral, os testes de hipóteses são mais precisos para grandes valores de κ; (ii) o teste de Wald é mais liberal para pequenos valores de κ (apresentando uma probabilidade estimada do erro tipo I maior do que o nível nominal adotado); (iii) o teste escore foi o mais conservador entre os três.

O segundo modelo, construído sob suposição de distribuição Watson e denotado por $\mathrm{TD}_2(\kappa)$, apresenta um único parâmetro e permite modelar observações axiais que se concentram tanto nos polos quanto nas regiões equatoriais. Para $\mathrm{TD}_2(\kappa)$, foram apresentadas expressões para computar as estimativas por máxima verossimilhança e obter a matriz de informação de Fisher. Os estudos de simulação mostraram que: (i) as estimativas pontuais são mais precisas para valores de $\kappa > 0$; o teste da razão de verossimilhanças tendem a ser mais liberal para $\kappa > 1$, enquanto que os testes Wald e escore são mais conservadores para $\kappa > 0$.

Uma vez estudadas e propostas as distribuições, elas foram utilizadas como elementos centrais no desenvolvimento de estatísticas de testes para dados direcionais (a saber uma função de $\mathrm{TD}_1(\kappa)$) e axiais (uma função de $\mathrm{TD}_2(\kappa)$). Adicionalmente foram obtidas as distribuições exatas dessas estatísticas. Duas aplicações foram feitas para ilustrar as propostas em dados esféricos. Os resultados mostraram que o uso dos novos paradigmas propostos conseguem detectar de modo simples (transferindo o problema de uma esfera real para o intervalo $[0, 1]$) e eficiente alta concentração em amostras esféricas.

Um segunda parte desta tese se dedicou a proposta de métodos baseados na mediana extrínseca como alternativa a média extrínseca de Fréchet, que tem fórmula analítica intratável. Foram apresentadas fórmulas matemáticas para computar a mediana extrínseca projetada e procedimentos para detecção de *outliers* baseados nessa medida. Estudos numéricos por simulação de Monte Carlo foram realizados para quantificar a robustez da nova mediana em termos da distribuição Bingham complexa (para o caso de formas planares). Os resultados mostraram que a mediana proposta é mais robusta que a forma média, principalmente para pequenos tamanhos de amostras. Uma aplicação feita aos dados de microfósseis mostrou que a mediana proposta é adequada para lidar com *outliers* nos dados.

TRABALHOS FUTUROS

Como pesquisas futuras, espera-se

- ✓ explorar outras distribuições de probabilidade com propriedade de simetria rotacional sobre a localização $\boldsymbol{\mu}$, tais como as distribuições de: Arnold (ARNOLD, 1941), Selby (SELBY, 1964) e Purkayastha (PURKAYASTHA, 1991), úteis para modelar dados esféricos;
- ✓ implementar no software R uma família de distribuições baseadas na função angular e as funções desta tese como um pacote;
- ✓ desenvolver um teste baseado em mediana para dados axiais;
- ✓ propor uma medida robusta, baseada na mediana extrínseca, para dados de forma em três dimensões.

REFERÊNCIAS

AMARAL, G. J. A.; WOOD, A. T. A. Empirical likelihood methods for two-dimensional shape analysis. *Biometrika*, v. 97, n. 3, p. 757–764, 2010.

AMIRI ABOUBACAR; THIAM, B.; VERDEBOUT, T. On the estimation of the density of a directional data stream. *Scandinavian Journal of Statistics*, v. 44, n. 1, p. 249–267, 2016.

ARNOLD, K. J. *On Spherical Probability Distributions*. Tese (Doutorado) — Massachusetts Institute of Technology, 1941.

AZNAR, J. C.; GLOAGUEN, E.; TAPSOBA, D.; HACHEM, S.; CAYA, D.; BéGIN, Y. Interpolation of monthly mean temperatures using cokriging in spherical coordinates. *International Journal of Climatology*, p. 1–12, 2012.

BARROS, C. M.; AMARAL, G. J. A.; NASCIMENTO, A. D. C.; CYSNEIROS, A. H. M. A. Detecting influential observations in watson data. *Communications in Statistics - Theory and Methods*, p. 1–31, 2016.

BHATTACHARYA, A.; BHATTACHARYA, R. *Nonparametric Inference on Manifolds (With Applications to Shape Spaces)*. [S.l.]: Cambridge University Press, 2012. ISBN 978-1107019584.

BINGHAM, M. A.; SCRAY, M. L. A permutation test for comparing rotational symmetry in three-dimensional rotation data sets. *Journal of Statistical Distributions and Applications*, Springer Open, v. 4, n. 19, p. 1–8, 2017.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2010. ISBN 9788585818821.

BOWMAN, A. W.; AZZALINI, A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. OUP Oxford, 1997. (Oxford Statistical Science Series). ISBN 9780191545696. Disponível em: <https://books.google.com.br/books?id=7WBMrZ9umRYC>.

BROCKWELL, P.; DAVIS, R. *Time Series: Theory and Methods: Theory and Methods*. [S.l.]: Springer New York, 1991. (Springer Series in Statistics). ISBN 9780387974293.

BROCKWELL, P. J.; DAVIS, R. A. *Time Series: Theory and Methods*. [S.l.]: Springer Series in Statistics, 1987. ISBN 978-1-4899-0006-7.

CABRERA, J.; WATSON, G. S. On a spherical median related distribution. *Communications in Statistics - Theory and Methods*, v. 19, n. 6, p. 1973–1986, 1990.

CARDOZO, N.; ALLMENDINGER, R. W. Spherical projections with osxstereonet. *Computers & Geosciences*, v. 51, p. 193–205, 2013.

CHIKUSE, Y. Concentrated matrixlangevin distributions. *Journal of Multivariate Analysis*, Academic Press, v. 85, n. 2, p. 375–394, 2003.

CORDEIRO, G. M.; SILVA, R. B.; NASCIMENTO, A. D. C. *Recent Advances in Lifetime and Reliability Models*. [S.l.]: Bentham Books, 2020.

DIGGLE, P. J.; FISHER, N. I. Sphere: A contouring program for spherical data. *Computers & Geosciences*, Pergamon Press Ltd, v. 11, n. 6, p. 725–766, 1985.

DIMROTH, E. 'Untersuchungen zum mechanismus von blastesis und syntexis in phylitten und hornfelsen des südwestlichen fichtelgebirges i. die statistische auswertung einfacher gürteldiagramme'. *Tschermaks Mineralogische und Petrographische Mitteilungen*, v. 8, p. 248–274, 1962.

DRYDEN, I. L.; MARDIA, K. V. *Statistical Shape Analysis with Applications in R*. [S.l.]: Wiley, 2016. (Wiley Series in Probability and Statistics). ISBN 0470699620.

FIGUEIREDO, A. Two-way analysis of variance for data from a concentrated bipolar watson distribution. *Journal of Applied Statistics*, Taylor & Francis, v. 33, n. 6, p. 575–581, 2006.

FIGUEIREDO, A. Multi-sample tests for axial data from watson distributions. *Advances in Statistical Analysis*, Springer-Verlag, v. 93, n. 4, p. 371–386, 2009.

FISHER, N. I. Robust tests for comparing the dispersions of several fisher or watson distributions on the sphere. *Geophysical Journal International*, v. 85, n. 3, p. 563–572, 1986.

FISHER, N. I.; LEWIS, T.; EMBLETON, B. J. J. *Statistical Analysis of Spherical Data*. [S.l.]: Cambridge University Press, 1993. ISBN 9780521456999.

FISHER, R. Dispersion on a sphere. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 217, n. 1130, p. 295–305, 1953.

FRANK, W. J. O.; DANIEL, W. L.; RONALD, F. B.; CHARLES, W. C. *NIST Handbook of Mathematical Functions*. [S.l.]: Cambridge University Press, 2010.

GRADSHTEYN, I.; RYZHIK, I. *Table of Integrals, Series, and Products*. Elsevier Science, 2000. ISBN 9780080542225. Disponível em: <https://books.google.com.br/books?id=h4y-36vKIZgC>.

HOAGLIN, D. C.; IGLEWICZ, B. Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, [American Statistical Association, Taylor & Francis, Ltd.], v. 82, p. 1147–1149, 1987. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2289392>.

HUO, Z.; ZHOU, J. An accelerated direct demodulation method for image reconstruction using spherical data from the hard x-ray modulation telescope. *Research in Astronomy and Astrophysics*, v. 13, n. 8, p. 991–1012, 2013.

KEILSON, J.; PETRONDAS, D.; SUMITA, U.; WELLNER, J. Significance points for some tests of uniformity on the sphere. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 17, n. 3, p. 195–218, 1983.

KENDALL, D. G. *The diffusion of shape. Advances in Applied Probability*. [S.l.]: Advances in Applied Probability, 1977.

KENT, J. T. The complex bingham distribution and shape analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 56, n. 2, p. 285–299, 1994.

KENT, J. T.; CONSTABLE, P. D.; ER, F. Simulation for the complex bingham distribution. *Statistics and Computing*, v. 14, p. 53–57, 2004.

KO, D. Robust estimation of the concentration parameter of the Von Mises-Fisher distribution. *Annals of Statistics*, The Institute of Mathematical Statistics, v. 20, n. 2, p. 917–928, 06 1992.

KUMMER, E. De integralibus quibusdam definitis et seriebus infinitis. *Journal Für Die Reine Und Angewandte Mathematik*, University of Arizona, v. 17, p. 228–242, 1837.

LEY, C.; VERDEBOUT, T. *Modern Directional Statistics*. [S.l.]: Chapman and Hall/CRC, 2017. ISBN 9781498706643.

LI, K.-H.; WONG, C. K.-F. Random sampling from the watson distribution. *Communications in Statistics - Simulation and Computation*, v. 22, n. 4, p. 997–1009, 1993.

LOHMANN, G. P. Eigenshape analysis of microfossils: A general morphometric procedure for describing changes in shape. *Mathematical Geology*, v. 15, n. 6, p. 659–672, 1983.

MARDIA, K. V.; JUPP, P. E. *Directional Statistics*. [S.l.]: Wiley, 1999. ISBN 0471953334.

MUJICA, A.; NAVA, M.; VARGAS, A. Dispersion of emerita analoga (stimpson, 1857) larvae in northern coast of chile (25°-31.5°s). *Latin American Journal of Aquatic Research*, v. 42, n. 3, p. 418–426, 2014.

NASCIMENTO, A. D. C.; AMARAL G. J. A., A. B. B.; CRUZ, J. T. M. Influential observation in complex normal data for problems in allometry. *Communications in Statistics - Theory and Methods*, v. 45, n. 9, p. 2714–2729, 2016.

NASCIMENTO, A. D. C.; SILVA, R. C. da; AMARAL, G. J. A. Distance-based hypothesis tests on the watson distribution. *Communications in Statistics - Simulation and Computation*, p. 1–14, 2018.

NEYMAN, J.; PEARSON, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 20A, n. 1/2, p. 175–240, 1928. ISSN 00063444.

PAINDAVEINE, D.; VERDEBOUT, T. Detecting the direction of a signal on high-dimensional spheres: non-null and le cam optimality results. *Probability Theory and Related Fields*, Springer, v. 176, p. 1165–1216, 2020.

PLAMONDON, R.; SRIHARI, S. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, p. 63–84, 2000.

PURKAYASTHA, S. A rotationally symmetric directional distribution obtained through maximum likelihood characterization. *Sankhyā: The Indian Journal of Statistics*, Indian Statistical Institute, v. 53, n. 1, p. 70–83, 1991.

R Core Team. *R: A language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>.

RAO, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, v. 44, n. 1, p. 50–57, 1948.

RAYLEIGH, L. On the problem of random vibrations, and of random flights in one, two, or three dimensions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, v. 37, n. 220, p. 321–347, 1919.

RELTON, F. E. *APPLIED BESSEL FUNCTIONS*. [S.l.]: Dover Publications, 1965.

RIZZO, M. L. *Statistical Computing with R*. [S.l.]: Chapman & Hall/CRC, 2007. ISBN 1584885459.

ROSS, S. M. *A First Course in Probability*. [S.l.]: Pearson Prentice Hall, 2010. ISBN 9780136033134.

SAU, M. F.; RODRIGUEZ, D. Minimum distance method for directional data and outlier detection. *Advances in Data Analysis and Classification*, Springer-Verlag, 2017.

SCHEIDEGGER, A. E. On the statics of the orientation of bedding planes, grain axes and similar sedimentological data. *US Geological Survey*, Geology, v. 525, p. 164–167, 1965.

SEBER, G. A. F. *A Matrix Handbook for Statisticians*. [S.l.]: Wiley-Interscience, 2008. ISBN 9780471748694.

SELBY, B. Girdle distributions on a sphere. *Biometrika*, University of Liverpoo, v. 51, n. 3-4, p. 381–392, 1964.

SER, G. Directional data analysis and an application. *Journal of Agricultural Sciences*, Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi, v. 24, n. 2, p. 121–126, 2014.

SMALL, C. G. A survey of multidimensional medians. *International Statistical Review*, v. 58, n. 3, p. 263–277, 1990.

SRA, S.; KARP, D. The multivariate watson distribution: Maximum-likelihood estimation and other aspects. *Journal of Multivariate Analysis*, ELSEVIER, v. 114, p. 256–269, 2013.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, American Mathematical Society, v. 54, n. 3, p. 426–482, 1943. ISSN 00029947. Disponível em: <http://www.jstor.org/stable/1990256>.

WATSON, G. S. The theory of concentrated langevin distributions. *Journal of Multivariate Analysis*, Academic Press, v. 14, p. 74–82, 1984.

ZHANG, Y.; WU, Y.; Y. JIANGUO; W., H.; R. RODRIGUEZ, J. A. P.; QIU, Y. 3d inversion of full gravity gradient tensor data in spherical coordinate system using local north-oriented frame. *Earth, Planets and Space*, v. 70, n. 1, p. 1–23, 2018.

APÊNDICE A – PRODUTO INTERNO

A definição de produto interno em um espaço vetorial real (complexo) $\mathbb{H}$ é tratada como uma generalização do produto escalar de dois vetores definidos em $\mathbb{R}^n$. Ou seja, no espaço Euclidiano com n dimensões.

Um espaço vetorial real $\mathbb{H}$ é dito ser um espaço com produto interno se para cada par de elementos $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}$ existe um número real $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ satisfazendo as condições (BROCKWELL; DAVIS, 1987):

1. Simetria: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle \ \forall \ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}$;
2. Distributividade: $\langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle + \langle \boldsymbol{y}, \boldsymbol{z} \rangle \ \forall \ \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{H}$;
3. Homogeneidade: $\langle \alpha \boldsymbol{x}, \boldsymbol{y} \rangle = \alpha \langle \boldsymbol{x}, \boldsymbol{y} \rangle \ \forall \ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}$ e $\alpha \in \mathbb{R}$;
4. Positividade: $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0 \ \forall \ \boldsymbol{x} \in \mathbb{H}$, com $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ se e somente se $\boldsymbol{x} = 0$,

em que $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ é chamado de produto interno de $\boldsymbol{x}$ e $\boldsymbol{y}$. Analogamente, um espaço vetorial complexo é dito ser um espaço munido de produto interno hermitiano se para cada par de elementos $\boldsymbol{x}$ e $\boldsymbol{y}$ em $\mathbb{H}$, existe um número complexo $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, satisfazendo as condições (2)–(4), e a condição (1) assumindo $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \overline{\langle \boldsymbol{y}, \boldsymbol{x} \rangle}$, em que o segundo termo dessa igualdade denota o complexo conjugado.

O conjuntos dos vetores colunas $\boldsymbol{x} \in \mathbb{R}^n$, com representação dada por

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = (x_1, \ldots, x_n)^\top. \tag{A.1}$$

é um espaço de produto interno real se definimos

$$\boldsymbol{x} = \sum_{i=1}^{n} x_i y_i. \tag{A.2}$$

Em termos matriciais (A.2) pode ser reescrito como

$$\boldsymbol{x} = \sum_{i=1}^{n} x_i y_i = \boldsymbol{y}^\top \boldsymbol{x} = \boldsymbol{y}^\top I_n \boldsymbol{x}, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \tag{A.3}$$

em que I_n é a matriz identidade de ordem n.

Analogamente, $\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathbb{C}^n$ (conjunto dos vetores complexos), é um espaço de produto interno complexo se definimos

$$\langle \boldsymbol{z}, \boldsymbol{w} \rangle = \sum_{i=1}^{n} z_i \bar{w}_i = \boldsymbol{w}^* \boldsymbol{z} = \boldsymbol{w}^* I_n \boldsymbol{z}, \quad \forall \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^n, \tag{A.4}$$

em que $\boldsymbol{w}^*$ é o transposto conjugado do vetor complexo $\boldsymbol{w}$ e I_n é a matriz identidade de ordem n.

Definição 1 (Norma) *(BROCKWELL; DAVIS, 1987) A norma de um elemento $\boldsymbol{x}$ de um espaço de produto interno é definido para ser*

$$||\boldsymbol{x}|| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}.$$

No espaço Euclidiano a norma de um vetor corresponde ao seu tamanho.

APÊNDICE B – FUNÇÕES ESPECIAIS

Ao longo desta tese, são apresentadas algumas funções especiais como: a função de Kummer e a função Bessel.

B.1 FUNÇÃO BESSEL

Em muitos problemas da física matemática, a função de Bessel costuma aparecer como solução das equações de Laplace para coordenadas cilíndricas. Além disso, existem diferentes tipos de funções Bessel com propriedades em comuns (RELTON, 1965).

A função de cilindro, denotada por $C_n(x)$, pode ser definida por duas formulas recorrentes

$$C_{n-1}(x) + C_{n+1}(x) = \frac{2n}{x} C_n(x) \tag{B.1}$$

e

$$C_{n-1}(x) - C_{n+1}(x) = 2\frac{d}{dx} C_n(x), \tag{B.2}$$

em que o parâmetro n é real e x é real positivo. Além disso, é possível obter as seguintes derivadas

$$C_n'(x) = \frac{d}{dx} C_n(x)$$

e

$$C_n'(ax) = \frac{d}{d(ax)} C_n(ax) = \frac{1}{a}\frac{d}{dx} C_n(ax).$$

As equações a seguir são obtidas por adição e subtração de (B.1) e (B.2).

$$xC_{n-1}(x) = nC_n(x) + xC_n'(x), \tag{B.3}$$

$$xC_{n+1}(x) = nC_n(x) - xC_n'(x). \tag{B.4}$$

B.1.1 Equações de Bessel

Derivando (B.3) obtém-se

$$xC_n''(x) + (n+1)C_n'(x) = xC_{n-1}'(x) + C_{n-1}(x). \tag{B.5}$$

Multiplicando (B.3) por n e subtraindo de (B.5) multiplicado por x, tem-se

$$x^2C_n''(x) + xC_n'(x) - n^2C_n(x) = x\left\{xC_{n-1}'(x) - (n-1)C_{n-1}(x)\right\} = -x^2C_n(x). \quad \text{(B.6)}$$

Utilizando (B.4), o resultado mostra que $C_n(x)$ satisfaz

$$x^2y'' + xy' + (x^2 - n^2)y = 0, \quad \text{(B.7)}$$

conhecida como equação de Bessel de ordem n.

B.1.2 Função Bessel modificada

Em algumas modelagens fenômenos físicos, estão presentes as funções Bessel que lidam com argumentos complexos. Neste caso, as soluções de uma equação Bessel são chamadas de função Bessel modificada. Considere a seguinte equação

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} - \left(1 + \frac{n^2}{x^2}\right)y = 0. \quad \text{(B.8)}$$

A solução da equação (B.8) é dada por

$$J_n(ix) = \frac{((1/2)ix)^n}{\Gamma(n+1)}\left\{1 - \frac{((1/2)ix)^2}{1(n+1)} + \cdots\right\} \quad \text{(B.9)}$$

Em seguida, é definida uma nova função para remover a parte imaginária indesejada, dada por

$$I_n(x) = \underset{100}{(i)}^{-n} J_n(ix) \quad \text{(B.10)}$$

com

$$I_n(x) = \frac{((1/2)x)^n}{\Gamma(n+1)}\left\{1 + \frac{((1/2)x)^2}{1(n+1)} + \frac{((1/2)x)^4}{1.2(n+1)(n+2)} + \cdots\right\}, \quad \text{(B.11)}$$

em que $I_n(x)$ é a função Bessel modificada de primeiro tipo e satisfaz a equação (B.8) (RELTON, 1965).

B.2 FUNÇÃO DE KUMMER

A função Kummer de primeiro tipo, denotada por $M(a, b, z)$ e introduzida por Kummer (1837), é tratada como a solução da equação diferencial de Kummer, dada por

$$z\frac{d^2w}{dz^2} + (b - z)\frac{dw}{dz} - aw = 0. \quad \text{(B.12)}$$

A equação (B.12) tem uma singularidade irregular no infinito de classficação 1 e uma singularidade regular na origem, com índices 0 e $1-b$ (FRANK et al., 2010). A primeira solução padrão de (B.12) é

$$M(a,\ b,\ z) = \sum_{s=0}^{\infty} \frac{(a)_s}{(b)_s s!} z^s = 1 + \frac{az}{b} + \frac{a(a+1)}{b(b+1)2!} z^2 + \cdots , \tag{B.13}$$

em que $a^{(0)} = 1$ e $a^{(s)} = a(a+1)(a+2)\cdots(a+n-1)$, contudo $M(a,b,z)$ não existe quando b é um inteiro não positivo (FRANK et al., 2010).

APÊNDICE C – ENCONTRANDO A DISTRIBUIÇÃO DE PROBABILIDADE DE TD_1 E TD_2

Seja $T = \boldsymbol{X}^\top \boldsymbol{\mu}$ e $\boldsymbol{X} \sim \text{vMF}_p(\boldsymbol{\mu}, \kappa)$ com densidade dada por

$$f_{T_p}(t) = a_p^{*-1}(\kappa)\exp(\kappa t)(1-t^2)^{(p-3)/2}, \text{ com } t \in (-1,1),$$

A distribuição da distância estocástica $D = 1 - T^2$, sob suposição de von Mises-Fisher, denotada por TD_1, é expressa em função da distribuição T. Logo,

$$\begin{aligned}\Pr(D \leq d) &= \Pr(1 - T^2 \leq d) \\ &= 1 - \Pr(-\sqrt{1-d} \leq T \leq \sqrt{1-d}) \\ &= 1 - \left[\Pr(T \leq \sqrt{1-d}) - \Pr(T \leq -\sqrt{1-d})\right].\end{aligned}$$

Daí,

$$F_{D_p}(d) = 1 - \left[F_{T_q}(\sqrt{1-d}) - F_{T_q}(-\sqrt{1-d})\right]. \tag{C.1}$$

Ao derivar (C.1) com respeito a d, valor observado de D, é obtido a função densidade de probabilidade para a variável aleatória D. Segue que

$$\begin{aligned}\frac{\partial F_{D_p}(d)}{\partial d} &= \frac{\partial\left\{1 - \left[F_{T_p}(\sqrt{1-d}) - F_{T_p}(-\sqrt{1-d})\right]\right\}}{\partial d} = \\ f_{D_p}(d) &= -\left[-f_{T_p}(\sqrt{1-d})\frac{1}{2\sqrt{1-d}} - f_{T_p}(-\sqrt{1-d})\frac{1}{2\sqrt{1-d}}\right] \\ f_{D_p}(d) &= \frac{1}{2\sqrt{1-d}}\left[f_{T_p}(\sqrt{1-d}) + f_{T_p}(-\sqrt{1-d})\right].\end{aligned}$$

A densidade da projeção $T = \boldsymbol{X}^\top \boldsymbol{\mu}$, denotada por f_{T_p}, explorada por meio da decomposição de vetores aleatórios na esfera, é essencial para determinar a distribuição de TD_1. Procedimento similar pode ser utilizado considerando outras distribuições rotacionalmente simétricas sobre uma locação $\mu \in \boldsymbol{S}^{p-1}$.

Agora, seja $T = \boldsymbol{X}^\top \boldsymbol{\mu}$ e $\boldsymbol{X} \sim \text{W}_p(\boldsymbol{\mu}, \kappa)$ com densidade dada por

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = c_{fa,p}\exp(\kappa[\boldsymbol{x}^T\boldsymbol{\mu}]^2), \tag{C.2}$$

Ao explorar a propriedade de simetria rotacional para $\boldsymbol{X} \sim \text{W}_p(\boldsymbol{\mu}, \kappa)$ é encontrado a seguinte densidade de T

$$f(t) = w_{p-1}c_{fa,p}f_a(t)(1-t^2)^{(p-3)/2}, \tag{C.3}$$

em que $w_p = 2\pi^{p/2}/\Gamma(p/2)$ e $c_{fa,p}$ refere-se a constante de normalização da distribuição Watson. A densidade de D, pode ser encontrada de forma similar ao caso von Mises-Fisher. Segue que

$$\begin{aligned}
\frac{\partial F_{D_p}(d)}{\partial d} &= \frac{\partial \left\{1 - \left[F_{T_p}(\sqrt{1-d}) - F_{T_p}(-\sqrt{1-d})\right]\right\}}{\partial d} = \\
f_{D_p}(d) &= -\left[-f_{T_p}(\sqrt{1-d})\frac{1}{2\sqrt{1-d}} - f_{T_p}(-\sqrt{1-d})\frac{1}{2\sqrt{1-d}}\right] \\
f_{D_p}(d) &= \frac{1}{2\sqrt{1-d}}\left[f_{T_p}(\sqrt{1-d}) + f_{T_p}(-\sqrt{1-d})\right],
\end{aligned}$$

em que f_{T_p} é a densidade da projeção $T = \boldsymbol{X}^\top \boldsymbol{\mu}$ para $\boldsymbol{X} \sim W(\boldsymbol{\mu}, \kappa)$. Então, a distribuição de probabilidade de D é definida por

$$F_{D_p}(d) = 1 - \left[F_{T_q}(\sqrt{1-d}) - F_{T_q}(-\sqrt{1-d})\right], \tag{C.4}$$

quando $\boldsymbol{X} \sim \mathrm{W}(\boldsymbol{\mu}, \kappa)$.

APÊNDICE D – FUNÇÃO GERADORA DE MOMENTOS PARA A DISTRIBUIÇÃO TD$_1$

Seja $\boldsymbol{X} \sim \text{vMF}(\boldsymbol{\mu}, \kappa)$. A função geradora de momentos de $D \sim TD_1(\kappa)$ é dada pela seguinte integração no suporte $[0, 1]$

$$
\begin{aligned}
M_D(t) &= \int_0^1 e^{tx} \frac{1}{2\sqrt{1-x}} \left[f_{T_p}(\sqrt{1-x}) + f_{T_p}(-\sqrt{1-x}) \right] dx \\
&= \frac{a_p^{*-1}(\kappa)}{2} \int_0^1 \frac{e^{tx}}{\sqrt{1-x}} \Big[e^{(\kappa\sqrt{1-x})} (1 - (\sqrt{1-x})^2)^{(p-3)/2} \\
&+ e^{(-\kappa\sqrt{1-x})} (1 - (-\sqrt{1-x})^2)^{(p-3)/2} \Big], dx \\
&= \frac{a_p^{*-1}(\kappa)}{2} \int_0^1 \frac{e^{tx}}{\sqrt{1-x}} \left[e^{(\kappa\sqrt{1-x})} (x)^{(p-3)/2} + e^{(-\kappa\sqrt{1-x})} (x)^{(p-3)/2} \right] dx \\
&= \frac{a_p^{*-1}(\kappa)}{2} \left[\int_0^1 \frac{e^{tx} e^{(\kappa\sqrt{1-x})} x^{(p-3)/2}}{\sqrt{1-x}} dx + \int_0^1 \frac{e^{tx} e^{(-\kappa\sqrt{1-x})} x^{(p-3)/2}}{\sqrt{1-x}} dx \right].
\end{aligned} \tag{D.1}
$$

Tomando $p = 3$, tem-se

$$
M_D(t) = \frac{a_3^{*-1}(\kappa)}{2} \left[\int_0^1 \frac{e^{tx} e^{(\kappa\sqrt{1-x})}}{\sqrt{1-x}} dx + \int_0^1 \frac{e^{tx} e^{(-\kappa\sqrt{1-x})}}{\sqrt{1-x}} dx \right]. \tag{D.2}
$$

Para obter a forma fechada de (D.2) utilizamos *softwares* de computação simbólica como o SageMath e o wxMaxima. Como resultado é obtido a seguinte expressão para a fgm

$$
M_D(t) = \frac{a_3^{*-1}(\kappa)\sqrt{\pi}}{2\sqrt{t}} \left[e^{t+\frac{\kappa^2}{4t}} \operatorname{erf}\left(\frac{2t+\kappa}{2\sqrt{t}} \right) + e^{t+\frac{\kappa^2}{4t}} \operatorname{erf}\left(\frac{2t-\kappa}{2\sqrt{t}} \right) \right], \tag{D.3}
$$

em que

$$
erf(d) = \Phi(d) = \frac{2}{\sqrt{\pi}} \int_0^d e^{t^2} \, \mathrm{d}t
$$

é a função de erro e $\Phi(\cdot)$ é a função de distribuição cumulativa da distribuição normal padrão (GRADSHTEYN; RYZHIK, 2000).

APÊNDICE E – FUNÇÃO GERADORA DE MOMENTOS PARA A DISTRIBUIÇÃO TD$_2$

Seja $\boldsymbol{X} \sim \mathrm{W}(\boldsymbol{\mu}, \kappa)$. A função geradora de momentos de $D \sim TD_2(\kappa)$ é dada por:

$$\begin{aligned} M_D(t) &= \frac{c_p(\kappa)}{2}\left[\int_0^1 \frac{e^{tx}e^{(\kappa(\sqrt{1-x})^2)}x^{(p-3)/2}}{\sqrt{1-x}}\,dx + \int_0^1 \frac{e^{tx}e^{(\kappa(-\sqrt{1-x})^2)}x^{(p-3)/2}}{\sqrt{1-x}}\,dx\right] \\ &= \frac{c_p(\kappa)}{2}\left[\int_0^1 \frac{e^{tx}e^{(\kappa(1-x))}x^{(p-3)/2}}{\sqrt{1-x}}\,dx + \int_0^1 \frac{e^{tx}e^{(\kappa(1-x))}x^{(p-3)/2}}{\sqrt{1-x}}\,dx\right] \\ &= \frac{c_p(\kappa)}{2}\left[2\int_0^1 \frac{e^{tx}e^{(\kappa(1-x))}x^{(p-3)/2}}{\sqrt{1-x}}\,dx\right] \\ &= c_p(\kappa)\left[\int_0^1 \frac{e^{tx}e^{(\kappa(1-x))}x^{(p-3)/2}}{\sqrt{1-x}}\,dx\right], \end{aligned} \tag{E.1}$$

em que

$$c_p(\kappa) = \frac{\Gamma(p/2)}{2\pi^{p/2}M(1/2, p/2, \kappa)}.$$

Tomando $p = 3$

$$M_D(t) = c_3(\kappa)\left[\int_0^1 \frac{e^{tx}e^{(\kappa(1-x))}}{\sqrt{1-x}}\,dx\right], \tag{E.2}$$

Devido à complexidade da integral em (E.1) não foi possível encontrar uma solução analítica para a fgm. No entanto, utilizando o *sagemath* ou *wxmaxima*, é possível obter o k-ésimo momento ordinário de D com $k \in \mathbb{N}^*$ e $p = 3$:

$$\mathbb{E}(D) = -\frac{c_p(\kappa)\,\left(2\kappa\, e^{\kappa} + \sqrt{\pi}\,\sqrt{-\kappa}\,\left(2\,\mathrm{erf}\left(\sqrt{-\kappa}\right)\kappa + \mathrm{erf}\left(\sqrt{-\kappa}\right)\right)\right)}{2\kappa^2}$$

$$\begin{aligned} \mathbb{E}(D^2) = &-\frac{c_p(\kappa)\sqrt{\pi}(-4\kappa^2 erf(\sqrt{-\kappa}) - 4\kappa erf(\sqrt{-\kappa}) - 3erf(\sqrt{-\kappa}))}{4\kappa^2\sqrt{-\kappa}} \\ &- \frac{c_p(\kappa)\sqrt{-\kappa}(4\kappa+6)e^{\kappa}}{4\kappa^2\sqrt{-\kappa}}, \end{aligned}$$

$$\begin{aligned} \mathbb{E}(D^3) = &-\frac{c_p(\kappa)\sqrt{\pi}\sqrt{-\kappa}(8\kappa^3 erf(\sqrt{-\kappa}) + 12\kappa^2 erf(\sqrt{-\kappa}) + 18\kappa erf(\sqrt{-\kappa}))}{8\kappa^4} \\ &- \frac{c_p(\kappa)\sqrt{\pi}\sqrt{-\kappa}(15erf(\sqrt{-\kappa}))}{8\kappa^4} - \frac{(c_p(\kappa)(8\kappa^3+16\kappa^2+30\kappa)e^{\kappa})}{8\kappa^4}, \end{aligned}$$

e

$$\mathbb{E}(D^4) = -\frac{c_p(\kappa)\sqrt{\pi}(-16\kappa^4 erf(\sqrt{-\kappa}) - 32\kappa^3 erf(\sqrt{-\kappa}) - 72\kappa^2 erf(\sqrt{-\kappa}))}{16\kappa^4\sqrt{-\kappa}}$$
$$- \frac{c_q(\kappa)\sqrt{\pi}(-120\kappa erf(\sqrt{-\kappa}) - 105 erf(\sqrt{-\kappa}))}{16\kappa^4\sqrt{-\kappa}} - \frac{c_p(\kappa)\sqrt{-\kappa}(16\kappa^3 + 40\kappa^2)e^\kappa}{16\kappa^4\sqrt{-\kappa}}$$
$$- \frac{c_p(\kappa)\sqrt{-\kappa}(+100\kappa + 200)e^\kappa}{16\kappa^4\sqrt{-\kappa}}.$$

APÊNDICE F – MATRIZ DE INFORMAÇÃO DE FISHER DO MODELO TD_2

Seja $D_1, \ldots, D_n$ amostra aleatória de n-pontos provenientes de $D \sim \mathrm{TD}_2(\kappa)$ com fdp definida por (4.3) e função de log-verossimilhança dada a seguir

$$\ell(\kappa) = -n\log 2 - n\log M(1/2, 3/2, \kappa) + \sum_{i=1}^{n} \kappa(1-d_i) - \frac{1}{2}\sum_{i=1}^{n}\log(1-d_i).$$

Considere as seguintes parcelas de $\ell(\kappa)$ derivadas com respeito a κ,

$$\frac{\partial - n\log 2}{\partial \kappa} = 0 \quad (\text{não depende de} \quad \kappa),$$

$$\frac{\partial - n\log M(1/2, 3/2, \kappa)}{\partial \kappa} = -n\frac{\partial \log M(1/2, 3/2, \kappa)}{\partial \kappa},$$

$$\frac{\partial \sum_{i=1}^{n} \kappa(1-d_i)}{\partial \kappa} = \sum_{i=1}^{n}(1-d_i) \tag{F.1}$$

e

$$\frac{\partial - \frac{1}{2}\sum_{i=1}^{n}\log(1-d_i)}{\partial \kappa} = 0 \quad (\text{não depende de} \quad \kappa),$$

em que a derivada de $\partial\left[M(1/2, 3/2, \kappa)\right]/\partial\kappa$, é dada por

$$\frac{\partial M(1/2, 3/2, \kappa)}{\partial \kappa} = \frac{M(3/2, 5/2, \kappa)}{3}.$$

Logo,

$$-n\frac{\partial \log M(1/2, 3/2, \kappa)}{\partial \kappa} = \frac{-nM(3/2, 5/2, \kappa)}{3M(1/2, 3/2, \kappa)} \tag{F.2}$$

e a função escore é expressa pela soma das parcelas (F.1) e (F.2), dada a seguir

$$U(\kappa) = -\frac{nM(3/2,\ 5/2,\ \kappa)}{3M(1/2,\ 3/2,\ \kappa)} + \sum_{i=1}^{n}(1-d_i),$$

Agora, considere as seguintes parcelas de $U(\kappa)$ derivadas com respeito a κ,

$$\frac{\partial}{\partial \kappa}\left[-\frac{nM(3/2,\ 5/2,\ \kappa)}{3M(1/2,\ 3/2,\ \kappa)}\right] = -n\left(\frac{-M(3/2,\ 5/2,\ \kappa)}{9M(1/2,\ 3/2,\ \kappa)} + \frac{M(5/2,\ 7/2,\ \kappa)^2}{5M(1/2,\ 3/2,\ \kappa)^2}\right) \tag{F.3}$$

e

$$\frac{\partial \sum_{i=1}^{n}(1-d_i)}{\partial \kappa} = 0 \quad (\text{não depende de} \quad \kappa), \tag{F.4}$$

em que

$$\frac{\partial M(3/2,\ 5/2,\ \kappa)}{\partial \kappa} = \frac{3M(5/2, 7/2, \kappa)}{5}. \tag{F.5}$$

Logo, a matriz de informação de Fisher, definida pela expressão $K(\kappa) = -\mathbb{E}\left(\frac{\partial^2 ln}{\partial \kappa^2}\right)$, é dada por

$$K(\kappa) = -\left[-n\left(\frac{-M(3/2,\ 5/2,\ \kappa)^2}{9M(1/2,\ 3/2,\ \kappa)^2} + \frac{M(5/2,\ 7/2,\ \kappa)}{5M(1/2,\ 3/2,\ \kappa)}\right)\right],$$

correspondendo a soma das parcelas (F.3) e (F.4).