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

Identificação de Pontos Influentes em uma Amostra da Distribuição de Watson

CRISTIANY DE MOURA BARROS

RECIFE
2014

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

Identificação de Pontos Influentes em uma Amostra da Distribuição de Watson

CRISTIANY DE MOURA BARROS

Orientador: Prof. Dr. GETÚLIO JOSÉ AMORIM DO AMARAL

Área de Concentração: ESTATÍSTICA APLICADA

Dissertação submetida como requerimento parcial para obtenção do grau de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, 26 de fevereiro de 2014

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Barros, Cristiany de Moura
Identificação de pontos influentes em uma amostra da distribuição de Watson / Cristiany de Moura Barros. - Recife: O Autor, 2014.
114 f.: il., fig., tab.

Orientador: Getúlio José Amorim do Amaral.
Dissertação (mestrado) - Universidade Federal de Pernambuco. CCEN, Estatística, 2014.

Inclui referências e apêndices.

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

310 CDD (23. ed.) MEI2014 – 031

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

26 de fevereiro de 2014

Nós recomendamos que a Dissertação de Mestrado de autoria de

Cristiany de Moura Barros

Intitulada:

"Identificação de Pontos Influentes em uma Amostra da Distribuição de WATSON"

Seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

_____
Coordenador da Pós-Graduação em Estatística

Banca Examinadora:

Getúlio José Amorim do Amaral	Orientador/UFPE
Audrey Helen Mariz de Aquino Cysneiros	UFPE
Renata Maria Cardoso Rodrigues de Souza	UFPE-CIn

Este documento será anexado à versão final da tese.

A Deus, por cada dom que recebemos dele... Isso é uma dádiva...

Aos meus pais, Francisco de Holanda e Maria de Moura, pelo amor incondicional.

A Allan, por todo o seu amor.

Aprendi com o Mestre dos Mestres que...

Aprendi com o Mestre dos Mestres que a arte de pensar é o tesouro dos sábios.

Aprendi um pouco mais a pensar antes de reagir, a expor - e não impor - minhas idéias e a entender que cada pessoa é um ser único no palco da existência.

Aprendi com o Mestre da Sensibilidade a navegar nas águas da emoção, a não ter medo da dor, a procurar um profundo significado para a vida e a perceber que nas coisas mais simples e anônimas se escondem os segredos da felicidade.

Aprendi com o Mestre da Vida que viver é uma experiência única, belíssima, mas brevíssima.

E, por saber que a vida passa tão rápido, sinto necessidade de compreender minhas limitações e aproveitar cada lágrima, sorriso, sucesso e fracasso como uma oportunidade preciosa de crescer.

Aprendi com o Mestre do Amor que a vida sem amor é um livro sem letras, uma primavera sem flores, uma pintura sem cores. Aprendi que o amor acalma a emoção, tranquiliza o pensamento, incendeia a motivação, rompe obstáculos intransponíveis e faz da vida uma agradável aventura, sem tédio, angústia ou solidão. Por tudo isso Jesus Cristo se tornou, para mim, um Mestre "InesquecíveL".

Augusto Cury.

Agradecimentos

Aos meus pais, Francisco de Holanda e Maria de Moura, pelo amor incondicional, dedicação, confiança, paciência, e por terem me ensinado os valores e princípios que até hoje norteiam a formação do meu caráter como pessoa.

Aos meus irmãos, pelo carinho, confiança e amizade.

Ao meu esposo Allankardec Silva Sabino, pelo amor, pela compreensão e paciência.

Ao professor Getúlio José Amorim do Amaral, pela orientação.

Aos meus colegas de mestrado Cleiton, Priscila, Raabe, Danielle, Jéssica, Claúdio, Cleber, Gianinni, Pedro, Renilma, e Raquel pela amizade, carinho e momentos de alegria compartilhados.

À Valéria Bittencourt, pelo carinho, paciência e amizade com que sempre tratou a mim e aos demais alunos do mestrado.

À banca examinadora, pelas valiosas críticas e sugestões.

À CAPES, pelo apoio financeiro.

Resumo

A análise estatística na esfera unitária é mais complexa do que se possa imaginar: a concepção elegante dos modelos probabilísticos é simples, porém usá-los na prática, muitas vezes se torna mais difícil. Esta dificuldade normalmente decorre da normalização complicada das constantes associadas com distribuições direcionais. No entanto, devido à respectiva capacidade poderosa de modelagem, distribuições esféricas continuam encontrando inúmeras aplicações. A distribuição direcional fundamental é a distribuição Von-Mises-Fisher, cujo os modelos para dados concentrados em torno de uma média. Mas para os dados que tem uma estrutura adicional, essa distribuição pode não ser adequada: em particular para os dados axialmente simétricos é mais conveniente abordarmos a distribuição de Watson (1965), que é o foco desta dissertação.

Na distribuição de Watson, são utilizados métodos tais como: ponto de corte para distância proposto por Cook (1977), teste de outlier para discordância proposto por Fisher *et al.* (1985), quantil de uma qui-quadrado proposto por Cook (1977) e distância geodésica. As contribuições dessa dissertação são: a derivação da distância de Cook, o uso da distância geodésica para detecção de outliers e um método de calculo do ponto de corte.

Palavras-chave: Distância de Cook. Distância Geodésica. Distribuição de Watson. Estimação de Máxima Verossimilhança. Estudos de Simulação.

Abstract

Statistical analysis on the unit sphere is more complex than you might imagine: the sleek design of the probabilistic models is simple, but use them in practice, it often becomes more difficult. This difficulty usually arises from the complicated normalizing constants associated with directional distributions. However, due to its powerful modeling capabilities, spherical distributions keep finding numerous applications . The fundamental directional distribution is the Von-Mises-Fisher data for which the models concentrated around an average distribution. But for the data that has an additional structure, this distribution may not be appropriate: in particular for axially symmetric data is more convenient approach the distribution Watson of (1965), which is the focus of this dissertation.

In Watson distribution, methods are used such as: cutoff for distance proposed by Cook (1977), outlier test for disagreement proposed by Fisher *et al.* (1985), quantile a chi-square proposed by Cook (1977) and geodesic distance. The contributions of this dissertation are: the derivation of the distance from Cook, the use of geodesic distance for detecting outliers and a method of calculating the cutoff.

Keywords: Cook's Distance. Geodesic Distance. Distribution Watson. Maximum Likelihood Estimation. Simulation Studies.

Sumário

CAPÍTULO 1

Introdução

1.1 Introdução

A análise estatística na esfera unitária é mais complexa do que se possa imaginar: a concepção elegante dos modelos probabilísticos é simples, porém usá-los na prática, muitas vezes se torna mais difícil. Esta dificuldade normalmente decorre da normalização complicada das constantes associadas com distribuições direcionais. No entanto, devido à respectiva capacidade poderosa de modelagem, distribuições hiperesferas continuam encontrando inúmeras aplicações,ver, por exemplo, Mardia e Jupp (2000). A distribuição direcional mais conhecida é a distribuição Von-Mises-Fisher, quais os modelos (VMF) cujo os dados são concentrados em torno de uma média (direção).

Mas para os dados que tem uma estrutura adicional é preciso definir que estrutura adicional é esta e essa distribuição pode não ser adequada: em particular para os dados axialmente simétricos é mais conveniente abordarmos a distribuição de Watson (1965), que é o foco desta dissertação. Três razões principais motivam o nosso estudo da distribuição multivariada de Watson, são elas: é fundamental para as estatísticas de direção; não tem recebido muita atenção para a análise de dados modernos que envolvem dados

de grande dimensão, um procedimento para análise da expressão genética Dhillon *et al.* (2003).

Uma razão pode ser que os domínios tradicionais da estatísticas direcionais são eixos tridimensionais e bidimensionais, por exemplo, círculos ou esferas. A distribuição de Watson é formada por quatro parâmetros: μ, ν, τ e k, onde k é conhecido como constante de normalização e sua densidade será definida em (2.1).

A distância definida por Cook (1977) mede quanto a deleção de uma observação altera as estimativas dos parâmetros de um modelo. Com base nesta distância , pode-se decidir se esta observação é influente ou não. De posse dos dados gerados de uma distribuição de Watson, utilizamos quatro critérios para detectar observações influentes, são eles: ponto de corte de distância proposto por Cook (1977), teste de outlier para discordância proposto por Fisher *et al.* (1985), quantil de uma qui-quadrado proposto por Cook (1977) e a distância geodésica.

Um conjunto básico de resumo de estatísticas em análise exploratória de dados consiste na mediana da amostra e nos extremos (também conhecidos como "dobradiças"), que são os quantis amostrais aproximados. Usando quantis não muito extremos, como a mediana, temos relativa insensibilidade de resumos e cálculos posteriores às observações incomuns. Assim, na tentativa de detecção de outliers, é vantajoso basear-se na regra de detecção de informações de amostra que é improvável que seja prejudicada por observações extremas. Tais abordagens tendem a minimizar essas dificuldades. As regras resistentes que estudam baseiam-se nos termos da amostra. Para estabelecer a notação, denotamos a amostra por $X_1, \ldots, X_n$ e a ordem correspondente as estatísticas por $X_1, \ldots, X_n$. Por definição , os termos são $FL = X(f)$ e $FU = X(n + l - f)$ com $f = \frac{1}{2}\lceil\frac{(n+3)}{2}\rceil$, onde $\lceil . \rceil$ indica o maior inteiro menor que o argumento real.

Conhecida como regra resistente à outliers, f localiza cada termo a partir do fim da amostra ordenada. O termo mais baixo, FL, é a mediana da metade inferior (incluindo a mediana geral quando n é ímpar) da amostra ordenada, e o termo superior, FU, é a mediana da metade superior (similarmente definido) da amostra ordenada. Usamos a notação estatística de ordem $X_{(f)}$ com os dois inteiros e não inteiros de f. A diferença da propagação entre esses termos é dada por

$$F = FU - FL,$$

é aproximadamente o intervalo interquartil. Ele fornece uma medida resistente e conveniente de propagação. A principal regra para identificar potenciais de outliers configura cercas internas $1.5 \times F$, além dos termos:

$$IFL = FL - 1.5 \times F \quad e \quad IFU = FL + 1.5 \times F.$$

Além de identificar observações que exigem uma atenção especial, a motivação inicial para esta técnica incluiu a detecção de comportamento pesado de cauda não homogênea (contra caudas bem pesadas). Segundo Tukey (1970), uma regra secundária configura cercas exteriores $3 \times F$ para além dos termos, que são mais seriamente extremas. Além disso, para a regra atual podemos usar os "valores secundários": $FL - 1.0 \times F$ ou $FU + 1.0 \times F$, $FL - 1.5 \times F$ ou $FU + 1.5 \times F$, $FL - 3.0 \times F$ ou $FU + 3.0 \times F$, $FL - 4.0 \times F$ ou $FU + 4.0 \times F$ e $FL - 6.0 \times F$ ou $FU + 6.0 \times F$, onde os valores: -1.0, -1.5, -3.0, -4.0, -6.0, 1.0, 1.5, 3.0, 4.0 e 6.0, são conhecidos como pontos de corte. Uma pesquisa sobre os métodos de teste para outliers produziu uma ampla literatura, discutida em livros por Barnett e Lewis (1978, 1984) e Hawkins (1980) e nos artigos por Barnett (1983) e Beckman e Cook (1983).

Num plano de duas dimensões, a geodésica é a menor distância que une dois pontos tal que, para pequenas variações da forma da curva, o seu comprimento é estacionário.

A representação da geodésica em um plano representa a projeção de um circulo máximo sobre uma esfera. Assim, tanto na superfície de uma esfera ou deformada num plano, a reta é uma curva, já que a menor distância possível entre dois pontos somente poderá ser curvada, pois uma reta necessariamente precisaria, permanecer sempre num plano, para ser a menor distância entre pontos. Do ponto de vista prático, na maioria dos casos, a geodésica é a curva de menor comprimento que une dois pontos.

Em uma "geometria plana"(espaço euclidiano), essa curva é um segmento de reta, mas em "geometrias curvas"(geometria riemaniana), muito utilizadas por exemplo na Teoria da Relatividade Geral, a curva de menor distância entre dois pontos pode não ser uma reta. Para entender isso, peguemos como exemplo a curvatura do globo terrestre e seus continentes. Se traçarmos uma linha ligando duas capitais de continentes distintos, perceberemos que a linha não é reta, mas sim um arco do círculo máximo, entretanto, se a distancia entre as duas cidades for pequena a linha que cobre o segmento do arco de circulo máximo será realmente uma reta .

Na relatividade geral, geodésicas descrevem o movimento de partículas pontuais sob a influência da gravidade. Em particular, o caminho tomado por uma pedra em queda, uma órbita de satélite , ou na forma de uma órbita planetária são todos geodésicas em curva-tempo-espaço.

1.2 Objetivo da Dissertação

O objetivo da dissertação é propor quatro medidas para detectar pontos influentes e outliers. Essas medidas são: distância de Cook, teste outlier para discordância, quantil de uma qui-quadrado e distância geodésica comparadas numericamente em estudos de experimentos de simulação e na análise de dados reais.

1.3 Organização da Dissertação

Além do capítulo de introdução, esta dissertação é composta por mais quatro capítulos. No Capítulo 2 nos concentramos em análise de dados na esfera e na distribuição de Watson. Apresentamos uma revisão geral sobre os dados na esfera, sobre a distribuição de Watson discutindo suas principais características e estimação dos seus parâmetros.

No Capítulo 3 abordamos como calcular a distância de Cook de uma determinada amostra da distribuição de Watson, e a partir dessa distância é verificado se a observação é influente ou não, usando quatro critérios: ponto de corte de distância proposto por Cook (1977), teste de outlier para discordância proposto por Fisher *et al.* (1985), quantil de uma qui-quadrado proposto por Cook (1977) e a distância geodésica.

No Capítulo 4, apresentamos uma análise dos dados simulados, calculando a taxa de detecção e a taxa de erro para diferentes tamanhos de amostras e diferentes valores de k, conhecido como constante de normalização, para podermos detectar se aquela determinada observação é influente ou não. Além disso, usamos também os critérios mencionados anteriormente para constatarmos se essas observações de fato seriam influentes ou não.

Com relação aos dados reais, tomamos três amostras de tamanhos, $n = 50$, $n = 72$ e $n = 75$, a primeira amostra refere-se as posições dos pólos determinada a partir do estudo de solos paleomagnéticos de Nova Caledônia, onde a latitude representa a variável θ e a longitude a variável ϕ. Notamos que todas as latitudes nesse conjunto de dados são negativas ver, (Apêndice (A.3)). A Figura 1.1 apresenta o gráfico das posições dos pólos a partir do estudo de solos paleomagnéticos de Nova Caledônia.

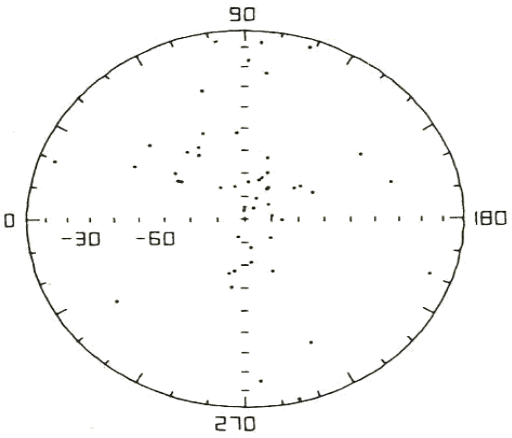

Figura 1.1: Posições dos pólos da partir do estudo de solos paleomagnéticos de Nova Caledônia.

A segunda e terceira amostra, ambas referem-se orientações de superfícies de clivagem de plano axial de dobras em ordovician turbiditosas, compostas pelas variáveis: mergulho e direção de mergulho, denotadas latitude (θ) e longitude (ϕ), respectivamente. A Figura 1.2 mostra o gráfico da projeção de 72 pólos para superfícies de clivagem do plano axial. A Figura 1.3 apresenta o gráfico da projeção de 75 pólos para superfícies de clivagem do plano axial.

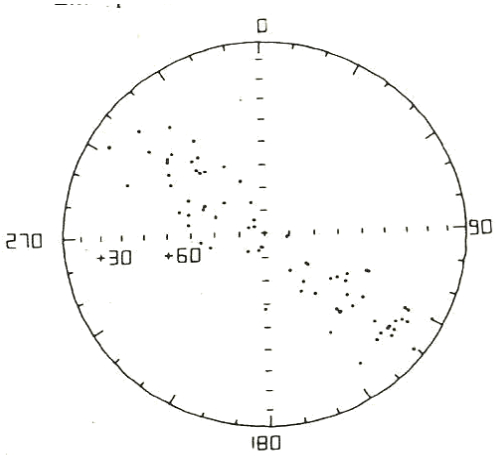

Figura 1.2: Projeção de 72 pólos para superfícies de clivagem do plano axial.

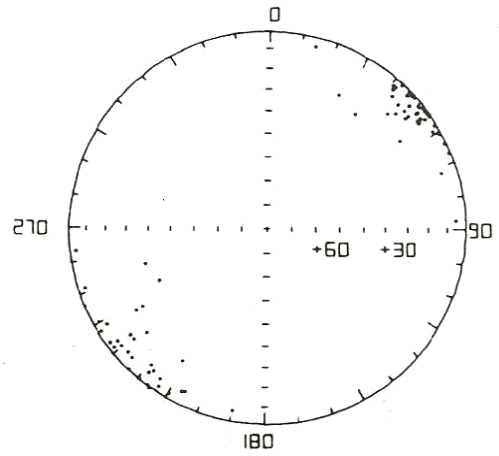

Figura 1.3: Projeção de 75 pólos para superfícies de clivagem do plano axial.

Os três conjuntos de dados foram obtidos a partir de Fisher *et al.* (1987). Ainda neste capítulo, apresentamos um estudo comparativo, algumas contribuições desenvolvidas entre os critérios para a taxa de detecção e a taxa de erros dos pontos influentes. Por fim, no Capítulo 5, apresentamos as conclusões extraídas do presente trabalho e apontamos direções para trabalhos futuros.

1.4 Suporte computacional

O sistema tipográfico usado neste trabalho foi, integralmente, o LaTeX[1], o qual consiste em um conjunto de macros para o processador de textos TeX, desenvolvido por Donald Knuth, em 1986. Sua principal utilização é na produção de textos matemáticos e científicos, devido à alta qualidade tipográfica. Adotou-se o software MikTeX: uma implementação do LaTeX para a utilização em ambiente Windows.

[1]Para mais informações e detalhes sobre o sistema de tipografia LaTeX ver De Castro (2003) ou acesse o site http://www.tex.ac.uk/CTAN/latex

Quanto a parte computacional, foi usado o software R, que é uma linguagem e um ambiente para computação estatística e para preparação de gráficos de alta qualidade. É um projeto GNU semelhante a S-PLUS e, ainda que haja diferenças significativas entre eles. R oferece uma grande variedade de técnicas estatísticas e gráficas. Por ser um software livre é permitida contribuições de novas funcionalidades por meio da criação de pacotes. Sua documentação e tutoriais estão disponíveis no sítio http://www.r-project.org.

CAPÍTULO 2

Dados na Esfera e Distribuição de Watson

2.1 Sistemas de Coordenadas Esféricas

2.1.1 Introdução

Existem vários problemas estatísticos que surgem na análise de dados quando as observações são dados direcionais. O assunto tem recebido cada vez mais atenção, mas o campo é tão antigo como a própria estatística matemática. Com efeito, a teoria dos erros foi desenvolvido por Gauss principalmente para analisar certas medições direcionais em astronomia. O avanço no assunto é marcado por um artigo pioneiro de Fisher (1953), o seu trabalho foi motivado por um problema paleomagnético representada pelo geofísico, J. Hospers.

Desde então, graças principalmente a GS Watson e MA Stephens, o desenvolvimento do tema tem sido rápido. É interessante notar que Karl Pearson estava envolvido com um problema de pássaros migratórios que leva ao passeio aleatório isotrópico em um círculo tão cedo quanto 1905. Além disso, Von Mises introduziu uma importante distribuição circular em 1918 para estudar o desvio de pesos atômicos dos valores integrais.

2.1.2 Direção média e centro de comprimento de massa resultante

Considere a coleção de pontos $P_1, \ldots, P_n$ na superfície da esfera unitária de centro O, com P_i correspondendo ao vetor unitário de coordenadas polares (θ_i, ϕ_i) onde $x_i = sen\theta_i \cos\phi_i$, $y_i = sen\theta_i sen\phi_i$ e $z_i = \cos\theta_i$, $i = 1, \ldots, n$. A direção dos cossenos pode ser escrito como um vetor

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} \quad \text{ou seu transposto} \quad \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}^\top = \begin{pmatrix} x_i & y_i & z_i \end{pmatrix}.$$

O ângulo ψ entre os vetores unitários $\overrightarrow{OP_1}$ e $\overrightarrow{OP_2}$ ver, Figura 2.1, é então dada (em radianos) por

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

com $0° \leq \psi \leq 180°$.

O vetor resultante dos vetores unitários n, é um vetor de direção $(\hat{\theta}, \hat{\phi})$ e com comprimento R (ver, Figura 2.2) é dado por

$$S_x = \sum_{i=1}^{n} x_i, \quad S_y = \sum_{i=1}^{n} y_i, \quad S_z = \sum_{i=1}^{n} z_i.$$

Depois,

$$R^2 = S_x^2 + S_y^2 + S_z^2$$

de modo que $(\hat{\theta}, \hat{\phi})$ tem cossenos de direção $(\hat{x}, \hat{y}, \hat{z}) = (S_x/R, S_y/R, S_z/R)$.

Assim,

$$sen\hat{\theta}\cos\hat{\phi} = \hat{x},$$
$$sen\hat{\theta}sen\hat{\phi} = \hat{y},$$
$$\cos\hat{\theta} = \hat{z}.$$

de modo que

$$\hat{\theta} = \arccos(\hat{z}), \hat{\phi} = \arctan(\hat{y}/\hat{x}), \quad com$$
$$0^\circ \leq \hat{\theta} \leq 180^\circ \quad e \quad 0^\circ \leq \hat{\phi} \leq 360^\circ.$$

A Figura 2.1 apresenta o gráfico dos vetores unitários no espaço tridimensional, com ψ o ângulo entre os vetores unitários $\overrightarrow{OP_1}$ e $\overrightarrow{OP_2}$, em que P_1, P_2, P_3, P_4 e P_5, são pontos que representam vetores unitários no espaço tridimensional.

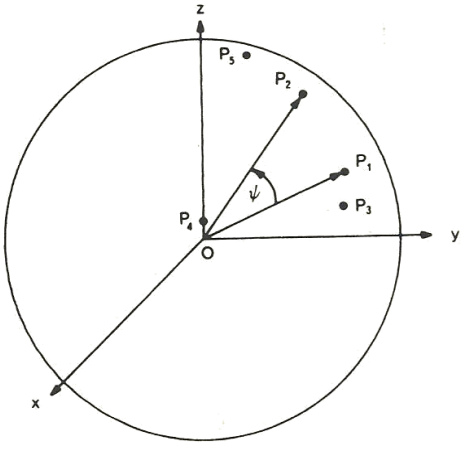

Figura 2.1: Vetores unitários no espaço tridimensional.

Vale salientar que

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

É necessário a programação para os casos excepcionais em que $\hat{x} = 0$ e $\hat{y} = 0$.

Aqui, $(\hat{\theta}, \hat{\phi})$ é chamado a média de direção dos n vetores unitários, R é o comprimento resultante e $\overline{R} = R/n$ é a média do comprimento resultante. Note que este centro de massa de $P_1, \ldots, P_n$ (consideradas massas unitárias) tem coordenadas $(\sum_{i=1}^n x_i/n, \sum_{i=1}^n y_i/n, \sum_{i=1}^n z_i/n)$ na direção $(\hat{\theta}, \hat{\phi})$ numa distância R a partir da origem.

Obviamente, o centro de massa de pontos distintos na superfície da esfera unitária deve ser interior à esfera. Assim, $0 \leq \overline{R} \leq 1$, $R = 1$ correspondendo para todos sendo coincidente. Note que o centro de massa é esse quando estamos usando direções. No caso de eixos da (distribuição de Watson), o centro de massa é o autovetor associado ao maior autovalor da matriz de covariância.

Note que $\overline{R}$ não é sempre um indicador de colocação confiável de dispersão simétrica dos pontos, por exemplo, em dois grupos iguais em lados opostos de um eixo podem resultar em $\overline{R} = 0$. A Figura 2.2 apresenta o gráfico do vetor unitário resultante e o vetor média de comprimento resultante.

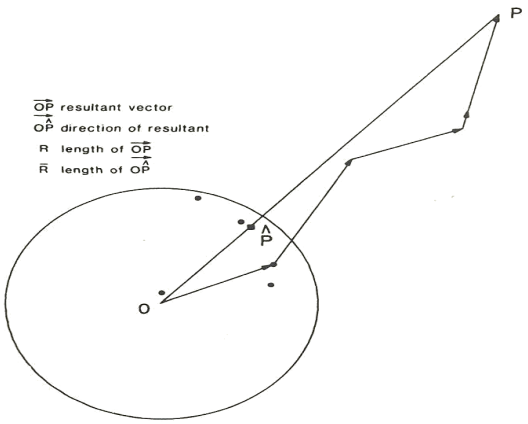

Figura 2.2: Vetor resultante e vetor média de comprimento resultante.

onde $\overrightarrow{OP}$ é o vetor resultante de $\overrightarrow{OP_1}, \overrightarrow{OP_2}, \overrightarrow{OP_3}, \overrightarrow{OP_4}$ e $\overrightarrow{OP_5}$ com comprimento R.

2.1.3 Rotação de vetores unitários e eixos

Seja (θ, ϕ) as coordenadas polares de um vetor unitário medido em relação a um pólo na direção $(0, 0)$. Será frequentemente necessário encontrar as coordenadas (θ', ϕ') em relação a alguma outra direção (θ_0, ϕ_0) como pólo. Isto pode ser conseguido através de rotação. Seja ψ_0 um ângulo arbitrário. A fórmula geral da matriz de rotação de x, y e z e eixos (θ_0, ϕ_0) com nova direção (θ', ϕ') é dada por

$$\begin{pmatrix} \cos\theta_0 \cos\phi_0 \cos\psi_0 - senn\theta_0 senn\psi_0 & \cos\theta_0 senn\phi_0 \cos\psi_0 + \cos\phi_0 sen\psi_0 & -sen\theta_0 \cos\psi_0 \\ -\cos\theta_0 \cos\phi_0 sen\psi_0 - sen\phi_0 sen\psi_0 & -\cos\theta_0 sen\phi_0 sen\psi_0 + \cos\phi_0 \cos\psi_0 & sen\theta_0 sen\psi_0 \\ sen\theta_0 \cos\phi_0 & sen\theta_0 sen\phi_0 & \cos\theta_0 \end{pmatrix}$$

$= \mathbf{A}(\theta_0, \phi_0, \psi_0)$.

Em seguida, a direção dos cossenos é dada por:

$$\begin{pmatrix} sen\theta' \cos\phi' \\ sen\theta' sen\phi' \\ \cos\theta' \end{pmatrix} = \mathbf{A}(\theta_0, \phi_0, \psi_0) \begin{pmatrix} sen\theta \cos\phi \\ sen\theta sen\phi \\ \cos\theta \end{pmatrix}.$$

O ângulo ψ_0 é uma rotação em torno do eixo polar através de (θ_0, ϕ_0). Se ele é arbitrariamente definido como zero, podemos simplificar a fórmula da matriz de rotação por:

$$\mathbf{A}(\theta_0, \phi_0, 0) = \begin{pmatrix} \cos\theta_0 \cos\phi_0 & \cos\theta_0 sen\phi_0 & -sen\theta_0 \\ -sen\phi_0 & \cos\phi_0 & 0 \\ sen\theta_0 \cos\phi_0 & sen\theta_0 sen\phi_0 & \cos\theta_0 \end{pmatrix}.$$

Note que $\theta^{'}$ é o ãngulo entre (θ, ϕ) e (θ_0, ϕ_0).

Para os dados axiais, supomos que o eixo (x, y, z) é medido em relação ao eixo polar $(0, 0, 1)$ e nós procuramos encontrar as coordenadas $(x^{'}, y^{'}, z^{'})$ em relação a (x_0, y_0, z_0). Por conveniência, primeiro deve-se determinar uma extremidade do eixo de coordenadas polares (x_0, y_0, z_0), para obter (θ_0, ϕ_0). Em seguida através dos vetores obtém-se $\mathbf{A}(\theta_0, \phi_0, \psi_0)$ ou $\mathbf{A}(\theta_0, \phi_0, 0)$, e portanto

$$\begin{pmatrix} x^{'} \\ y^{'} \\ z^{'} \end{pmatrix} = \mathbf{A}(\theta_0, \phi_0, \psi_0) \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

$z^{'}$ é o cosseno do menor dos ângulos entre (x, y, z) e (x_0, y_0, z_0).

2.2 Distribuição de Watson

2.2.1 Introdução

Seja $\mathbb{S}^{p-1} = \{\mathbf{x} | \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_2 = 1\}$ ser $p-1$ uma hiperesfera de unidade-dimensional centrada na origem de norma unitária. Nós nos concentramos em vetores axialmente simétricos, ou seja, $\pm\mathbf{x} \in \mathbb{S}^{p-1}$ são equivalentes, o que também é denotada por $\mathbf{x} \in \mathbb{P}^{p-1}$, em que $\mathbb{P}^{p-1}$ é uma hiperplana projetiva da dimensão de $p-1$. A escolha natural para a modelagem desses dados é através da distribuição multivariada de Watson (ver, Mardia e Jupp, (2000)).

Considere $(\pm x_1, \cdots, \pm x_n)$ uma amostra aleatória de uma distribuição de Watson, com função de densidade de probabilidade dada por

$$f(\pm x; \mu, k) = M\left(\frac{1}{2}, \frac{p}{2}, k\right)^{-1} \exp\{k(x^\top \mu)^2\}, \tag{2.1}$$

em que M é a função confluente hipergeométrica de Kummer definida por Erdélyi et al. (1953) e Andrews et al. (1999), cuja a fórmula é dada por

$$M(a, b, k) = \sum_0^\infty \frac{\Gamma(a+n)\Gamma(b)k^n}{\Gamma(a)\Gamma(b+n)n!} \tag{2.2}$$

onde a, b, $k \in \mathbb{R}$ e μ um vetor. Observe-se que para $k > 0$, a densidade é mais concentrada com o aumento dos parâmetros μ e k, enquanto que para $k < 0$, a densidade concentra-se em torno do grande círculo ortogonal a μ.

2.2.2 Estimação de Máxima Verossimilhança

Considere $(\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathbb{P}^{p-1}$, uma amostra aleatória com função de densidade de uma distribuição de Watson com média μ e parâmetro de concentração k. O logaritmo de verossimilhança dessa distribuição é dada por

$$\ell(\mu; k, x_1, \ldots, x_n) = n\left(k\mu^\top \mathbf{S}\mu - lnM\left(\frac{1}{2}, \frac{p}{2}, k\right) + \gamma\right), \tag{2.3}$$

em que $\mathbf{S} = \sum_{i=1}^n x_i x_i^\top$ é a matriz de dispersão de amostra e γ é um termo constante, que pode ser ignorado. Maximizando a equação (2.3), obtemos as estimativas dos parâmetros para o vetor média μ dado por

$$\hat{\mu} = \mathbf{s}_1 \quad se \quad \hat{k} > 0, \quad \hat{\mu} = \mathbf{s}_p \quad se \quad \hat{k} < 0, \tag{2.4}$$

onde $(\mathbf{s}_1, \ldots, \mathbf{s}_p)$ são os autovetores normalizados $(\in \mathbb{P}^{p-1})$ da matriz de dispersão $\mathbf{S}$ correspondentes aos autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. A estimativa do parâmetro de concentração k é obtida resolvendo (para ser mais preciso, precisamos $\lambda_1 > \lambda_2$ para assegurar um único EMV, para $k > 0$, e $\lambda_{p-1} > \lambda_p$, para $k < 0$):

$$g\left(\frac{1}{2}, \frac{p}{2}; \hat{k}\right) := \frac{M'\left(\frac{1}{2}, \frac{p}{2}, \hat{k}\right)}{M\left(\frac{1}{2}, \frac{p}{2}, \hat{k}\right)} = \hat{\mu}^T \mathbf{S}\hat{\mu} := r \quad (0 \leq r \leq 1), \tag{2.5}$$

em que M' denota a derivada de M com relação a $\hat{k}$, (veja Abramowitz *et al.* (1965)) dada por

$$M'(a,b,k) = \frac{a}{b} M(a+1, b+1, k).$$

Observe que as equações (2.4) e (2.5) são acopladas, por isso precisamos decidir resolver: $g\left(\frac{1}{2}, \frac{p}{2}; \hat{k}\right) = \lambda_1$ ou resolver $g\left(\frac{1}{2}, \frac{p}{2}; \hat{k}\right) = \lambda_p$. Uma alternativa fácil seria resolver as duas equações e selecionar a solução que gera o maior logaritmo da função de verossimilhança. Porém resolver estas equações é muito mais difícil, pois, nem sempre essas equações possuem forma fechada. Pode-se resolver (2.5), utilizando um método numérico de busca de raízes (por exemplo, de Newton-Raphson). Mas, a situação não é tão simples. Vamos, portanto, considerar uma equação um pouco mais geral dada por

$$g(a,c;k) := \frac{M'(a,c;k)}{M(a,c;k)} = r, \qquad c > a > 0, \quad 0 \leq r \leq 1. \tag{2.6}$$

Para gerar os dados aleatórios da distribuição de Watson, para $k \neq 0$, usamos o algoritmo proposto por Kim e Carl (1993), com os seguintes passos:

1. Seja $\rho = \frac{4k}{2k+3+\sqrt{(2k+3)^2-16k}}$ e $r = \left(\frac{3\rho}{2k}\right)^3 \exp\left(-3 + \frac{2k}{\rho}\right)$.
2. Gere U_0 e U_1 uniformes independentes de $U(0,1)$.
3. Seja $S = \frac{U_0^2}{1-\rho(1-U_0^2)}$.
4. Seja $W = kS$ e $V = \frac{rU_1^2}{(1-\rho S)^3}$.
5. Se $V > \exp(2W)$, volta para o passo 2.
6. Seja $\theta = cos^{-1}(\sqrt{S})$.
7. Gere U_2 de uma distribuição uniforme $U(0,1)$.
8. Se $U_2 < 0.5$, considere $\theta = \pi - \theta$ e $\phi = 4\pi U_2$, caso contrário, $\phi = 2\pi(2U_2 - 1)$.

2.2.3 Resolvendo k

Existem diferentes soluções para (2.6). A primeira é baseada no método de Newton-Raphson para raízes. O segundo método é baseado no cálculo de uma solução de forma fechada aproximado para $(2, 6)$, exigindo assim apenas alguns pontos infuentes. Esse método é conhecido como método de precisão assintótica das aproximações. Vamos agora olhar mais precisamente a forma como esse método se comporta ao limitar os valores de r, que no nosso caso como o valor de k tomado inicialmente nos processos de simulação foi positivo, segundo Suvrit *et al.* (2013), tomamos o autovetor associado ao maior autovalor r. Podemos calcular r de três formas: $r \to 0$, $r \to \frac{a}{c}$ ou $r \to 1$. Para a escolha do r, e posteriormente para a escolha do $k(r)$, utilizamos o seguinte teorema:

Teorema 2.2.1 (Critério para escolha de k)**.** *Seja $c > a > 0$, $r \in (0, 1)$; seja $k(r)$ a solução para $g(a, c; k) = r$, em seguida temos que,*

$$k(r) = \frac{-a}{c} + (c - a - 1) + \frac{(c - a - 1)(1 + a)}{a} r + O(r^2), \quad \text{r}\to 0, \tag{2.7}$$

$$k(r) = \left(r - \frac{a}{c}\right)\left\{\frac{c^2(1 + c)}{a(c - a)} + \frac{c^3(1 + c)^2(2a - c)}{a^2(c - a)^2(c + 2)}\left(r - \frac{a}{c}\right) + O\left(\left(r - \frac{a}{c}\right)\right)\right\}, \quad \text{r}\to \text{a/c}, \tag{2.8}$$

$$k(r) = \frac{c - a}{1 - r} + 1 - a + \frac{(a - 1)(a - c - 1)}{c - a}(1 - r) + O((1 - r)^2), \quad \text{r}\to 1. \tag{2.9}$$

Nos nossos experimentos de simulação escolhemos autovalores próximos de 1, denotados por $r1$ (ver Apêndice A). O que justifica o fato de usarmos a equação (2.9), e não as equações (2.8) e (2.9). Consideramos o caso em que $a = \frac{1}{2}$ e dimensionalidade $c = \frac{p}{2}$, onde $p = 3$. Este caso ilustra como as nossas aproximações não-lineares se comportam para a equação geral (2.6).

Vamos agora derivar os limites dos dois lados o que levará a uma aproximação de forma fechada para a solução de (2.5). Esta aproximação, é mais rápida para calcular, uma vez que está na forma fechada. Antes de avançar para os detalhes, vamos olhar para um pouco de história. Para dados dimensionais ou tridimensionais, ou sob hipóteses restritivas sobre k ou r, algumas aproximações tinham sido previamente obtida por Mardia e Jupp (2000). Devido às suas suposições restritivas, estas aproximações têm aplicabilidade limitada, onde estes pressupostos são frequentemente violados (Banerjee *et al.* (2005)). Recentemente Bijral *et al.* (2007), seguido da técnica de Banerjee *et al.* (2005) utilizam a aproximação ad-hoc (na verdade, em particular para o caso de $a = 1/2$), dada por

$$k(r) = \frac{cr - a}{r(1-r)} + \frac{r}{2c(1-r)}, \tag{2.10}$$

que observado ser muito preciso. No entanto, (2.10), precisa de justificativa teórica, embora novamente apenas aproximação ad-hoc. A seguir, apresentamos novas aproximações para k que são, teoricamente, bem motivados e também numericamente mais precisas. A chave para a obtenção destes aproximações são um conjunto de limites que localizam k, utilizando o seguinte teorema:

Teorema 2.2.2 (Outras fórmulas de estimação de k na forma fechada.)**.** *Considere a solução de $g(a, c; k) = r$ ser denotado por $k(r)$, dado pelas equações*

$$k(r) = \frac{cr - a}{r(1-r)} \left(1 + \frac{1-r}{c-a}\right), \tag{2.11}$$

$$k(r) = \frac{cr - a}{2r(1-r)} \left(1 + \sqrt{1 + \frac{4(c+1)r(1-r)}{a(c-a)}}\right), \tag{2.12}$$

$$k(r) = \frac{cr - a}{r(1-r)} \left(1 + \frac{r}{a}\right). \tag{2.13}$$

CAPÍTULO 3

Distância de Cook

3.1 Introdução

A distância de Cook é uma medida da influência de uma observação e é proporcional à soma dos quadrados das diferenças entre as previsões feitas com todas as observações da análise e previsões feitas deixando a observação em questão. Se as previsões são as mesmas com ou sem a observação em questão, então a observação não tem nenhuma influência sobre o modelo em pauta. Se as previsões diferem grandemente quando a observação não é incluída na análise, então a observação é influente.

A regra comum é que uma observação com um alto valor da distância de Cook tenha muita influência. A influência de uma observação é dada por dois fatores: o quanto o valor da observação sobre a variável de previsão difere da média da variável de previsão e a diferença entre a pontuação prevista para a observação e sua pontuação real.

Algumas vezes, por observação dos valores que constituem a amostra ou pela análise de alguns gráficos, é fácil identificar as observações que se afastam da maioria. Em outros casos, é necessária a aplicação de técnicas mais sofisticadas. Em ambos os casos, esta

análise prévia tem de ser seguida de testes apropriados para confirmar as suspeitas de existência de observações outliers. Quando se passa para um conjunto de dados em que foram observadas, não uma mas p variáveis, há um acréscimo significativo de dificuldades. Se o conjunto de p variáveis é não correlacionado, é possível trabalhar com cada variável individualmente. Porém, se existe uma correlação entre as variáveis, é preciso usar métodos de análise multivariada. No mundo real existem muitas situações onde é possível obter p variáveis para o objeto, no entanto, a luta contra essas dificuldades é justificada pela necessidade de obter conhecimentos, uma vez que em termos práticos é muito usual e necessário trabalhar com dados multidimensionais em vez de dados com uma dimensão apenas.

Em dados multidimensionais, uma observação é considerada outlier se está "muito" distante das restantes no espaço p-dimensional definido pelas variáveis. Um grande problema na identificação de outliers multivariados surge pelo fato de que uma observação pode não ser "anormal"em nenhuma das variáveis originais estudadas isoladamente e ser na análise multivariada, ou pode ainda ser outlier por não seguir a estrutura de correlação dos restantes dados. É impossível detectar este tipo de outlier observando cada uma das variáveis originais isoladamente.

A distância de Cook generalizada é dada por

$$DC(i) = (\hat{\theta}_i - \hat{\theta})^\top J(\hat{\theta})(\hat{\theta}_i - \hat{\theta}), \tag{3.1}$$

onde $\hat{\theta}_i$ é o vetor retirando a i-ésima observação da matriz $S1$ (ver Apêndice (A.1)), $\hat{\theta}$ é o vetor da matriz S com todas as observações (ver Apêndice (A.1)) e $J(\hat{\theta})$ é matriz de Informação de Fisher Observada, (ver Apêndice (A.1)), dada por

$$J(\hat{\theta}) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_1} & \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_2} & \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_3} & \frac{\partial^2 \ell}{\partial \mu_1 \partial k} \\ \frac{\partial^2 \ell}{\partial \mu_2 \partial \mu_1} & \frac{\partial^2 \ell}{\partial \mu_2 \partial \mu_2} & \frac{\partial^2 \ell}{\partial \mu_2 \partial \mu_3} & \frac{\partial^2 \ell}{\partial \mu_2 \partial k} \\ \frac{\partial^2 \ell}{\partial \mu_3 \partial \mu_1} & \frac{\partial^2 \ell}{\partial \mu_3 \partial \mu_2} & \frac{\partial^2 \ell}{\partial \mu_3 \partial \mu_3} & \frac{\partial^2 \ell}{\partial \mu_3 \partial k} \\ \frac{\partial^2 \ell}{\partial k \partial \mu_1} & \frac{\partial^2 \ell}{\partial k \partial \mu_2} & \frac{\partial^2 \ell}{\partial k \partial \mu_3} & \frac{\partial^2 \ell}{\partial k \partial k} \end{pmatrix} \tag{3.2}$$

Para calcularmos a matriz de Informação de Fisher $J(\hat{\theta})$ em (3.2), é necessário encontrarmos as segundas derivadas do logaritmo da verosssimilhança da distribuição de Watson em (2.3), com relação a cada um de seus quatro parâmeros: μ_1, μ_2, μ_3 e k. Note que $\mu^\top \mathbf{S} \mu$, para $p = 3$, é dado por

$$\begin{aligned} k\mu^\top \mathbf{S}\mu &= k \left(\begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \end{pmatrix} \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \mathbf{S}_{13} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \mathbf{S}_{23} \\ \mathbf{S}_{31} & \mathbf{S}_{32} & \mathbf{S}_{33} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \right) \\ &= k \left(\begin{pmatrix} \mathbf{S}_{11}\mu_1 + \mathbf{S}_{21}\mu_2 + \mathbf{S}_{31}\mu_3 \\ \mathbf{S}_{12}\mu_1 + \mathbf{S}_{22}\mu_2 + \mathbf{S}_{32}\mu_3 \\ \mathbf{S}_{13}\mu_1 + \mathbf{S}_{23}\mu_2 + \mathbf{S}_{33}\mu_3 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \right) \\ &= k[\mathbf{S}_{11}{\mu_1}^2 + \mathbf{S}_{21}\mu_1\mu_2 + \mathbf{S}_{31}\mu_1\mu_3 + \mathbf{S}_{12}\mu_1\mu_2 + \mathbf{S}_{22}{\mu_2}^2 + \mathbf{S}_{32}\mu_2\mu_3 + \mathbf{S}_{13}\mu_1\mu_3 \\ &+ \mathbf{S}_{23}\mu_2\mu_3 + \mathbf{S}_{33}{\mu_3}^2], \end{aligned} \tag{3.3}$$

Os termos da matriz de Informação de Fisher observada são obtidos da seguinte maneira:

- $\frac{\partial \ell}{\partial \mu_1} = k\left(2\mathbf{S}_{11}\mu_1 + \mathbf{S}_{21}\mu_2 + \mathbf{S}_{31}\mu_3 + \mathbf{S}_{12}\mu_2 + \mathbf{S}_{13}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_1} = 2k\mathbf{S}_{11}$

- $\frac{\partial \ell}{\partial \mu_2} = k\left(\mathbf{S}_{21}\mu_1 + \mathbf{S}_{12}\mu_1 + 2\mathbf{S}_{22}\mu_2 + \mathbf{S}_{32}\mu_3 + \mathbf{S}_{23}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_2} = k(\mathbf{S}_{21} + \mathbf{S}_{12})$

- $\frac{\partial \ell}{\partial \mu_3} = k\left(\mathbf{S}_{31}\mu_1 + \mathbf{S}_{32}\mu_2 + \mathbf{S}_{13}\mu_1 + \mathbf{S}_{23}\mu_2 + 2\mathbf{S}_{33}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_3} = k(\mathbf{S}_{31} + \mathbf{S}_{13})$

- $\frac{\partial \ell}{\partial \mu_2} = k\left(\mathbf{S}_{21}\mu_1 + \mathbf{S}_{12}\mu_1 + 2\mathbf{S}_{22}\mu_2 + \mathbf{S}_{32}\mu_3 + \mathbf{S}_{23}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial \mu_2 \partial \mu_2} = 2k\mathbf{S}_{22}$

- $\frac{\partial \ell}{\partial \mu_3} = k\left(\mathbf{S}_{31}\mu_1 + \mathbf{S}_{32}\mu_2 + \mathbf{S}_{13}\mu_1 + \mathbf{S}_{23}\mu_2 + 2\mathbf{S}_{33}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial \mu_3 \partial \mu_3} = 2k\mathbf{S}_{33}$

- $\frac{\partial \ell}{\partial \mu_3} = k\left(\mathbf{S}_{31}\mu_1 + \mathbf{S}_{32}\mu_2 + \mathbf{S}_{13}\mu_1 + \mathbf{S}_{23}\mu_2 + 2\mathbf{S}_{33}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial \mu_2 \partial \mu_3} = k(\mathbf{S}_{32} + \mathbf{S}_{23})$

- $\frac{\partial \ell}{\partial \mu_1} = k\left(2\mathbf{S}_{11}\mu_1 + \mathbf{S}_{21}\mu_2 + \mathbf{S}_{31}\mu_3 + \mathbf{S}_{12}\mu_2 + \mathbf{S}_{13}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial k \partial \mu_1} = 2\mathbf{S}_{11}\mu_1 + \mathbf{S}_{21}\mu_2 + \mathbf{S}_{31}\mu_3 + \mathbf{S}_{12}\mu_2 + \mathbf{S}_{13}\mu_3$

- $\frac{\partial \ell}{\partial \mu_2} = k\left(\mathbf{S}_{21}\mu_1 + \mathbf{S}_{12}\mu_1 + 2\mathbf{S}_{22}\mu_2 + \mathbf{S}_{32}\mu_3 + \mathbf{S}_{23}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial k \partial \mu_2} = \mathbf{S}_{21}\mu_1 + \mathbf{S}_{12}\mu_1 + 2\mathbf{S}_{22}\mu_2 + \mathbf{S}_{32}\mu_3 + \mathbf{S}_{23}\mu_3$

- $\frac{\partial \ell}{\partial \mu_3} = k\left(\mathbf{S}_{31}\mu_1 + \mathbf{S}_{32}\mu_2 + \mathbf{S}_{13}\mu_1 + \mathbf{S}_{23}\mu_2 + 2\mathbf{S}_{33}\mu_3\right) \Rightarrow \frac{\partial^2 \ell}{\partial k \partial \mu_3} = \mathbf{S}_{31}\mu_1 + \mathbf{S}_{32}\mu_2 + \mathbf{S}_{13}\mu_1 + \mathbf{S}_{23}\mu_2 + 2\mathbf{S}_{33}\mu_3$

- $\frac{\partial \ell}{\partial k} = \mu^\top \mathbf{S}\mu - \frac{1}{M(a,b,k)}\frac{\partial M(a,b,k)}{\partial k} = \mu^\top \mathbf{S}\mu - \frac{1}{M(a,b,k)}\left[\frac{a}{b}M(a,b,k)\right] \Rightarrow$

$$
\begin{aligned}
&\frac{\partial^2 \ell}{\partial k \partial k} = -\frac{\partial}{\partial k}\left[\frac{\frac{a}{b}M(a+1,b+1,k)}{M(a,b,k)}\right] = -\frac{a}{b}\frac{\partial}{\partial k}\left[\frac{M(a+1,b+1,k)}{M(a,b,k)}\right] = -\frac{a}{b}\frac{\partial}{\partial k}\left[\frac{M(a',b',k)}{M(a,b,k)}\right] \\
&= -\frac{a}{b}\left[\frac{M(a,b,k)\frac{\partial}{\partial k}M(a',b',k)] - M(a',b',k)\frac{\partial}{\partial k}M(a,b,k)}{(M(a,b,k))^2}\right] \\
&= -\frac{a}{b}\left[\frac{M(a,b,k)\frac{a'}{b'}M(a'+1,b'+1,k)] - M(a',b',k)\frac{a}{b}M(a+1,b+1,k)}{(M(a,b,k))^2}\right] \\
&= -\frac{a}{b}\left[\frac{M(a,b,k)\frac{a+1}{b+1}M(a+2,b+2,k)] - M(a+1,b+1,k)\frac{a}{b}M(a+1,b+1,k)}{(M(a,b,k))^2}\right] \\
&= -\frac{a}{b}\left[\frac{a+1b+1M(a,b,k)M(a+2,b+2,k) - \frac{a}{b}(M(a+1,b+1,k))^2}{(M(a,b,k))^2}\right] \\
&= \frac{a}{b}\left[\frac{\frac{a}{b}(M(a+1,b+1,k))^2 - \frac{a+1}{b+1}M(a,b,k)M(a+2,b+2,k)}{(M(a,b,k))^2}\right] \\
&= \frac{a}{b}\left[\frac{\frac{a(b+1)(M(a+1,b+1,k))^2 - (a+1)bM(a,b,k)M(a+2,b+2,k)}{b(b+1)}}{(M(a,b,k))^2}\right] \\
&= \frac{a}{b}\left[\frac{a(b+1)(M(a+1,b+1,k))^2 - (a+1)bM(a,b,k)M(a+2,b+2,k)}{b(b+1)}\frac{1}{(M(a,b,k))^2}\right] \\
&= \frac{a^2(b+1)(M(a+1,b+1,k))^2 - a(a+1)bM(a,b,k)M(a+2,b+2,k)}{b^2(b+1)(M(a,b,k))^2}.
\end{aligned}
$$

Como satisfazem as condições de regularidade, temos que:

- $\frac{\partial^2 \ell}{\partial \mu_2 \partial \mu_1} = \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_2} = k(\mathbf{S}_{21} + \mathbf{S}_{12})$
- $\frac{\partial^2 \ell}{\partial \mu_3 \partial \mu_1} = \frac{\partial^2 \ell}{\partial \mu_1 \partial \mu_3} = k(\mathbf{S}_{31} + \mathbf{S}_{13})$
- $\frac{\partial^2 \ell}{\partial \mu_3 \partial \mu_2} = \frac{\partial^2 \ell}{\partial \mu_2 \partial \mu_3} = k(\mathbf{S}_{32} + \mathbf{S}_{23})$
- $\frac{\partial^2 \ell}{\partial \mu_1 \partial k} = \frac{\partial^2 \ell}{\partial k \partial \mu_1} = 2\mathbf{S}_{11}\mu_1 + \mathbf{S}_{21}\mu_2 + \mathbf{S}_{31}\mu_3 + \mathbf{S}_{12}\mu_2 + \mathbf{S}_{13}\mu_3$
- $\frac{\partial^2 \ell}{\partial \mu_2 \partial k} = \frac{\partial^2 \ell}{\partial k \partial \mu_2} = \mathbf{S}_{21}\mu_1 + \mathbf{S}_{12}\mu_1 + 2\mathbf{S}_{22}\mu_2 + \mathbf{S}_{32}\mu_3 + \mathbf{S}_{23}\mu_3$
- $\frac{\partial^2 \ell}{\partial \mu_3 \partial k} = \frac{\partial^2 \ell}{\partial k \partial \mu_3} = \mathbf{S}_{31}\mu_1 + \mathbf{S}_{32}\mu_2 + \mathbf{S}_{13}\mu_1 + \mathbf{S}_{23}\mu_2 + 2\mathbf{S}_{33}\mu_3$

3.2 Critérios

A preocupação com observações outliers é antiga e data das primeiras tentativas de analisar um conjunto de dados. Inicialmente, pensava-se que a melhor forma de lidar com esse tipo de observação seria através da sua eliminação da análise. Atualmente este procedimento é ainda muitas vezes utilizado, existindo, no entanto, outras formas de lidar com tal tipo de fenômeno. Conscientes deste fato e sabendo que tais observações poderiam conter informações importantes em relação aos dados, sendo por vezes as mais importantes, é nosso propósito apresentar os principais aspectos da discussão deste assunto.

Grande parte dos autores que estudam este fenômeno ver, por exemplo, Barnett e Lewis (1994), como sendo uma das primeiras e mais importantes referências a observações

outliers. Esses comentários indicam que a prática de rejeitar tal tipo de observação era comum naquela altura (século XVIII). A discussão sobre as observações outliers centrava-se na justificativa da rejeição daqueles valores. As opiniões não eram unânimes: uns defendiam a rejeição das observações "inconsistentes com as restantes", enquanto outros afirmavam que as observações nunca deviam ser rejeitadas simplesmente por parecerem inconsistentes com os restantes dados e que todas as observações deviam contribuir com igual peso para o resultado final. Em qualquer dos casos está presente uma certa subjectividade na tomada de decisão sobre o que fazer com os outliers.

Antes de decidir o que deverá ser feito às observações outliers é conveniente ter conhecimento das causas que levam ao seu aparecimento. Em muitos casos as razões da sua existência determinam as formas como devem ser tratadas. Assim, as principais causas que levam ao aparecimento de outliers são: erros de medição, erros de execução e variabilidade inerente dos elementos da população.

O estudo de outliers, independentemente da(s) sua(s) causa(s), pode ser realizado em várias fases. A fase inicial é a da identificação das observações que são potencialmente aberrantes. A identificação de outliers consiste na detecção, com métodos subjetivos, das observações surpreendentes. A identificação é feita, geralmente, por análise gráfica ou, no caso de o número de dados ser pequeno, por observação direta dos mesmos. São assim identificadas as observações que têm fortes possibilidades de virem a ser designadas por outliers.

Na segunda fase, tem-se como objetivo a eliminação da subjetividade inerente à fase anterior. Pretende-se saber se as observações identificadas como outliers potenciais o são, efetivamente. São efetuados testes à ou às observações "preocupantes". Devem ser escolhidos os testes mais adequados para a situação em estudo. Estes dependem do tipo de outlier em causa, do seu número, da sua origem, do conhecimento da distribuição subja-

cente à população de origem das observações, etc. As observações suspeitas são testadas quanto à sua discordância. Se for não for rejeitada a hipótese de algumas observações serem outliers, elas podem ser designadas como discordantes. Uma observação diz-se discordante se puder considerar-se inconsistente com os restantes valores depois da aplicação de um critério estatístico objetivo. Muitas vezes o termo discordante é usado como sinônimo de outlier.

Na terceira e última fase é necessário decidir o que fazer com as observações discordantes. A maneira mais simples de lidar com essas observações é eliminá-las. Como já foi dito, esta abordagem, apesar de muito utilizada, não é aconselhada. Caso contrário, as observações consideradas como outliers devem ser tratadas cuidadosamente pois contêm informação relevante sobre características subjacentes aos dados e poderão ser decisivas no conhecimento da população à qual pertence a amostra em estudo.

A forma alternativa à eliminação é a acomodação dos outliers (accommodation of outliers). Tenta-se "conviver"com os outliers. A acomodação passa pela inclusão na análise de todas as observações, incluindo os possíveis outliers. Independentemente de existirem ou não outliers, opta-se por construir proteção contra eles. Para tal, são efetuadas modificações no modelo básico e/ou nos métodos de análise.

Uma linha de pesquisa importante é a proposta de medidas de influência com base no tamanho de amostras de um determinado conjunto de dados. Assim a novidade deste trabalho é a apresentação de alguns métodos de detecção de outliers e pontos influentes em amostras desta distribuição. Não há uma definição amplamente aceita de um outlier (ver Barnett e Lewis, 1994). No entanto, pode-se definir um outlier como uma observação que é incompatível com os outros pontos de amostra. Por outro lado , os pontos influentes podem ser entendidos como a informação que representa uma mudança forte em inferência de processos quando removidos do conjunto de dados em estudo . Assim, todos os pontos

influentes são discrepantes, mas nem todos os outliers são pontos influentes (Barnett e Lewis, 1994). Alguns critérios para detecção de outliers e observações influentes dos dados de uma distribuição de Watson dados têm sido propostos. São eles: ponto de corte para distância proposto por Cook (1977), teste de outlier para discordância proposto por Fisher *et al.* (1985), quantil de uma qui-quadrado proposto por Cook (1977) e a distância geodésica.

3.2.1 Ponto de Corte para Distância

Considere $(x_1, x_2, \ldots, x_n)$, uma amostra aleatória n-dimensional de x. Neste trabalho, utilizou-se o seguinte algoritmo com base nas propostas de Hoaglin *et al.* (1986), como uma forma alternativa de definir os limiares de medidas influentes analíticas:

1. Da amostra aleatória, obter $GD_i, \ldots, GD_n$, para $i = 1, \ldots, n$.

2. Seja $d(1), \ldots, d(n)$, o conjunto de distâncias que representam o conjuntos de dados. Encontre as estatísticas de ordem sobre este conjunto de medidas, $d(1) \leq \ldots \leq d(n)$.

3. Calcular $FL = d(f)$ e $FU = d(i + 1 - f)$ em que $f = \frac{1}{2}\lceil\frac{(i+3)}{2}\rceil$, em que $\lceil . \rceil$ denota o maior inteiro menor do que um argumento real. Neste caso, a observação será considerada como outlier ou ponto influente se:

$$d[i] > FU + C \times (FU - FL), \tag{3.4}$$

onde o valor C, é conhecido como ponte de corte. Não existe um critério rígido para o valor de C. Nessa dissertação nós investigamos os valores $C = 1.5$, 3, 4 e 6.

3.2.2 Teste Outlier para Discordância

Seja

$$T_{i-1} = T_i - \begin{pmatrix} x_i^2 & x_i y_i & x_i z_i \\ x_i y_i & y_i^2 & y_i z_i \\ x_i z_i & y_i z_i & z_i^2 \end{pmatrix},$$

onde $T_i - 1$ é a matriz de orientação com a i-ésima observação omitida e T_i é a matriz de orientação com todas as observações.

A matriz T_i é dada por

$$T_i = \begin{pmatrix} \sum x_i^2 & \sum x_i y_i & \sum x_i z_i \\ \sum x_i y_i & \sum y_i^2 & \sum y_i z_i \\ \sum x_i z_i & \sum y_i z_i & \sum z_i^2 \end{pmatrix}.$$

O eixo em torno do qual o momento é menor é chamado de eixo principal para completar o conjunto de três coordenadas ortogonais que existem dois eixos menores. Estes eixos correspondem aos autovetores (também conhecido com vetor latente ou vetor característico) de T_i, que denotamos $\hat{\mu_1}$, $\hat{\mu_2}$ e $\hat{\mu_3}$. Associados com os autovetores, temos os autovalores $\hat{\tau_1}$, $\hat{\tau_2}$ e $\hat{\tau_3}$ respectivamente, que satisfazem

$$\hat{\tau_1} \geq 0, \quad \hat{\tau_2} \geq 0, \quad \hat{\tau_3} \geq 0, \quad \hat{\tau_1} + \hat{\tau_2} + \hat{\tau_3} = n.$$

Vamos supor que $0 \leq \hat{\tau_1} \leq \hat{\tau_2} \leq \hat{\tau_3}$. Se $\hat{\tau_1} = \hat{\tau_2} = \hat{\tau_3} = \frac{n}{3}$, então não há eixo maior do que qualquer outro momento de inércia; caso contrário $\hat{\mu_3}$ é o eixo principal. Os autovetores e autovalores fornecem a decomposição espectral da matriz de orientação dada por

$$T = \hat{\tau_1}\hat{\mu_1}\hat{\mu_1}' + \hat{\tau_2}\hat{\mu_2}\hat{\mu_2}' + \hat{\tau_3}\hat{\mu_3}\hat{\mu_3}'.$$

Programas de computador usados para calcular os autovetores e autovalores de uma matriz positiva definida simétrica estão disponíveis em pacotes de software matemáticos padrão,

por exemplo, o R tem disponível a função *eigen* que fornece os autovalores e autovetores. Alternativamente, um algoritmo adequado está listado no programa descrito por Diggle e Fisher (1985).

Vamos usar muitas vezes os autovalores normalizados

$$\bar{\tau_1} = \frac{\hat{\tau_1}}{n}, \quad \bar{\tau_2} = \frac{\hat{\tau_2}}{n} \quad \bar{\tau_3} = \frac{\hat{\tau_3}}{n},$$

$$\bar{\tau_1} + \bar{\tau_2} + \bar{\tau_3} = 1.$$

Seja $\hat{\tau_{3,i}}$ e $\hat{\tau_{3,i-1}}$ os autovalores das matrizes T_i e T_{i-1}, respectivamente. A estatística de teste dada por

$$H_i = \frac{(i-2)(1+\hat{\tau}_{3,i-1} - \hat{\tau}_{3,i})}{i-1-\hat{\tau}_{3,i-1}} \tag{3.5}$$

O outlier (x_i, y_i, z_i) é considerado discordante se o valor da estatística H em (3.5), for muito grande. Nessa dissertação nós propomos o critério abaixo para detectar os valores grandes de $H[i]$. Neste caso, a observação será considerada como outlier ou ponto influente se

$$H[i] > FU + C \times (FU - FL), \tag{3.6}$$

em quw o valor C, é conhecido como ponte de corte, assumindo os valores: 1.5, 3, 4 e 6.

3.2.3 Quantil de uma Qui-quadrado

Cook (1977) apresenta uma distância para medir a influência no contexto da análise de regressão. Este critério consiste em detectar se uma observação é influente ou não, usando o valor da distância de Cook e o quantil de uma qui-quadrado de grau 3, com nível de

confiança de 95%. Considere $(\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathbb{P}^{p-1}$, uma amostra aleatória com função de densidade de uma distribuição de Watson com média μ e parâmetro de concentração k dada pela equação (2.1) e logaritmo da verossimilhança dada por (2.3). A matriz de informação observada e esperada são dadas, respectivamente, por

$$\begin{aligned} J(\hat{\theta}) &= \frac{\partial^2 \ell(\hat{\theta})}{\partial \hat{\theta} \hat{\theta}^\top}, \\ K(\theta) &= -\mathbb{E}\left(\frac{\partial^2 \ell(\theta)}{\partial \theta \theta^\top}\right). \end{aligned}$$

Através da distância de Cook generalizada dada pela equação (3.1), podemos detectar se a obervação possui um ou mais pontos influentes (outliers). Note que $J(\hat{\theta})$ é uma matriz positiva definida e assintoticamente equivalente a matriz de informação de θ, ver Cook (1977) e Cook e Weisberg (1982). De acordo com Cook (1977), a região de confiança para a distância de Cook é dada por

$$\{\theta : (\hat{\theta}_i - \hat{\theta})^\top J(\hat{\theta})(\hat{\theta}_i - \hat{\theta}) \leq \chi^2_{p,\alpha}\} \tag{3.7}$$

em que todos os pontos fora da região de confiança dada por (3.7), são considerados influentes e devem ser cuidadosamente examinados. Portanto, uma observação é considerada um outlier baseada nesse critério se

$$dcook[i] > qchisq(0.95, 3), \tag{3.8}$$

sendo 0.95 o intervalo de confiança e o valor 3 é a dimensão dos vetores sob estudo.

3.2.4 Distância Geodésica

A representação da geodésica em uma esfera ou em um plano deformada, é uma curva, já que a menor distância possível entre dois pontos somente poderá ser uma curva, porque temos um espaço não-euclideano. Se fosse uma reta necessariamente precisaria de um

plano sem nenhuma deformação para ter a menor distância entre os pontos. Do ponto de vista prático, na maioria dos casos, a geodésica é a curva de menor comprimento que une dois pontos. Em uma "geometria plana"(espaço euclidiano), essa curva é um segmento de reta, mas em "geometrias curvas"(geometria riemaniana), muito utilizadas por exemplo na Teoria da Relatividade Geral, a curva de menor distância entre dois pontos pode não ser uma reta. A fórmula da distancia geodésica entre dois vetores é dada por

$$DG = \arccos(t(v_1) \times v_2),$$

onde DG significa distância geodésica e v_1 e v_2 são vetores pertencentes a esfera.

O critério de distância geodésica usado para calcular a taxa de detecção de outliers e a taxa de erro nos dados da distribuição de Watson é dado por

$$distancia[i] > FU + C \times (FU - FL), \tag{3.9}$$

em que o valor C, é conhecido como ponte de corte, assumindo os valores: 1.5, 3, 4 e 6.

No caso a função que calcula a distância geodésica (ver Apêndice A), considera-se como resposta a menor da distâncias entre os vetores u e v.

CAPÍTULO 4

Análise de Dados

4.1 Resultados Numéricos

Nesta seção, nosso objetivo é comparar, através de simulações, os desempenhos dos vários critérios usados para determinar a existência de algum outlier. Os critérios a serem avaliados são: ponto de corte para a distância ($D1$), teste outlier para discordância ($D2$), quantil de uma qui-quadrado ($D3$) e distância geodésica ($D4$). As estatísticas usadas para comparar estes critérios são dadas no Capítulo 3 em (3.4), (3.6), (3.8) e (3.9), respectivamente. O número de réplicas foi fixado em 1000 e foi considerado o nível nominal de $\alpha = 0,05\%$. As simulações foram realizadas usando a linguagem de programação R, cujo os tutoriais estão disponíveis no sítio http://www.r-project.org.

Para cada tamanho de amostra, calculamos as taxas de detecção e as taxas de erro. A taxa de detecção refere-se a taxa de sucesso de uma observação que é influente em um determinado conjunto de observações, enquanto a taxa de erro verifica se uma observação que não é influente, é influente. Para a taxa de detecção, fixamos um valor de k para gerarmos as $n-1$ observações e fixamos um outro valor de k, para gerarmos a n-ésima observação. Notamos ainda que essa observação gerada separadamente das demais, seria

um candidato a outlier. Porém para a taxa de erro, fixamos um único valor de k, para gerarmos todas as observações.

A Tabela 4.1 apresenta a taxa de detecção dos critérios para a situação em que $n = 20$, onde as 19 observações foram geradas com $k = 10$ e a vigésima observação gerada com $k = 0,01$. Notou-se que todos os critérios apresentam uma boa taxa de detecção para o ponto de corte, $C = 1,5$. É possível notar que para o ponto de corte $C = 3$, os critérios $D1$ e $D3$, continaram apresentando uma boa taxa de detecção, porém os critérios $D2$ e $D4$, diminuíram as taxas de detecção para esse esse ponto de corte. Notamos ainda que a medida que aumentamos o ponto de corte, a taxa de detecção tende a diminuir, o que é esperado. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de detecção dos critérios $D1$ e $D3$, continuaram altas, quando comparados aos critérios $D2$ e $D4$.

Para o tamanho de amostra $n = 50$, onde as 49 observações foram geradas com $k = 10$ e a vigésima observação gerada com $k = 0,01$. Notou-se que todos os critérios apresentaram uma boa taxa de detecção para o ponto de corte, $C = 1,5$, e suas taxas de detecção com esse mesmo ponto de corte foram maiores que as taxas de detecção para o tamanho de amostra $n = 20$. É possível notar que o mesmo ocorreu para o ponto de corte $C = 3$, onde os critérios $D1$ e $D3$, continuam apresentando uma boa taxa de detecção, porém os critérios $D2$ e $D4$, diminuíram as taxas de detecção para esse esse ponto de corte. O mesmo acontece para os pontos de cortes iguais a $C = 4$ e $C = 6$, quando comparados aos critérios $D2$ e $D4$. Porém as taxas de detecção de todos os critérios para esse tamanho de amostra foram maiores, quando comparados ao tamanho de amostra $n = 20$. Para o tamanho de amostra $n = 80$, onde as 79 observações foram geradas com $k = 10$ e a vigésima observação gerada com $k = 0,01$. Notou-se que todos os critérios apresentaram uma boa taxa de detecção para o ponto de corte, $C = 1,5$, e suas taxas de detecção com esse mesmo ponto de corte foram maiores que as taxas de detecção para o tamanho de amostra $n = 20$ e $n = 50$. Os critérios $D2$ e $D4$, diminuíram as taxas de detecção para

o ponto de corte $C = 3$. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de detecção dos critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, quando comparados aos critérios $D2$ e $D4$. Porém as taxas de detecção de todos os critérios para esse tamanho de amostra foram maiores, quando comparados ao tamanho de amostra $n = 20$ e $n = 50$.

Tabela 4.1: Taxa de detecção de outliers para diferentes tamanhos de amostras e valores de $k = 10$ e $k = 0,01$

C	n	Critérios			
		$D1$	$D2$	$D3$	$D4$
1,5	20	1	0,797	1	0,847
	50	1	0,809	1	0,857
	80	1	0,798	1	0,862
3	20	1	0,676	1	0,712
	50	1	0,713	1	0,727
	80	1	0,688	1	0,748
4	20	1	0,579	1	0,604
	50	1	0,615	1	0,615
	80	1	0,603	1	0,640
6	20	1	0,405	1	0,389
	50	1	0,418	1	0,399
	80	1	0,394	1	0,459

A Figura 4.1 apresenta o gráfico das distâncias de Cook, H e geodésica, denotadas por $D1$, $D2$ e $D4$, respectivamente com $n = 20$ observações para valores de $k = 10$ e $k = 0,01$. Claramente, podemos observar que para estes valores, a distância H tem melhor desempenho que as distâncias de Cook e geodésica com respeito à taxa de detecção de outliers. As Figuras 4.2 e 4.3 apresentam os gráficos das distâncias de Cook, H e geodésica para $n = 50$ e $n = 80$ observações respectivamente, para valores de $k = 10$ e $k = 0,01$. Em ambas figuras, podemos observar que o desempenho das três distâncias com respeito à taxa de detecção de outliers é similar a Figura 4.1. Notamos ainda que em todas as figuras, a distância H apresenta valores mais dispersos, enquanto que na distância de Cook os valores estão mais concentrados em torno de 0 e 1.

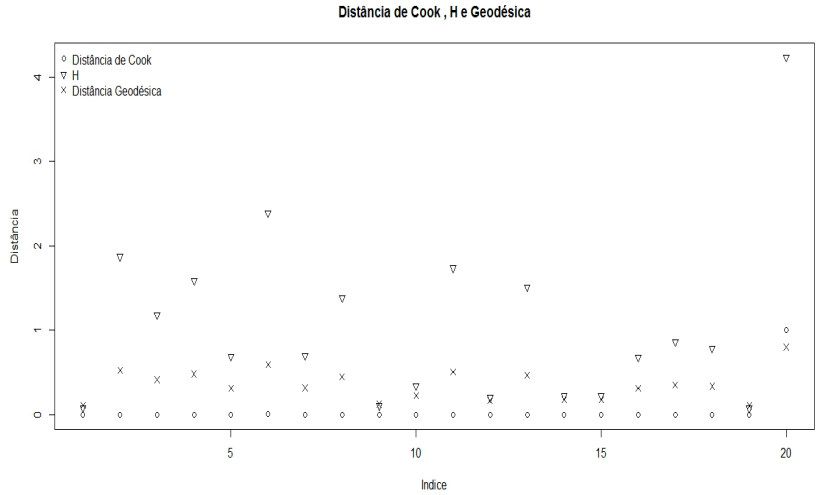

Figura 4.1: Distâncias de Cook, H e geodésica para $n = 20$ observações para valores de $k = 10$ e $k = 0,01$.

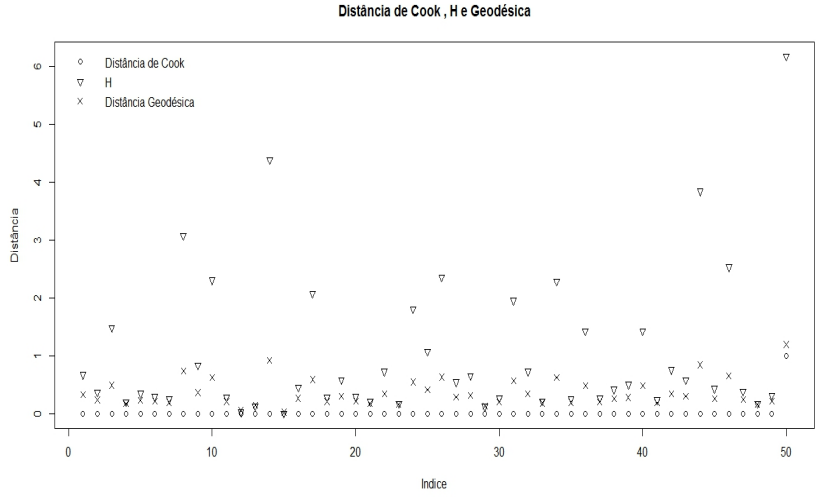

Figura 4.2: Distâncias de Cook, H e geodésica para $n = 50$ observações para valores de $k = 10$ e $k = 0,01$.

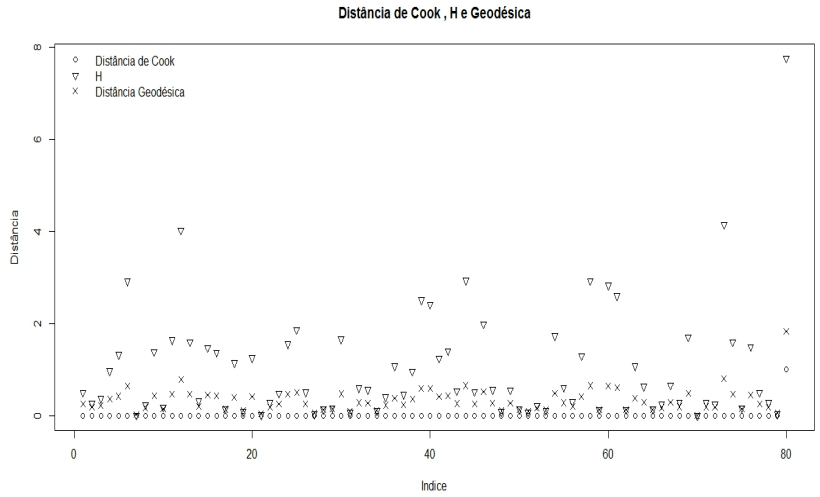

Figura 4.3: Distâncias de Cook, H e geodésica para $n = 80$ observações para valores de $k = 10$ e $k = 0,01$.

A Tabela 4.2 apresenta a taxa de detecção dos critérios para a situação em que $n = 20$, onde as 19 observações foram geradas com $k = 12$ e a vigésima observação gerada com $k = 0,05$. Notou-se que todos os critérios apresentaram uma boa taxa de detecção para o ponto de corte, $C = 1,5$. É possível notar que para o ponto de corte $C = 3$, os critérios $D1$ e $D3$, continaram apresentando uma boa taxa de detecção, porém os critérios $D2$ e $D4$, diminuíram as taxas de detecção para esse esse ponto de corte.

Notamos ainda que a medida que aumentamos o ponto de corte, a taxa de detecção tende a diminuir. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de detecção dos critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, quando comparados aos critérios $D2$ e $D4$.

Para o tamanho de amostra $n = 50$, onde as 49 observações foram geradas com $k = 12$ e a vigésima observação gerada com $k = 0,05$. Notou-se que todos os critérios apresentaram uma boa taxa de detecção para o ponto de corte, $C = 1,5$, e suas taxas de detecção com esse mesmo ponto de corte foram maiores que as taxas de detecção para o

tamanho de amostra $n = 20$. É possível notar que o mesmo ocorreu para o ponto de corte $C = 3$, onde os critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, porém os critérios $D2$ e $D4$, diminuíram as taxas de detecção para esse esse ponto de corte.

Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de detecção dos critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, quando comparados aos critérios $D2$ e $D4$. Porém as taxas de detecção de todos os critérios para esse tamanho de amostra foram maiores, quando comparados ao tamanho de amostra $n = 20$.

Para o tamanho de amostra $n = 80$, onde as 79 observações foram geradas com $k = 12$ e a vigésima observação gerada com $k = 0,05$. Notou-se que todos os critérios apresentaram uma boa taxa de detecção para o ponto de corte $C = 1,5$, e suas taxas de detecção com esse mesmo ponto de corte foram maiores que as taxas de detecção para o tamanho de amostra $n = 20$ e $n = 50$.

Tabela 4.2: Taxa de detecção de outliers para diferentes tamanhos de amostras e valores de $k = 12$ e $k = 0,05$

C	n	Critérios			
		$D1$	$D2$	$D3$	$D4$
1,5	20	1	0,827	1	0,844
	50	1	0,848	1	0,884
	80	1	0,863	1	0,884
3	20	1	0,736	1	0,717
	50	1	0,768	1	0,789
	80	1	0,770	1	0,795
4	20	1	0,682	1	0,617
	50	1	0,705	1	0,686
	80	1	0,703	1	0,709
6	20	1	0,548	1	0,440
	50	1	0,573	1	0,505
	80	1	0,571	1	0,532

A Figura 4.4 apresenta o gráfico das distâncias de Cook, H e geodésica para $n = 20$ observações para valores de $k = 12$ e $k = 0,05$. Podemos observar que para estes valores, a distância H tem melhor desempenho que as distâncias de Cook e geodésica com respeito à taxa de detecção de outliers.

Por outro lado, a distancia de Cook apresenta melhor desempenho quando comparada à distância geodésica. As Figuras 4.5 e 4.6 apresentam os gráficos das distâncias de Cook, H e geodésica para $n = 50$ e $n = 80$ observações respectivamente, para valores de $k = 12$ e $k = 0,05$. Em ambas figuras, podemos observar que o desempenho das três distâncias com respeito à taxa de detecção de outliers é similar a Figura 4.4.

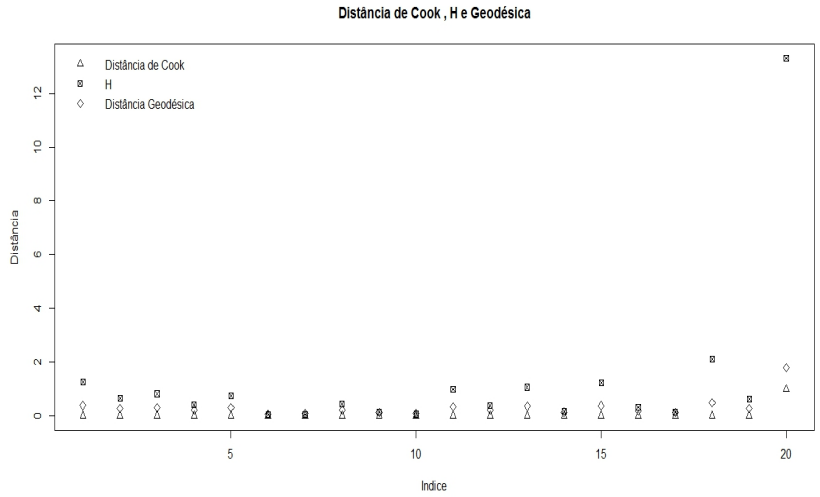

Figura 4.4: Distâncias de Cook, H e geodésica para $n = 20$ observações para valores de $k = 12$ e $k = 0,05$.

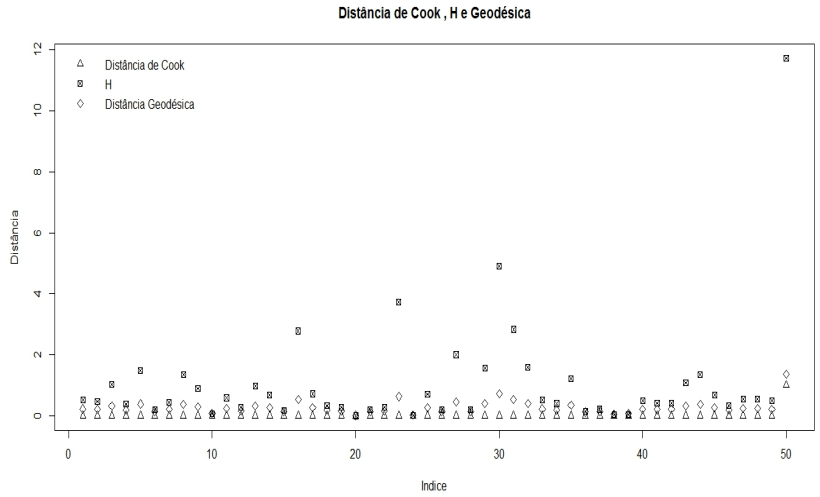

Figura 4.5: Distâncias de Cook, H e geodésica para $n = 50$ observações para valores de $k = 12$ e $k = 0,05$.

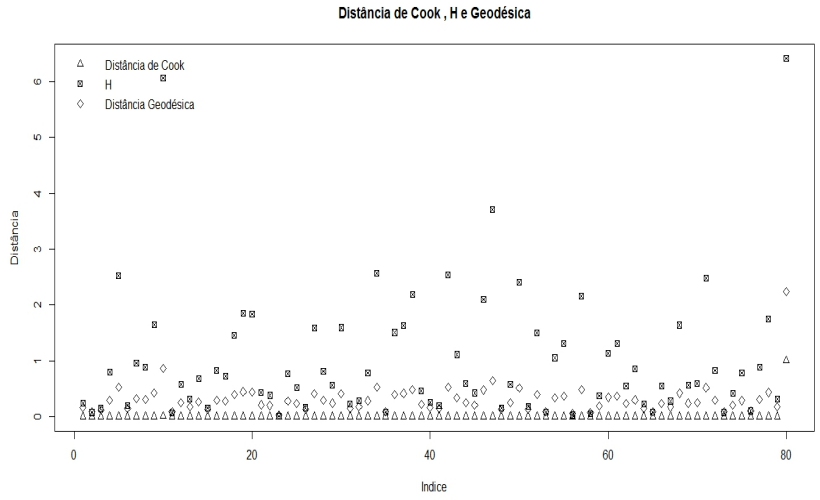

Figura 4.6: Distâncias de Cook, H e geodésica para $n = 80$ observações para valores de $k = 12$ e $k = 0,05$.

A Tabela 4.3 apresenta a taxa de detecção dos critérios para a situação em que $n = 20$, onde as 19 observações foram geradas com $k = 13$ e a vigésima observação gerada com $k = 1$. Notou-se que todos os critérios apresentam uma boa taxa de detecção para o ponto de corte, $C = 1,5$. É possível notar que para o ponto de corte $C = 3$, os critérios

$D1$ e $D3$, continaram apresentando uma boa taxa de detecção, porém os critérios $D2$ e $D4$, diminuíram as taxas de detecção para esse esse ponto de corte. Notamos ainda que a medida que aumentamos o ponto de corte, a taxa de detecção tende a diminuir. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de detecção dos critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, quando comparados aos critérios $D2$ e $D4$. Para o tamanho de amostra $n = 50$, onde as 49 observações foram geradas com $k = 13$ e a vigésima observação gerada com $k = 1$.

Notou-se que todos os critérios apresentaram uma boa taxa de detecção para o ponto de corte, $C = 1,5$, e suas taxas de detecção com esse mesmo ponto de corte foram maiores que as taxas de detecção para o tamanho de amostra $n = 20$. É possível notar que o mesmo ocorreu para o ponto de corte $C = 3$, onde os critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, porém os critérios $D2$ e $D4$, diminuíram as taxas de detecção para esse esse ponto de corte. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de detecção dos critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, quando comparados aos critérios $D2$ e $D4$. Porém as taxas de detecção de todos os critérios para esse tamanho de amostra foram maiores, quando comparados ao tamanho de amostra $n = 20$.

Para o tamanho de amostra $n = 80$, onde as 79 observações foram geradas com $k = 13$ e a vigésima observação gerada com $k = 1$. Notou-se que todos os critérios apresentaram uma boa taxa de detecção para o ponto de corte, $C = 1,5$, e suas taxas de detecção com esse mesmo ponto de corte foram maiores que as taxas de detecção para o tamanho de amostra $n = 20$ e $n = 50$. É possível notar que o mesmo ocorreu para o ponto de corte $C = 3$, onde os critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, porém os critérios $D2$ e $D4$, diminuíram as taxas de detecção para esse esse ponto de corte. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de detecção dos critérios $D1$ e $D3$, continuaram apresentando uma boa taxa de detecção, quando comparados aos

critérios $D2$ e $D4$. Porém as taxas de detecção de todos os critérios para esse tamanho de amostra foram maiores, quando comparados ao tamanho de amostra $n = 20$ e $n = 80$.

Tabela 4.3: Taxa de detecção de outliers para diferentes tamanhos de amostras e valores de $k = 13$ e $k = 1$

C	n	Critérios			
		$D1$	$D2$	$D3$	$D4$
1,5	20	1	$0,862$	1	$0,898$
	50	1	$0,881$	1	$0,908$
	80	1	$0,876$	1	$0,910$
3	20	1	$0,793$	1	$0,788$
	50	1	$0,820$	1	$0,806$
	80	1	$0,770$	1	$0,795$
4	20	1	$0,742$	1	$0,701$
	50	1	$0,775$	1	$0,724$
	80	1	$0,747$	1	$0,741$
6	20	1	$0,634$	1	$0,524$
	50	1	$0,642$	1	$0,549$
	80	1	$0,622$	1	$0,563$

A Figura 4.7 apresenta o gráfico das distâncias de Cook, H e geodésica para $n = 20$ observações para valores de $k = 13$ e $k = 1$. Claramente, podemos observar que para estes valores, a distância H tem melhor desempenho que as distâncias de Cook e geodésica com respeito à taxa de detecção de outliers. Por outro lado, a distancia de Cook apresenta melhor desempenho quando comparada à distância geodésica. As Figuras 4.8 e 4.9 apresentam os gráficos das distâncias de Cook, H e geodésica para $n = 50$ e $n = 80$ observações respectivamente, para valores de $k = 13$ e $k = 1$. Em ambas figuras, podemos observar que o desempenho das três distâncias com respeito à taxa de detecção de outliers é similar a Figura 4.7. Notamos ainda que em todas as figuras, a distância H apresenta valores mais dispersos, enquanto que na distância de Cook os valores estão mais concentrados.

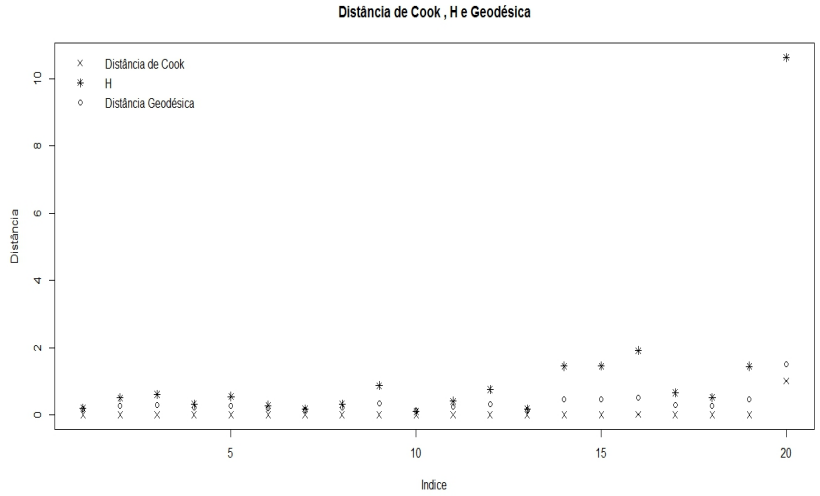

Figura 4.7: Distâncias de Cook, H e geodésica para $n = 20$ observações para valores de $k = 13$ e $k = 1$.

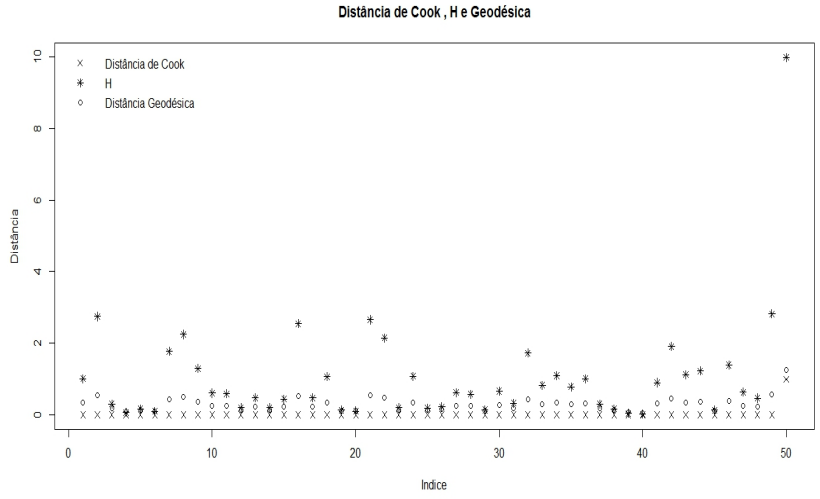

Figura 4.8: Distâncias de Cook, H e geodésica para $n = 50$ observações para valores de $k = 13$ e $k = 1$.

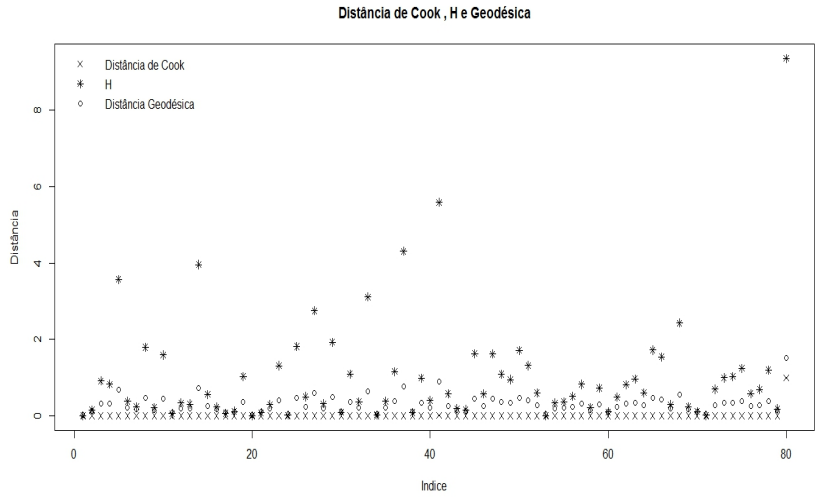

Figura 4.9: Distâncias de Cook, H e geodésica para $n = 80$ observações para valores de $k = 13$ e $k = 1$.

Para calcularmos a taxa de erro fixamos valores para $k = 0,01$, $k = 0,05$ e $k = 1$ e variamos os tamanhos das amostras para $n = 20$, $n = 50$ e $n = 80$. A tabela 4.4 apresenta a taxa de erro dos critérios para a situação em que $n = 20$, gerada com $k = 0,01$. Notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1.5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro baixas: $0,001$, $0,013$ e $0,009$.

É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já o critério $D3$, apresenta uma taxa de erro de $0,018$ e os critérios $D2$ e $D4$, taxas de erro iguais a zero, para esse esse ponto de corte. Notamos ainda que a medida que aumentamos o ponto de corte, a taxa de erro tende a diminuir.

Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, foi alta quando comparados aos critérios $D2$, $D3$ e $D4$. Para o tamanho de amostra $n = 50$, geradas com $k = 0,01$, notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$ de $0,481$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro

iguais a zero. É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero.

Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, continuaram apresentando uma alta taxa de erro, quando comparados aos critérios $D2$ e $D4$. Porém as taxas de erro dos critérios $D2$, $D3$ e $D4$ para esse tamanho de amostra foram menores, quando comparados ao tamanho de amostra $n = 20$, exceto o critério $D1$, que apresenta taxa de erro alta quando comparado a taxa de erro de amostra $n = 20$, para todos os pontos de corte.

Para o tamanho de amostra $n = 80$, geradas com $k = 0,01$, notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$ de $0,433$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero. É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero.

Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, continuaram apresentando uma alta taxa de erro, quando comparados aos critérios $D2$ e $D4$. Porém as taxas de erro dos critérios $D2$, $D3$ e $D4$ para esse tamanho de amostra foram menores, quando comparados aos tamanhos de amostra $n = 20$ e $n = 50$, exceto o critério $D1$, que apresenta taxa de erro alta quando comparado a taxa de erro de amostra $n = 20$.

Tabela 4.4: Taxa de erro para diversos tamanhos de amostras com $k = 0,01$

C	n	Critérios			
		$D1$	$D2$	$D3$	$D4$
1,5	20	$0,363$	$0,001$	$0,013$	0.009
	50	$0,481$	0	0	0
	80	$0,433$	0	0	0
3	20	$0,309$	0	$0,018$	0
	50	$0,347$	0	0	0
	80	$0,384$	0	0	0
4	20	$0,253$	0	$0,013$	0
	50	$0,295$	0	0	0
	80	$0,320$	0	0	0
6	20	$0,284$	0	$0,018$	0
	50	$0,332$	0	0	0
	80	$0,369$	0	0	0

A Tabela 4.5 apresenta a taxa de erro dos critérios para a situação em que $n = 20$, gerada com $k = 0,05$. Notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro baixas: 0, $0,012$ e $0,010$. É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já o critério $D3$, apresenta uma taxa de erro de $0,012$ e os critérios $D2$ e $D4$, taxas de erro iguais a zero, para esse esse ponto de corte. Notamos ainda que a medida que aumentamos o ponto de corte, a taxa de erro tende a diminuir. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, foi alta quando comparados aos critérios $D2$, $D3$ e $D4$. Para o tamanho de amostra $n = 50$, geradas com $k = 0,05$, notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$ de $0,489$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero. É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, continuaram apresentando uma alta taxa de erro, quando comparados aos critérios $D2$, $D3$ e $D4$. Porém as taxas de erro dos critérios $D2$, $D3$ e $D4$ para esse tamanho de

amostra foram iguais a zero, quando comparados ao tamanho de amostra $n = 20$, exceto o critério $D1$, que apresenta taxa de erro alta, para todos os pontos de corte, quando comparado a taxa de erro de amostra $n = 20$.

Para o tamanho de amostra $n = 80$, geradas com $k = 0,05$, notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$ de $0,427$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero. É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, continuaram apresentando uma alta taxa de erro, quando comparados aos critérios $D2$, $D3$ e $D4$. Porém as taxas de erro dos critérios $D2$, $D3$ e $D4$ para esse tamanho de amostra foram iguais a zero, quando comparados aos tamanhos de amostra $n = 20$, exceto o critério $D1$, que apresenta taxa de erro alta quando comparado a taxa de erro de amostra $n = 20$ e $n = 50$.

Tabela 4.5: Taxa de erro para diversos tamanhos de amostras com $k = 0,05$

C	n	Critérios			
		$D1$	$D2$	$D3$	$D4$
1,5	20	$0,365$	0	$0,012$	0.010
	50	$0,489$	0	0	0
	80	$0,427$	0	0	0
3	20	$0,271$	0	$0,012$	0
	50	$0,363$	0	0	0
	80	$0,326$	0	0	0
4	20	$0,253$	0	$0,012$	0
	50	$0,352$	0	0	0
	80	$0,388$	0	0	0
6	20	$0,240$	0	$0,012$	0
	50	$0,348$	0	0	0
	80	$0,382$	0	0	0

A Tabela 4.6 apresenta a taxa de erro dos critérios para a situação em que $n = 20$, gerada com $k = 1$. Notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro baixas: $0,004$, $0,082$ e $0,023$. É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já o critério $D3$, apresenta uma taxa de erro de $0,077$ e os critérios $D2$ e $D4$, taxas de erro iguais a zero, para esse esse ponto de corte. Notamos ainda que a medida que aumentamos o ponto de corte, a taxa de erro tende a diminuir. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, foi alta quando comparados aos critérios $D2$, $D3$ e $D4$.

Para o tamanho de amostra $n = 50$, geradas com $k = 1$, notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$ de $0,460$, já o critério $D2$ apresenta taxa de erro igual a zero, e os critérios $D3$ e $D4$, apresentam taxas de erro: $0,002$ e $0,001$, respectivamente. É possível notar que para o ponto de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já os critérios $D2$ e $D4$, apresentaram taxas de erro iguais a zero, e o critério $C3$, taxa de erro igual a $0,002$. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, continuaram apresentando uma alta taxa de erro, quando comparados aos critérios $D2$, $D3$ e $D4$. Porém as taxas de erro dos critérios $D2$ e $D4$ para esse tamanho de amostra foram iguais a zero, e o critério $D3$, com taxa de erro $0,002$. Quando comparados ao tamanho de amostra $n = 20$, os critérios $D2$, $D3$ e $D4$, apresentam taxas de erro baixas, exceto o critério $D1$, que apresenta taxa de erro alta, para todos os pontos de corte, quando comparado a taxa de erro de amostra $n = 20$.

Para o tamanho de amostra $n = 80$, geradas com $k = 1$, notou-se que o critério $D1$ apresenta uma taxa de erro alta para o ponto de corte $C = 1,5$ de $0,516$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero. É possível notar que para o ponto

de corte $C = 3$, o critério $D1$ apresenta uma taxa de erro menor quando comparado ao ponto de corte $C = 1,5$, já os critérios $D2$, $D3$ e $D4$, apresentaram taxas de erro iguais a zero. Para os pontos de cortes iguais a $C = 4$ e $C = 6$, as taxas de erro do critério $D1$, continuaram apresentando uma alta taxa de erro, quando comparados aos critérios $D2$, $D3$ e $D4$. Porém as taxas de erro dos critérios $D2$, $D3$ e $D4$ para esse tamanho de amostra foram iguais a zero, quando comparados aos tamanhos de amostra $n = 20$, exceto o critério $D1$, que apresenta taxa de erro alta quando comparado a taxa de erro de amostra $n = 20$ e $n = 50$.

Tabela 4.6: Taxa de erro para diversos tamanhos de amostras com $k = 1$

C	n	Critérios			
		$D1$	$D2$	$D3$	$D4$
1,5	20	$0,404$	$0,004$	$0,082$	0.023
	50	$0,460$	0	$0,002$	$0,001$
	80	$0,516$	0	0	0
3	20	$0,334$	0	$0,077$	0
	50	$0,363$	0	$0,002$	0
	80	$0,373$	0	$0,001$	0
4	20	$0,339$	0	$0,082$	0
	50	$0,381$	0	0.002	0
	80	$0,410$	0	0	0
6	20	$0,321$	0	$0,077$	0
	50	$0,356$	0	$0,002$	0
	80	$0,368$	0	$0,001$	0

4.2 Dados Reais

Inicialmente coletamos três conjuntos de dados reais. O primeiro conjunto de dados refere-se as posições de pólos determinados apartir de estudos dos solos paleomagnéticos de Nova Caledônia. As coordenadas usadas foram: latitude e longitude denotadas respectivamente por θ e ϕ para uma amostra de tamanho $n = 50$. Foram calculados 50 vetores vetores de uma certa distribuição a princípio desconhecida, cada um deles com coordenadas (x, y, z), de tal forma que $x = sen\theta\cos\phi$, $y = sen\theta\sin\phi$ e $z = \cos\theta$, e foram alocados em uma matriz u, ver (Apêndice (A.3)). Calculamos a matriz de covariância S, com todas as observações e as matrizes de covariância *somai*, omitindo a i-ésima observação. Destas matrizes foram calculados seus autovetores e autovalores. Para calcularmos o valor de k estimado da matriz de covariância S e os valores dos k estimados das matrizes de covariância *somai* , usamos a equação (2.11).

Como o valor dos k estimados foram maiores que zero, tomamos o autovetor associado ao maior autovalor. Os valores dos k estimados foram concatenados aos seus respectivos autovetores, denotados por θ_{chap} relacionado a matriz S e θ_{chap1} relacionado a matriz *somai*, ver (Apêndice (A.3)). Foram calculadas três distâncias: distância de Cook, usando a equação (3.1), estatística de teste H, usando a equação (3.6) e a distância geodésica (ver Apêndice (A.3)). A Tabela 4.7 apresenta o cálculo destas distâncias. É possível notar que os valores das distância de Cook, estão bem próximos uns dos outros, não apresenta nenhum valor exorbitante. A distância fornecida pela estatística de teste H, relacionada ao teste outlier para discordância, não apresenta nenhum valor de H distante dos demais, se uma dessas observações apresentasse um valor de H, muito grande, esta observação seria considerada um outlier. Para a distância geodésica, é também possível notar que os valores das distâncias estão bem próximos uns dos outros.

Tabela 4.7: Cálculo das Distâncias para $n = 50$ observações.

n	dcook	H	distância geodésica
1	0,0681997455	2,258353451	1,38284922
2	0,0008649093	0,898194664	0,66627856
3	0,0060233466	1,336841709	0,85479574
4	0,0700602323	2,275871870	1,40436048
5	0,0240463568	1,729657647	1,03327900
6	0,0324834537	1,855545440	1,09728184
7	0,0397158133	0,017077514	0,08511169
8	0,0008121648	1,027586970	0,72182842
9	0,0409482260	0,002136180	0,03006521
10	0,0660692383	2,237775662	1,35988801
11	0,0617484623	2,195248201	1,31896480
12	0,0663951907	2,240943068	1,36326194
13	0,0332322905	0,099814689	0,20717258
14	0,0015732999	0,824199660	0,63360279
15	0,0085762781	1,415532889	0,88932943
16	0,0395112908	0,019573775	0,09115272
17	0,0110734389	0,481211451	0,46841892
18	0,0400616576	1,954060901	1,15115965
19	0,0099021851	1,449395480	0,90390308
20	0,0404743624	0,007862928	0,05767393
21	0,0448314551	2,012112999	1,18605203
22	0,0023008261	1,174406072	0,78486941
23	0,0089255365	0,533474290	0,49598077
24	0,0198801312	0,303172064	0,36627558
25	0,0024153096	1,187547358	0,79113637
26	0,0286910684	0,163378290	0,26594191
27	0,0023507093	1,184378490	0,78983855
28	0,0035904859	0,710316824	0,58177754
29	0,0018043481	1,141627507	0,77087590
30	0,0685033431	2,261118126	1,38587918
31	0,0233472473	1,717287386	1,02682488
32	0,0007923082	1,047976938	0,73132888
33	0,0342193149	0,086707727	0,19290534
34	0,0381995567	0,035801423	0,12338617
35	0,0194568815	1,652358045	0,99650654
36	0,0371232814	0,049336318	0,14496760
37	0,0208748057	1,675680940	1,00688406
38	0,0759462898	2,330240038	1,50755780
39	0,0297538639	0,147850979	0,25291647
40	0,0195684908	0,307480407	0,36955869
41	0,0007550154	0,944919649	0,68604189
42	0,0053407476	0,645206196	0,55045665
43	0,0056318924	0,634811523	0,54550924
44	0,0009957945	0,878110162	0,65756141
45	0,0766244037	2,336349872	1,53377721
46	0,0334348943	0,097227864	0,20430509
47	0,0242403907	0,230112144	0,31758716
48	0,0255175173	0,210649087	0,30313985
49	0,0256663600	0,207974094	0,30139765
50	0,0236730624	0,238974881	0,32394987

A Figura 4.10 apresenta o gráfico das distâncias de Cook e da estatística H do teste de outlier para discordância para $n = 50$ observações. A Figura 4.10 mostra claramente que os valores das distâncias de Cook estão bem concentrados em torno de zero, já os valores das distãncias fornecidas pela estatística de teste H, estão mais instáveis, variando entre 0 e 2, porém nenhuma das distâncias comparadas apresentaram algum ponto influente. A Figura 4.11, mostra o gráfico das distâncias de Cook e distância Geodésica para $n = 50$ observações. Analisando a Figura 4.11, novamente, observamos que os valores das distâncias geodésica estão variando entre 0 e 1, entretanto os valores das distãnciasde Cook estão concentrados em torno de zero. Em ambas distâncias não foram detectados outliers. A Figura 4.12 apresenta o gráfico das distância H e da distância geodésica para $n = 50$ observações. Analisando a Figura 4.12, novamente, observamos que os valores de ambas as distâncias estão mais dispersos, porém nenhum desses valores possui uma discrepância significativa para ser considerado um outlier.

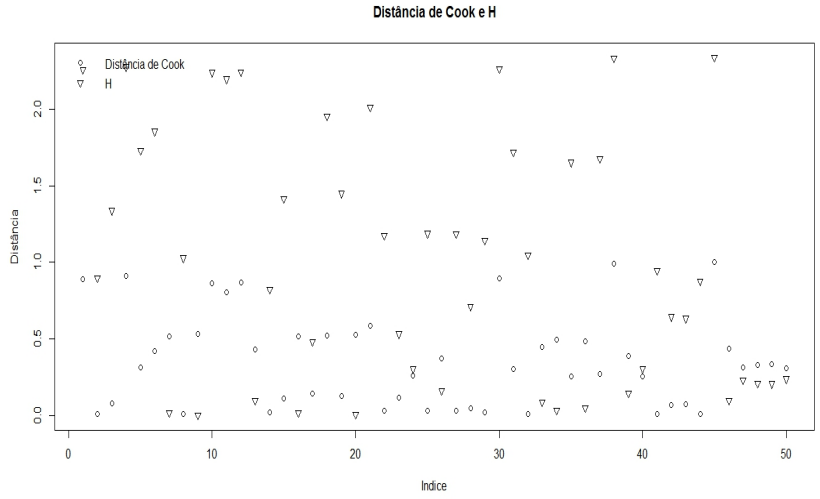

Figura 4.10: Distâncias de Cook e H para $n = 50$ observações.

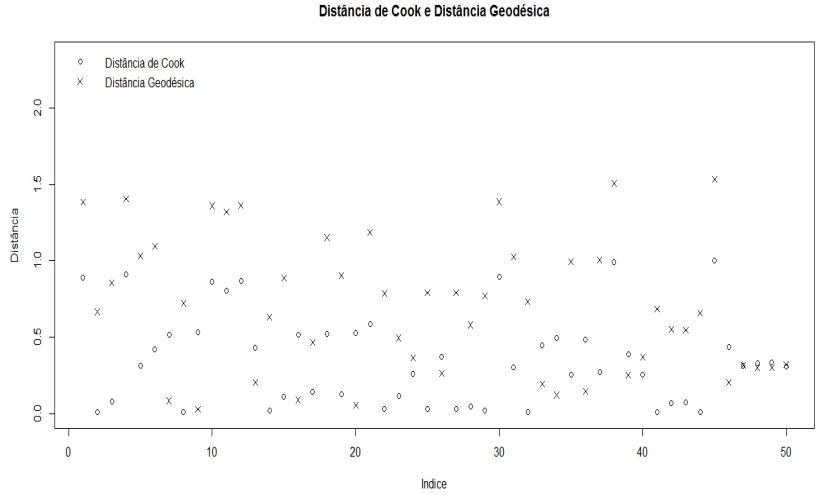

Figura 4.11: Distâncias de Cook e distância geodésica para $n = 50$ observações.

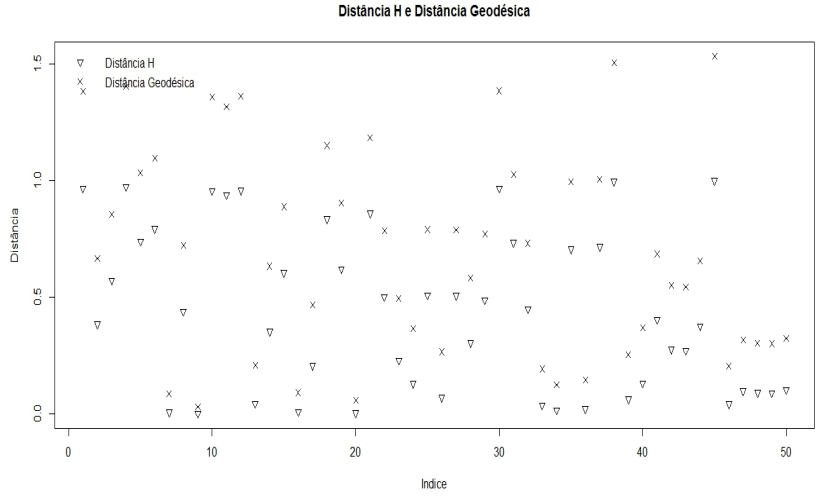

Figura 4.12: Distâncias H e geodésica para $n = 50$ observações.

O segundo conjunto da dados refere-se a orientações de superfícies de clivagem de plano axial de dobras em Ordoviciano turbiditos, cujas coordenadas em estudo são: mergulho e direção de mergulho, denotadas respectivamente por θ e ϕ para uma amostra de tamanho $n = 72$ observações. Foram calculados 72 vetores da distribuição de Watson, cada um deles com coordenadas (x, y, z), de tal forma que $x = sen\theta \cos\phi$, $y = sen\theta \sin\phi$ e $z = \cos\theta$, e alocados em uma matriz u, ver (Apêndice (A.4)).

Calculamos a matriz de covariância S, com todas as observações e as matrizes de covariância *somai*, omitindo a i-ésima observação. Destas matrizes foram calculados seus autovetores e autovalores. Como o valor dos k estimados foram maiores que zero, tomamos o autovetor associado ao maior autovalor. Os valores dos k estimados foram concatenados aos seus respectivos autovetores, denotados por θ_{chap} relacionado a matriz S e θ_{chap1} relacionado a matriz *somai*, ver (Apêndice (A.4)). Foram calculadas três distâncias: distância de Cook, usando a equação (3.1), estatística de teste H, usando a equação (3.6) e a distância geodésica (ver Apêndice (A.4)).

A Tabela 4.8 apresenta o cálculo destas distâncias. É possível notar que os valores das distância de Cook, estão bem próximos uns dos outros, não apresenta nenhum valor exorbitante. A distância fornecida pela estatística de teste H, relacionada ao teste outlier para discordância, não apresenta nenhum valor de H distante dos demais, se uma dessas observações apresentasse um valor de H, muito grande, esta observação seria considerada um outlier. Para a distância geodésica, é possível notar que os valores dessas distâncias estão concentrados em torno de 0 e 1.

Tabela 4.8: Cálculo das Distâncias para $n = 72$ observações.

n	dcook	H	distância geodésica
1	8,089763e−02	2,5173292085	1,47190320
2	2,407741e−02	0,1580651715	0,25186393
3	2,880634e−02	0,0798665520	0,17796676
4	1,679734e−03	1,1905716803	0,75438855
5	1,063936e+01	1,3407757623	0,81246829
6	1,454118e−02	0,3430051100	0,37679299
7	9,863905e−03	0,4576561467	0,43890174
8	7,737528e−02	2,4833561983	1,41832317
9	6,652534e−03	1,4167313920	0,84361511
10	2,847462e−02	1,8903788829	1,04064949
11	3,400792e−02	0,0009111994	0,01896851
12	3,378787e−02	0,0041165872	0,04025796
13	3,348376e−03	0,6872037867	0,54683668
14	5,437132e−04	1,0614015510	0,70216075
15	2,048906e−02	0,2229644632	0,30041426
16	1,064155e+01	1,4183848277	0,84335747
17	2,977817e−02	0,0644782105	0,15992786
18	1,148012e−03	1,1411683651	0,73416560
19	3,397012e−02	0,0014606251	0,02400193
20	3,060411e−02	0,0517739799	0,14309390
21	9,368008e−03	0,4743769400	0,44629541
22	2,245087e−02	0,1868755811	0,27432323
23	2,046235e−02	0,2224521661	0,30080918
24	3,389190e−02	0,0025990086	0,03200181
25	8,349560e−02	2,5419667539	1,56414373
26	8,352032e−04	1,1047067201	0,71942237
27	2,640858e−03	0,7202598000	0,56192947
28	5,939000e−02	2,2960023644	1,25428193
29	3,709968e−02	2,0194578082	1,10075416
30	7,992227e−02	2,5080659877	1,45513307
31	2,253569e−02	0,1854349827	0,27316609
32	7,535427e−02	2,4635188646	1,39425393
33	8,335718e−02	2,5406672924	1,54728064
34	7,679739e−02	2,4776297291	1,41079237
35	2,239565e−02	0,1876761190	0,27508710
36	2,194498e−02	0,1959923442	0,28114165
37	7,760153e−02	2,4855070985	1,42099692
38	4,818420e−02	2,1647221100	1,17526250
39	1,904903e−02	1,7218440654	0,96657892

Tabela 4.8: (Continuação.)

n	dcook	H	distância geodésica
40	9,689950e−03	1,5096453598	0,88086107
41	3,013749e−02	0,0589780478	0,15277889
42	2,055088e−02	0,2216184497	0,29961284
43	1,856097e−03	0,7745784280	0,58442091
44	2,447848e−03	0,7325823329	0,56704625
45	1,063653e+01	0,7955366913	0,59378433
46	3,691078e−03	0,6737162000	0,54031925
47	1,899333e−02	0,2517424042	0,31983124
48	7,034451e−04	1,1048057866	0,72054872
49	3,399884e−02	0,0010457196	0,02028265
50	2,178081e−02	1,7760160528	0,99062967
51	6,635566e−03	1,4169944905	0,84392401
52	4,947428e−02	2,1805744951	1,18411160
53	3,016008e−02	0,0585381029	0,15235604
54	2,136688e−02	1,7674062599	0,98647208
55	2,224577e−02	1,7818789070	0,99183316
56	2,444027e−02	1,8210085373	1,00879301
57	2,229472e−02	0,1892977692	0,27646702
58	1,152439e−02	1,5561059036	0,89888519
59	6,483483e−04	1,1024928123	0,71992548
60	3,371908e−02	0,0051242804	0,04489410
61	1,996878e−02	1,7395230506	0,97392719
62	7,520909e−03	1,4459671826	0,85548398
63	3,380400e−02	0,0038863897	0,03908220
64	2,394771e−02	1,8126193028	1,00525696
65	6,700293e−02	2,3783924157	1,31425079
66	3,353484e−02	0,0078083696	0,05548082
67	3,258490e−02	0,0218585241	0,09275952
68	3,112011e−02	0,0437684457	0,13172744
69	2,265964e−02	0,1826320041	0,27153448
70	2,500365e−02	1,8307308757	1,01303110
71	3,109899e−02	0,0442270601	0,13214801
72	2,155821e−02	0,2024022525	0,28635143

A Figura 4.13 apresenta o gráfico das distâncias de Cook e da estatística H do teste de outlier para discordância para $n = 72$ observações. A Figura 4.13 mostra que os valores das distâncias de Cook estão próximos de zero, já os valores das distãncias fornecidas pela estatística de teste H, estão mais dispersos, variando entre 0 e 2, isto é, quando compara-

mos essas duas distâncias, podemos notar que os valores da distância H são maiores que os valores das distâncias de Cook, porém em nenhuma das distâncias comparadas apresentaram algum ponto outlier. A Figura 4.14, mostra o gráfico das distâncias de Cook e distância Geodésica para $n = 72$ observações. Analisando a Figura 4.14, novamente, observamos que os valores das distâncias geodésica estão variando entre 0 e 1, entretanto os valores das distãnciasde Cook estão concentrados em torno de zero. Em ambas distâncias não foram detectados outliers. A Figura 4.15 apresenta o gráfico das distância H e da distância geodésica para $n = 72$ observações. Analisando a Figura 4.15, observamos que os valores de ambas as distâncias estão mais dispersos, e em algumas observações é possível notar claramente que os valores da distância geodésica supera os valores da distância H, porém nenhum desses valores possui uma discrepância significativa para ser considerado um outlier.

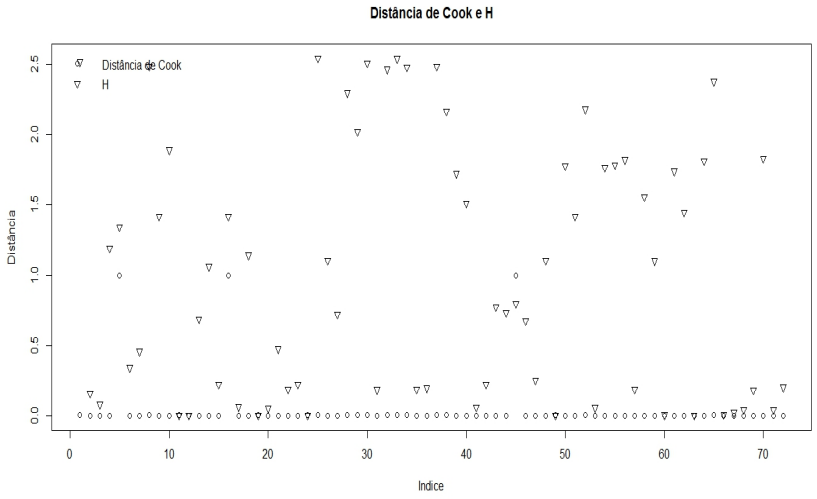

Figura 4.13: Distâncias de Cook e H para $n = 72$ observações.

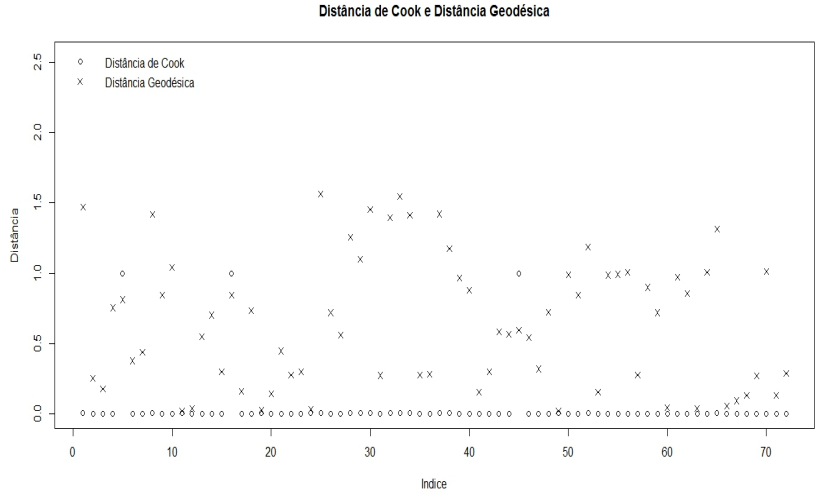

Figura 4.14: Distâncias de Cook e distância geodésica para $n = 72$ observações.

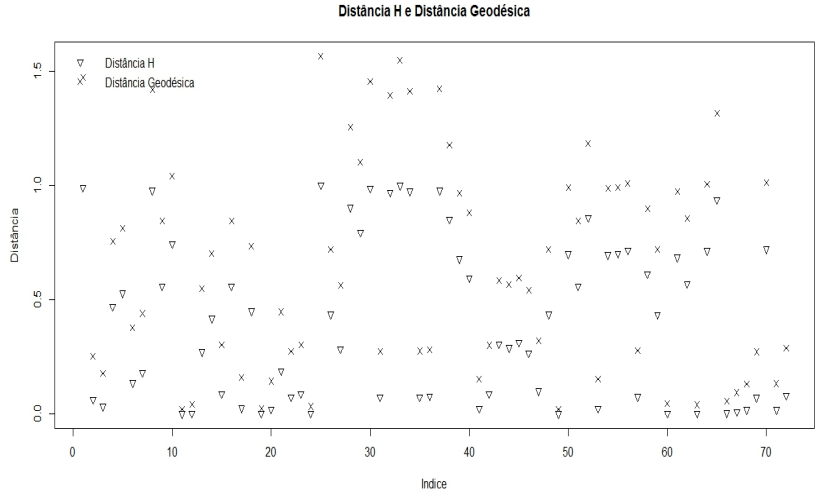

Figura 4.15: Distâncias H e geodésica para $n = 72$ observações.

O terceiro conjunto da dados refere-se a orientações de superfícies de clivagem de plano axial de dobras em Ordoviciano turbiditos, cujas coordenadas em estudo são: mergulho e direção de mergulho, denotadas respectivamente por θ e ϕ para uma amostra de tamanho $n = 75$. Foram calculados 75 vetores da distribuição de Watson, cada um deles com coordenadas (x, y, z), de tal forma que $x = sen\theta \cos\phi$, $y = sen\theta \sin\phi$ e $z = \cos\theta$, e alocados em uma matriz u, ver (Apêndice (A.5)).

Calculamos a matriz de covariância S, com todas as observações e as matrizes de covariância *somai*, omitindo a i-ésima observação. Destas matrizes foram calculados seus autovetores e autovalores. Como o valor dos k estimados foram maiores que zero, tomamos o autovetor associado ao maior autovalor. Os valores dos k estimados foram concatenados aos seus respectivos autovetores, denotados por θ_{chap} relacionado a matriz S e θ_{chap1} relacionado a matriz *somai*, (ver Apêndice (A.5)). Foram calculadas três distâncias: distância de Cook, usando a equação (3.1), estatística de teste H, usando a equação (3.6) e a distância geodésica (ver Apêndice (A.5)).

A Tabela 4.9 apresenta o cálculo destas distâncias. É possível notar que os valores das distância de Cook, estão concentrados em torno de zero, e apresenta nenhum valor exorbitante. A distância fornecida pela estatística de teste H, relacionada ao teste outlier para discordância, não apresenta nenhum valor de H distante dos demais, os valores dessas distâncias variam entre 0 e 2, porém se uma dessas observações apresentasse um valor de H, muito grande, esta observação seria considerada um outlier. Para a distância geodésica, é possível notar que os valores dessas distâncias estão concentrados em torno de 0 e 1.

Tabela 4.9: Cálculo das Distâncias para $n = 75$ observações.

n	dcook	H	distância geodésica
1	0,0087341445	0, 146627492	0, 26134909
2	0, 0054281960	0, 328388256	0, 39750201
3	0, 0071847978	0, 233473107	0, 33052115
4	0, 0038160487	0, 455511585	0, 47130093
5	0, 0073636199	0, 216548076	0, 31963759
6	0, 0103106285	0, 074177328	0, 18380240
7	0, 0061568558	0, 284081652	0, 36828561
8	0, 0112037012	0, 032741726	0, 12238294
9	0, 0012724240	1, 261714253	0, 86264007
10	0, 0110823632	0, 038036804	0, 13194254
11	0, 0063387936	0, 274934116	0, 36169061
12	0, 0103026038	0, 074584116	0, 18429467
13	0, 0049322460	1, 598653855	1, 02604077
14	0, 0054772870	0, 331003845	0, 39808474
15	0, 0163305007	2, 151238560	1, 46655922
16	0, 0126958208	2, 009402156	1, 29000412
17	0, 0168167529	2, 169131797	1, 52129923
18	0, 0024029638	0, 560136545	0, 53032767
19	0, 0083447616	0, 168508750	0, 27989895
20	0, 0016659037	1, 293620508	0, 87596973
21	0, 0117432360	0, 009647046	0, 06634889
22	0, 0102186242	0, 076587930	0, 18773443
23	0, 0049322460	1, 598653855	1, 02604077
24	0, 0094885852	1, 861678941	1, 17845962
25	0, 0037160100	1, 526642461	0, 99150498
26	0, 0102760436	1, 905585751	1, 21015779
27	0, 0011516921	1, 210609364	0, 83721379
28	0, 0077729403	0, 199380923	0, 30498311
29	0, 0077195733	0, 203382350	0, 30782094
30	0, 0057709465	1, 662310578	1, 06130535
31	0, 0105270295	0, 064075390	0, 17074475
32	0, 0006767142	0, 864740636	0, 67767690
33	0, 0009699761	1, 203735587	0, 83535679
34	0, 0109174522	0, 046421487	0, 14506985
35	0, 0161384411	2, 144265247	1, 45199556
36	0, 0052106490	0, 347258930	0, 40852419
37	0, 0017411176	1, 304384771	0, 88106477
38	0, 0037087495	0, 466169541	0, 47700605
39	0, 0050235051	0, 360322933	0, 41650756

Tabela 4.9: (Continuação.)

n	dcook	H	distância geodésica
40	0,0062841870	1,696792398	1,08084385
41	0,0070467709	0,241253775	0,33622366
42	0,0009259410	1,189763622	0,82865617
43	0,0107196366	0,055409284	0,15859136
44	0,0052106490	0,347258930	0,40852419
45	0,0104290840	0,068110570	0,17634712
46	0,0076246414	0,203109502	0,30908481
47	0,0103863632	1,908796873	1,21184309
48	0,0019124117	1,338884556	0,89829568
49	0,0004257325	1,071889827	0,77517689
50	0,0008373564	1,144237629	0,80667134
51	0,0090924197	0,130538428	0,24575144
52	0,0087321498	1,831873172	1,16120958
53	0,0014035275	0,678969765	0,59037544
54	0,0012635712	0,696484681	0,59916402
55	0,0072441282	1,750441778	1,11109897
56	0,0004997717	0,904045416	0,69682036
57	0,0010024526	1,186505296	0,82631255
58	0,0106611737	0,058038531	0,16236669
59	0,0015546033	1,319772255	0,89114014
60	0,0169363040	2,173506531	1,55092023
61	0,0149342514	2,098097639	1,38059795
62	0,0094667601	0,111987522	0,22746185
63	0,0165033060	2,157867873	1,48315938
64	0,0084131567	1,814502469	1,15000678
65	0,0033470386	1,478005439	0,96501666
66	0,0103578407	1,906231901	1,20965490
67	0,0029860174	1,441384170	0,94658396
68	0,0097411713	1,878163686	1,19062647
69	0,0118815018	1,972508374	1,25840137
70	0,0029860174	1,441384170	0,94658396
71	0,0112895992	1,947208199	1,23904949
72	0,0030428107	1,446354444	0,94897749
73	0,0119924565	1,977479872	1,26244327
74	0,0040300752	1,544114955	0,99948102
75	0,0039328410	1,532444328	0,99301254

A Figura 4.16 apresenta o gráfico das distâncias de Cook e da estatística H do teste de outlier para discordância para $n = 75$ observações. Na Figura 4.16, notou-se que os

valores das distâncias de Cook estão próximos de zero, já os valores das distâncias fornecidas pela estatística de teste H, estão mais dispersos, variando entre 0 e 2, isto é, quando comparamos essas duas distâncias, podemos notar que os valores da distância H são maiores que os valores das distâncias de Cook, porém em nenhuma das distâncias comparadas apresentaram algum ponto outlier. A Figura 4.17, mostra o gráfico das distâncias de Cook e distância Geodésica para $n = 75$ observações. Analisando a Figura 4.17, novamente, observamos que os valores das distâncias geodésica estão variando entre 0 e 1, entretanto, os valores das distâncias de Cook estão concentrados em torno de zero. Em ambas distâncias não foram detectados outliers. A Figura 4.18 apresenta o gráfico das distância H e da distância geodésica para $n = 75$ observações. Analisando a Figura 4.18, novamente, observamos que os valores de ambas as distâncias estão mais dispersos, e em algumas observações é possível notar claramente que os valores da distância geodésica supera os valores da distância H, porém nenhum desses valores possui uma discrepância significativa para ser considerado um outlier.

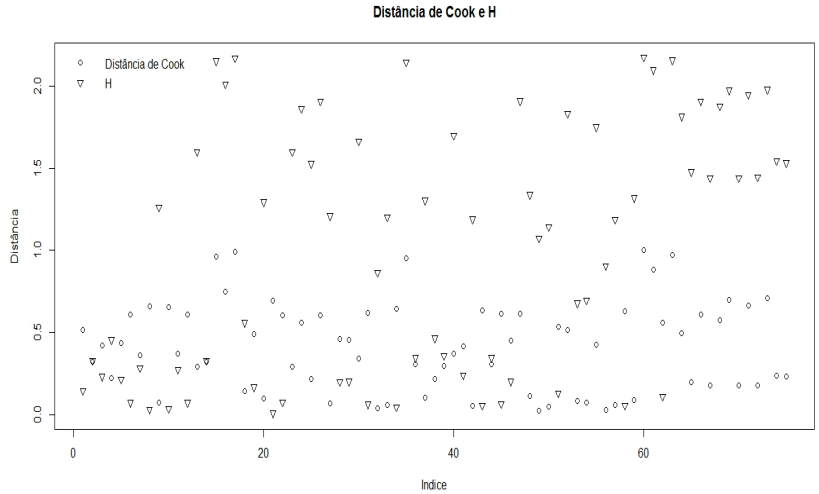

Figura 4.16: Distâncias de Cook e H para $n = 75$ observações.

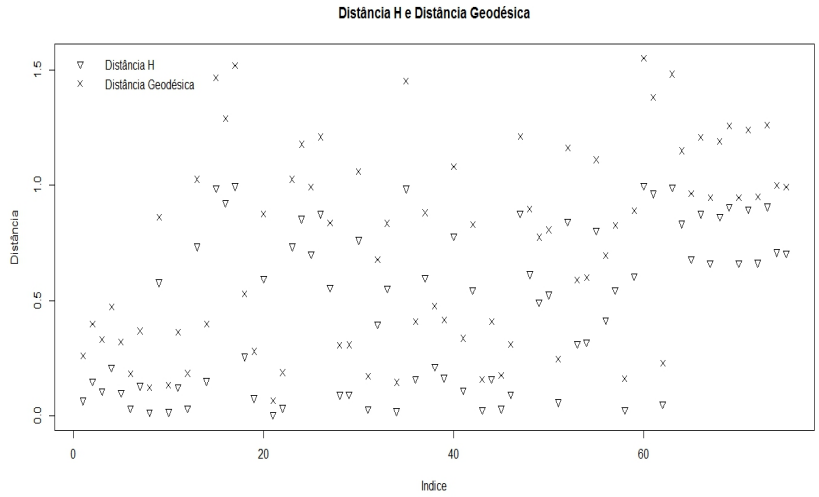

Figura 4.17: Distâncias H e distância geodésica para $n = 75$ observações.

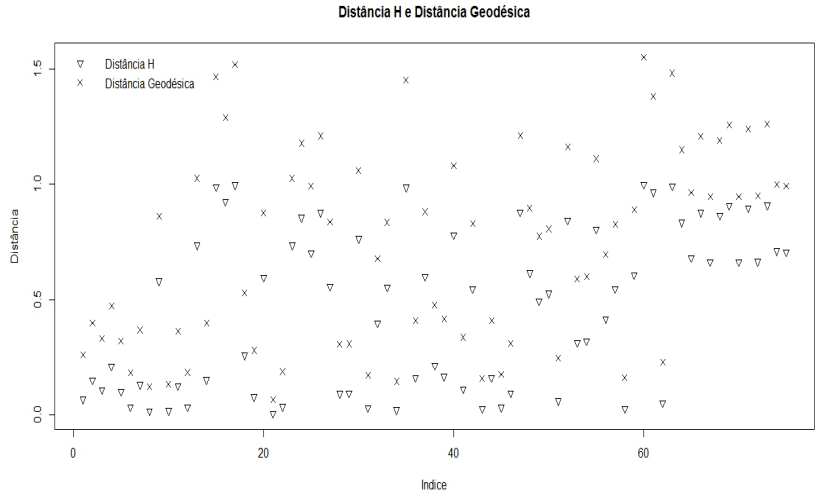

Figura 4.18: Distâncias H e geodésica para $n = 75$ observações.

A Tabela 4.10 apresenta os tamanhos das distâncias usando os pontos de corte: $C = 1,5$, $C = 3$, $C = 4$ e $C = 6$, baseado nos quatro critérios: distância de Cook, estatística H, qui-quadrado e distância geodésica, denotadas respectivamente, por $D1$, $D2$, $D3$ e $D4$, dos dados reais para $n = 50$ observações, relacionadas as posições dos pólos da partir do estudo de solos paleomagnéticos de Nova Caledônia. Notamos que à medida que aumentamos o ponto de corte em cada critério, com exceção do critério $D3$, que independente do ponto de corte permanceu fixo, houve um aumento significativo nos critérios $D2$ e $D4$, quando comparados ao critério $D1$, principalmente o critério $D2$.

As Tabelas 4.11 e 4.12 apresentam os tamanhos das distâncias usando os pontos de corte: $C = 1,5$, $C = 3$, $C = 4$ e $C = 6$, baseado nos quatro critérios para os dados reais para $n = 72$ e $n = 75$ observações, respectivamente. Ambas Tabelas estão relacionadas as orientações de clivagem de plano axial de dobras em superfícies ordovicianas turbiditosas. Analisando a Tabela 4.11, podemos observar que quanto maior for o ponto de corte, maior será a distância em cada um dos critérios. Dentre estes critérios, o critério $D2$ apresenta a maior distância. Note que no critério $D1$, os valores das distâncias continuam pequenos, mesmo quando aumentamos os pontos de corte. Analisando a Tabela 4.12, o critério $D2$ também apresenta um aumento siginficativo, quando comparado aos demais critérios, com exceção do critério $D3$, que permanece fixo. Analisando as três Tabelas, notamos que a medida que aumentamos o número de observações, o critério $D1$ apresenta distâncias menores, quando comparado aos outros critérios.

Tabela 4.10: Distâncias de Cook, H, qui-quadrado e geodésica para $n = 50$ observações.

$D1$	$D2$	$D3$	$D4$
C=1,5			
0,090841	3,977245	7,814728	2,112352
C=3			
0,141967	6,237202	7,814728	3,19788
C=4			
0,176051	7,743841	7,814728	3,921565
C=6			
0,244219	10,75712	7,814728	5,368935

Tabela 4.11: Distâncias de Cook, H, qui-quadrado e geodésica para $n = 72$ observações.

$D1$	$D2$	$D3$	$D4$
C=1,5			
0,069975	4,166092	7,814728	2,069272
C=3			
0,106146	6,556168	7,814728	3,147915
C=4			
0,130260	8,149552	7,814728	3,867110
C=6			
0,178488	11,336320	7,814728	5,305201

Tabela 4.12: Distâncias de Cook, H, qui-quadrado e geodésica para $n = 75$ observações.

$D1$	$D2$	$D3$	$D4$
C=1,5			
0,020874	3,805567	7,814728	2,157842
C=3			
0,031390	5,948823	7,814728	3,253658
C=4			
0,038401	7,377660	7,814728	3,984442
C=6			
0,052422	10,235340	7,814728	5,446011

CAPÍTULO 5

Conclusões

Resumimos as principais contribuições desta dissertação nos seguintes itens:

- No Capítulo 3, usamos o logartimo da verossimilhança da distribuição de Watson dada pela equação (2.3) no caso $p = 3$ e derivamos essa função com relação a cada um de seus parâmetros: μ_1, μ_2, μ_3 e k. Obtivemos a matriz de informação de Fisher observada, calculamos a distância de Cook e usamos quatro critérios com o objetivo de detectar a existência de observações influentes nos dados da distribuição de Watson. São eles: ponto de corte para a distância proposto por Cook (1977), teste de outlier para discordância proposto por Fisher *et al.* (1985), quantil de uma qui-quadrado proposto por Cook (1977) e distância geodésica.

- Utilizamos uma proposta sobre medidas de influência com base no tamanho de amostras de um determinado conjunto de dados. Este é o enfoque do Capítulo 4, onde desenvolvemos o estudo de alguns métodos de detecção de outliers e pontos influentes em amostras de tamanhos: $n = 20$, $n = 50$ e $n = 80$.

Além dessas contribuições, podemos tirar as seguintes conclusões:

- O estudo dos critérios para o caso dos dados simulados, com respeito a taxa de detecção mostraram-se eficientes para o ponto de corte $C = 1,5$. Em todos os critérios nesse caso, observou-se que as taxas de detecção foram altas, em contrapartida à medida que aumentamos os pontos de corte, as taxas de detecção diminuíram, devido a perda de sensibilidade destes critérios. Com relação à taxa de erro, observamos que os critérios $D2$ e $D4$, apresentaram taxas de erro baixas e à medida que aumentamos os pontos de corte, estas taxas diminuíram cada vez mais. Entretanto para valores de ponto de corte muito pequenos, pode acontecer do critério detectar um outlier, quando na verdade ele não é. E quando o ponto de corte for muito grande, o critério pode não detectar um outlier, quando na verdade ele é um outlier, isto é, estes critérios são bastantes flexíveis.

- Em relação ao estudo dos critérios para os três conjuntos de dados reais, observamos que o critério H permaneceu mais instável, ísto é, os seus valores estavam mais dispersos quando comparado a distância de Cook e a distância geodésica.

- Em suma, entre todos os critérios apresentados nesta tese, aqueles que produziram melhores desempenhos na taxa de detecção e na taxa de erro, foram os critérios $D2$ e $D4$, denotados pelos critérios: teste outlier para discordância e distância geodésica, respectivamente. Vale ressaltar que ambos os critérios apresentam taxas de detecção semelhantes para o ponto de corte $C = 1,5$.

- Nós sugerimos para os usuários utilizar os critérios $D2$ e $D4$ com o ponto de corte $C = 1,5$. Esses critérios forneceram as melhores taxas de detecção e erro dentre as

possibilidades consideradas.

Várias linhas de pesquisas podem ainda ser tratadas, tais como:

- Obter a estimação do parâmetro de concentração k, através do método iterativo de colocação de raiz de Newton-Raphson, usando o algoritmo com os seguintes passos:

  1. Apartir de k_0, usando o método de Newton-Raphson resolvemos a equação $g(a,c,k) - r = 0$ por iteração

$$k_{n+1} = k_n - \frac{g(a,c;k_n) - r}{g^{'}(a,c;k_n)}, \quad n = 0, 1, \ldots \tag{5.1}$$

  2. Esta iteração pode ser simplificada reescrevendo $g'(a,c;k)$ como

$$g'(a,c;k) = \frac{M''(a,c;k)}{M(a,c;k)} - \left(\frac{M''(a,c;k)}{M(a,c;k)}\right)^2, \tag{5.2}$$

  Usando as duas identidades seguintes

$$M''(a,c;k) = \frac{a(a+1)}{c(c+1)} M(a+2, b+2; k); \tag{5.3}$$

$$M(a+2, b+2; k) = \frac{(c+1)(-c+k)}{(a+1)k} M(a+1, c+1; k) + \frac{(c+1)c}{(a+1)k} M(a,c;k) \tag{5.4}$$

Usando as equações (5.3) e (5.4), podemos reescrever a derivada (5.2) como

$$g'(a,c;k) = (1-c/k)g(a,c;k) + (a/k) - (g(a,c;k))^2.$$

A principal consequência dessas simplificações é que iteração (5.1) pode ser implementado com apenas uma avaliação da relação de $g(a,c;k_n) = M'(a,c;k_n)/M(a,c;k_n)$, que é uma tarefa não trivial em si mesma. Uma visão sobre esta dificuldade é oferecido por meio de observações em Gautschi (1977) e Gil *et al.*, (2007). No pior dos casos, pode-se ter que calcular o numerador e o denominador separadamente (usando multi-ponto flutuante de precisão aritmética), e depois dividir. Se o fizer, pode exigir várias milhões de operações com ponto flutuante de precisão estendida, o que é muito indesejável.

- Combinar a Distância de Cook com a distância geodésica e analisar os quatro critérios discutidos.

- Usar a distância de Kullback-Leibler para detectar pontos influentes definida por

$$\begin{aligned} d_{KL} &= \frac{1}{2}[D_{KL}(x_1,x_2) + D_{KL}(x_2,x_1)] \\ &= \frac{1}{2}\left[\int_X f_{x1}(\underline{x};\theta_1)\log\frac{f_{x1}(\underline{x};\theta_1)}{f_{x2}(\underline{x};\theta_2)}d\underline{x} + \int_X f_{x2}(\underline{x};\theta_2)\log\frac{f_{x1}(\underline{x};\theta_2)}{f_{x2}(\underline{x};\theta_1)}d\underline{x}\right] \\ &= \frac{1}{2}\left[\int_X f_{x1}(\underline{x};\theta_1) - f_{x2}(\underline{x};\theta_2)\right]\log\frac{f_{x1}(\underline{x};\theta_1)}{f_{x2}(\underline{x};\theta_2)}d\underline{x}, \end{aligned}$$

onde x_1 e x_2 são vetores reais aleatórios p-dimensional em X e d$\underline{x}$ o elemento diferencial dado pela equação

$$d\underline{x} = \prod_{i=1}^{n} d\Re\{x_i\}d\Im\{x_i\},$$

onde x_i é o i-ésimo elemento de X, (ver Goodman, 1963).

Ao finalizar este trabalho, esperamos ter dado uma contribuição relevante à área de pesquisa que trata da identificação de pontos influentes de uma amostra aleatória da distribuição Watson.

Referências

[1] ABRAMOWITZ, M.; STEGUN, I. A. *Handbook of Mathematical Functions*. New York: Dover, 1965. 1044p.

[2] ANDREWS, G. E.; ASKEY, R.; ROY, R. *Special Functions*. Inglaterra: Cambridge University Press, 1999. 365p.

[3] BANERJEE, A.; DHILLON, I. S.; GHOSH, J.; SRA. S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, USA, v.6, p. 1345 – 1382, 2005.

[4] BARNETT, V.; LEWIS, T. *Outliers in Statistical Data*. New York: John Wiley & Sons, 1978. 365p.

[5] BARNETT, V. Principles e and Methods for Handling Outliers in Data Sets. *Academics Press*, USA, p. 131 – 166, 1983.

[6] BARNETT, V.; LEWIS, T. *Outliers in Statistical Data*. 2ed. New York: John Wiley & Sons, 1984. 463p.

[7] BARNETT, V.; LEWIS, T. *Outliers in Statistical Data*. New York: John Wiley & Sons, 1994. 604p.

[8] BECKMAN, R. J.; COOK, R. D. Outliers. *Technometrics*, USA, v. 25, p. 119 – 163, 1983.

[9] BIJRAL, A.; BREITENBACH, M.; GRUDIC, G. Z. Mixture of watson distributions: a generative model for hypersherical embeddings. *Artificial Inteligence and Statistics*, USA, p. 35 – 42, 2007.

[10] DE CASTRO, R. *El universo Latex*. Colombia: Universidade Nacional da Colombia, 2003. 470p.

[11] COOK, R. D. Detection of influential observations in linear regression. *Technometrics*, USA, v. 19, p. 15 – 18, 1977.

[12] COOK, R. D.; WEISBERG, S. *Residual and influence in Regression*. New York: Chapman & Hall, 1982. 230p.

[13] DIGGLE, P. J.; FISHER, N. I. A comparison of tests of uniformity for spherical data. *Austral J. Statist.*, AUSTRÁLIA, v. 27, p. 53 – 59, 1985.

[14] DHILLON, I. S.; MARCOTTE, E. M.; ROSHAN, U. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, SUIÇA, v. 19, p. 1612 – 1619, 2003.

[15] ERDÉLYI, A.; MAGNUS, W.; OBERHETTINGER, F.; TRICOMI, F. G. *Higher transcendental functions*. New York: McGraw Hill, 1953. 292p.

[16] FISHER, R. A. Dispersion on a sphere. *Proc. R. Soc.*, USA, v. 9, p. 295 – 305, 1953.

[17] FISHER, N. I.; HUNTINGTON, J. F.; JACKETT, D. R.; WILLCOX, M. E.; CREASEY, J. W. Spatial analysis or two-dimensional orientation data. *J. Math. Geol.*, GRÃ-BRETANHA, v. 17, p. 177 – 194, 1985. 329p.

[18] FISHER, N. I.; LEWIS, T.; EMBLETON, B. J. J. *Statistical analysis of spherical data*. Inglaterra: Cambridge University Press, 1987.

[19] GAUTSCHI, W. Anomalous convergence of a continued fraction for rations of kummer functions. *Mathematics of Computation*, USA, v. 31, p. 994 – 999, 1977.

[20] GIL, A.; SEGURA, J.; TEMME, N. M. *Numerical Methods for Special Functions*. Inglaterra: Cambridge University Press, 2007. 415p.

[21] GOODMAN, N. R. Satistical analysis based on a certain complex Gaussian distribution (an introduction). *The Analls of Mathematical Statistics*, USA, v. 34, p. 152 – 177, 1963.

[22] HAWKINS, D. M. *Identification of Outliers*. New York: Chapman & Hall, 1980. 188p.

[23] HOAGLIN, D. C.; IGLEWICZ, B.; TUKEY, J. W. Performance of some resistant rule for outliers labelling. *Journal of the American Statistical Association*, USA, v. 34, p. 991 – 999, 1986.

[24] KIM-HUNG, L.; CARL, K. W. Random sampling from the Watson Distribution. *Communications is Statistics*, INGLATERRA, v. 22, p. 997 – 1009, 1993.

[25] MARDIA, K. V.; JUPP, P. *Directional Statistics*. New York: John Wiley & Sons, 2000. 465p.

[26] SUVRIT, Sra.; KARP, D. The multivariate Watson distribution: Maximum-likelihood estimation and other aspects. *Journal of Multivariate Analysis*, USA, v. 114, p. 256 – 269, 2013.

[27] Tukey, W. J. *Exploraty Data Analysis*. USA: Addison-Wesley, 1970. 688p.

[28] Watson, G. S. Equatorial distributions on a sphere. *Biometrika*, USA, v. 52, p. 193 – 201, 1965.

APÊNDICE A

Programas

A.1 Taxa de deteccao

```
#################################################################################################
# Taxa de deteccao dos dados aleatorios de uma distribuicao watson para uma amostra n=20  #
#################################################################################################

rm(list=ls())
set.seed(1234)


#dados=matrix(NA,n1,2)
n1<-19
n<-n1+1

nrep = 1000


acerto1<-0
acerto2<-0
acerto3<-0
acerto4<-0


rotacao<-function(theta0,phi0,qsi0)
{

msaida<-matrix(0,3,3)

msaida[1,1]<-cos(theta0)*cos(phi0)*cos(qsi0)-sin(phi0)*sin(qsi0)
msaida[1,2]<-cos(theta0)*sin(phi0)*cos(qsi0)+cos(phi0)*sin(qsi0)
msaida[1,3]<-  -sin(theta0)*cos(qsi0)
msaida[2,1]<- -cos(theta0)*cos(phi0)*sin(qsi0)-sin(phi0)*cos(qsi0)
msaida[2,2]<- -cos(theta0)*sin(phi0)*sin(qsi0)+cos(phi0)*cos(qsi0)
msaida[2,3]<-  sin(theta0)*sin(qsi0)
msaida[3,1]<- sin(theta0)*cos(phi0)
msaida[3,2]<- sin(theta0)*sin(phi0)
msaida[3,3]<-cos(theta0)
```

```
return(msaida)
}


for(imc in 1:nrep)
{
##########################################
#                                        #
#Gera-se n-1 observacoes de uma Watson   #
#                                        #
##########################################
k=13
x = y = vector()
x=y=z=array(0,dim=c(n1,1))
u1=matrix(0,3,n1)
for(j in 1:n1)
{
rho=4*k/(2*k+3+sqrt((2*k+3)^2-16*k))
r=(3*rho/2*k)^3*exp(-3+(2*k)/rho)
bandeira=0

while(bandeira==0){
un=runif(3)
S=un[1]^2/(1-rho*(1-un[1]^2))
W=k*S
V=(r*un[2]^2)/(1-rho*S)^3
if(V>exp(2*W))
{
bandeira=0
}
else{
bandeira=1
}
}
theta=acos(sqrt(S))
if(un[3]<0.5){
theta = pi-theta
phi=4*pi*un[3]
}else{

phi=2*pi*(2*un[3]-1)
}
#print(theta)
#print(phi)

x[j] = sin(theta)*cos(phi)
y[j] = sin(theta)*sin(phi)
z[j] = cos(theta)

u1[,j] = c(x[j],y[j],z[j])# cada coluna é formada por cada vetor
#de 3 observacoes


} # aqui termina o for


##########################################
#                                        #
#Gera-se a n-esima observacao da Watson  #
#                                        #
##########################################

#set.seed(1234)
k=1
n2=1
#dados=matrix(NA,n2,2)
```

```
u2=matrix(0,3,n2)

rho=4*k/(2*k+3+sqrt((2*k+3)^2-16*k))
r=(3*rho/2*k)^3*exp(-3+(2*k)/rho)
bandeira=0

while(bandeira==0){
un=runif(3)
S=un[1]^2/(1-rho*(1-un[1]^2))
W=k*S
V=(r*un[2]^2)/(1-rho*S)^3
if(V>exp(2*W))
{
bandeira=0
}
else{
bandeira=1
}
}
theta=acos(sqrt(S))
if(un[3]<0.5){
theta = pi-theta
phi=4*pi*un[3]
}else{

phi=2*pi*(2*un[3]-1)
}
#print(theta)
#print(phi)

x = sin(theta)*cos(phi)
y = sin(theta)*sin(phi)
z = cos(theta)

u2 = c(x,y,z)# cada coluna e formada por cada vetor
#de 3 observacoes


#rotacao de u2

#u2=rotacao(90,100,0)%*%u2 #Bom resultado

u2=rotacao(90,100,0)%*%u2


######################################################################
#                                                                    #
# As (n-1) observacoes sao concatenadas com a n-esima observacao     #
#                                                                    #
######################################################################

u<-cbind(u1,u2)

#print("dados")
#print(u)


######################################################################
#             Determinando os valores dos k estimados                #
######################################################################


k_estimado = matrix(0,n,1)

k_estimado = function(a,c,autoval){
```

```
(c-a)/(1-autoval)+1-a+((a-1)*(a-c-1)*(1-autoval))/
(c-a)
}

k_estimado1<-function(a,c,r1)
{
(c*r1-a)/(r1*(1-r1))+r1/(2*c*(1-r1))
}

k_estimado2 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+(1-c/c-a))
}

k_estimado3 = function(a,c,r1){
(r1*c-a)/(2*r1*(1-r1))*(1+sqrt(1+(4*(c+1)*r1*(1-r1))/(a*(c-a))))
}

k_estimado4 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+r/a)
}


#################################################################
#       Determinando os autovalores e autovetores de S          #
#################################################################
soma<-array(0,dim=c(3,3))
for(j in 1:n){
soma = soma + u[,j]%*%t(u[,j])

}

S = (1/n)*soma
saida1 = eigen(S) #autovalores de S
saida1$vectors # autovetores
saida1$values # autovalores




#####################################################################
#Encontrando meu theta_chap (composto por 4 valores da matriz S)  #
#####################################################################
a = 1/2
c = 3/2
r1 =saida1$values[1]  ## pois r1 tende para 1 (maior autovalor de S)


k_matrizS<-k_estimado1(a,c,r1)
theta_chap = cbind(t(saida1$vectors[,1]),k_estimado1(a,c,r1))##armazena esses valores em
#uma coluna
t(theta_chap) ## transposto do vetor theta

#############################################################
#                                                           #
#Encontrando theta sem a i-esima observacao (theta_chap1)  #
#                                                           #
#############################################################
somai<-array(0,dim=c(3,3,n))

for(j in 1:n){
somai[,,j]=soma-u[,j]%*%t(u[,j])
}
somai = (1/(n-1))*somai ## fornece 20 matrizes 3 por 3


theta_chap1<-array(0,dim=c(n,4))
```

```
for(j in 1:(n-1))
{

saida = eigen(somai[,,j]) #autovalores de somai[,,1]
saida$vectors # autovetores
saida$values # autovalores
r1_autovec = saida$vectors[,1]
r1_autoval<-saida$values[1]

theta_chap1[j,] = cbind(t(r1_autovec),k_estimado1(a,c,r1_autoval)) # matriz 20

}


saida = eigen(somai[,,n]) #autovalores de somai[,,1]
saida$vectors # autovetores
saida$values # autovalores
r1_autovec = saida$vectors[,1]
r1_autoval<-saida$values[1]
#print("r1_autoval")
#print(r1_autoval)


theta_chap1[n,] = cbind(t(r1_autovec),k_estimado2(a,c,r1_autoval)) # matriz 20


#####################################################################
#              Valor de M(a,b,k_estimado1(a,c,r1))                  #
#####################################################################

mabk<-function(a,b,k_matrizS)
{
sum = 0
erro=1
i=0


while((erro > 0.01)&(i<90)){

sum1=sum
sum  = sum + (gamma(a+i)*gamma(b)*(k_matrizS^{i}))/
(gamma(a)*gamma(b+i)*factorial(i))
i=i+1

erro = abs(sum-sum1)

}

return(sum)
}

#################################
#                               #
#Valor numerico das constantes  #
#                               #
#################################
a<-0.5
b<-1.5
result_mabk<-mabk(1/2,3/2,k_matrizS)
result_ma1b1k<-mabk(3/2,5/2,k_matrizS)
result_ma2b2k<-mabk(5/2,7/2,k_matrizS)

#########################################################################
#Calculando a matriz de informacao de fisher observada (J(theta_chap))#
#########################################################################
```

```
IF = matrix(0,4,4) ## matriz de zeros 4 por 4
IF[1,1] = 2*(k_matrizS)*(S[1,1])
IF[1,2] = (k_matrizS)*(S[2,1]+S[1,2])
IF[1,3] = (k_matrizS)*(S[3,1]+S[1,3])
IF[1,4] = 2*(S[1,1])+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[2,1] = (k_matrizS)*(S[2,1]+S[1,2])
IF[2,2] = 2*(k_matrizS)*(S[2,2])
IF[2,3] = (k_matrizS)*(S[3,2]+S[2,3])
IF[2,4] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[3,1] = (k_matrizS)*(S[3,1]+S[1,3])
IF[3,2] = (k_matrizS)*(S[3,2]+S[2,3])
IF[3,3] =  2*(k_matrizS)*(S[3,3])
IF[3,4] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,1] = (2*(S[1,1]))+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[4,2] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[4,3] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,4] = ((a^{2})*(b+1)*((result_ma1b1k)^{2}))-((a)*(a+1)*(b)*(result_mabk)*(result_ma2b2k))/
((b^{2})*(b+1)*((result_mabk)^{2}))
#Inf_obs = solve(IF) # inversa da matriz de informacao de Fisher observada




######################################################################
#                   Calculando a distancia de Cook                   #
######################################################################

dcook=array(0,n,1)
for(j in 1:n){
dif_transp = theta_chap1[j,]-theta_chap
# diferenca entre theta_chap1 e theta_chap
#(vetor com menos uma observacao - vetor com todas as observacoes)
t(dif_transp) # transposto do vetor dif_transp
dcook[j] =abs((dif_transp)%*%(IF)%*%(t(dif_transp)))
#alocando os vetores de theta_chap1 da matriz S1 em uma nova matriz
}


########################
#                      #
#Distancia Geodesica   #
#                      #
########################
geodesica<-function(v1,v2)
{
return(acos(t(v1)%*%v2))
}

distancia<-array(0,dim=c(n,1))
tu<-array(0,dim=c(3,n))

for(i in 1:(n-1))
{
if(u[3,i]<0)
{
tu[,i]<- -u[,i]
}else{
tu[,i]<-u[,i]
}

distancia[i]<-min(geodesica(u[,i],saida1$vectors[,1]),geodesica(-u[,i],saida1$vectors[,1]))


#distancia[i]<-geodesica(tu[,i],saida1$vectors[,1])
}
```

```
tu[,n]<-u[,n]
distancia[n]<-geodesica(tu[,n],saida1$vectors[,1])


#########################################################################
#    1 Criterio: Ponto de corte para as distâncias                      #
#########################################################################
ddcook  = sort(dcook) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = ddcook[round(n+1-f)]
FL = ddcook[round(f)]


if(dcook[n]>FU+(4*(FU-FL)))
acerto1<-acerto1+1



#################################################################
#    2 Criterio: Teste outlier para discordancia             #
#################################################################
dados = array(0,dim=c(3,n))
dados1 = u # fornece os 20 vetores da matria u com todas as observacoes
soma_1 = 0
for(j in 1:n){
soma_1 = soma_1 + dados1[,j]%*%t(dados1[,j]) # fornece as 20 matrizes 3 por 3
# com todas as observacoes
}
# soma_1 = Tn # matriz de orientacao com todas as observacoes
soma_mat = array(0,dim=c(3,3,n))
for(i in 1:n){
dadost = dados1[,-i]
soma_mat1 = 0
for(j in 1:n-1){
soma_mat1 = soma_mat1 + dadost[,j]%*%t(dadost[,j])
}
soma_mat[,,i] = soma_mat1 # fornece as 20 matrizes 3 por 3 com n-1 observacoes

}


saida2 = eigen(soma_mat[,,i]) # autovalores de soma_mat[,,i]
saida2$vectors # autovetores
saida2$values # autovalores

saida2=vector("list", n) ## guardando os autovalores e autovetores de soma_mat
# em uma lista
for(j in 1:n){
saida2[[j]] = eigen(soma_mat[,,j]) #autovalores de soma_mat[,,j]

}


tau_chap = eigen(soma_1)$values[1]# autovalores de soma_1 # maior autovalor
#da matriz soma_1

tau_chap1 = matrix(0,n,1)
for(i in 1:n){
tau_chap1[i] = eigen(soma_mat[,,i])$values[1] # maior autovalor da matriz
#soma_mat
}


##########################################
#         Estatistica de teste(H)        #
##########################################
```

```
H = matrix(0,n,1)
for(i in 1:n){
H[i] = ((n-2)*(1+tau_chap1[i,]-tau_chap))/(n-1-tau_chap1[i,]) # estatistica de teste para
#detectar se o ponto seria um outlier (consideramos outlier se H for muito
#grande!)
}

SH  = sort(H) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = SH[round(n+1-f)]
FL = SH[round(f)]
if(H[n]>FU+(4*(FU-FL)))
acerto2<-acerto2+1


#######################################################
#   3 Criterio: Quantil de uma qui-quadrada          #
#######################################################

if(dcook[n]>qchisq(0.95,3))
acerto3<-acerto3+1

print(imc)


#######################################
#                                     #
# 4 Criterio: Distancia Geodesica     #
#                                     #
#######################################

Sd  = sort(distancia) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = Sd[round(n+1-f)]
FL = Sd[round(f)]

#||distancia[n]<FL-(1.5*(FU-FL))

if(distancia[n]>FU+(4*(FU-FL)))
{
acerto4=acerto4+1
}

} #fim do loop de geracao de amostras

print("acerto do metodo ponto de corte")
print(acerto1/nrep)
print("acerto teste de discordancia")
print(acerto2/nrep)
print("acerto com a quiquadrado")
print(acerto3/nrep)
print("acerto com a distancia geodesica")
print(acerto4/nrep)

resultadofinal = rbind(acerto1/nrep,acerto2/nrep,acerto3/nrep,acerto4/nrep)
write.table(resultadofinal,file='taxaacertoC4.txt',col.names=TRUE)
```

A.2 Taxa de erro

```
#############################################################################################
# Taxa de erro dos dados aleatorios de uma distribuicao watson para uma amostra n=20  #
```

```
##############################################################################################

rm(list=ls())
set.seed(1234)
#dados=matrix(NA,n1,2)
n1<-20
n<-n1

nrep = 1000

erro1<-0
erro2<-0
erro3<-0
erro4<-0

rotacao<-function(theta0,phi0,qsi0)
{

msaida<-matrix(0,3,3)

msaida[1,1]<-cos(theta0)*cos(phi0)*cos(qsi0)-sin(phi0)*sin(qsi0)
msaida[1,2]<-cos(theta0)*sin(phi0)*cos(qsi0)+cos(phi0)*sin(qsi0)
msaida[1,3]<-  -sin(theta0)*cos(qsi0)
msaida[2,1]<- -cos(theta0)*cos(phi0)*sin(qsi0)-sin(phi0)*cos(qsi0)
msaida[2,2]<- -cos(theta0)*sin(phi0)*sin(qsi0)+cos(phi0)*cos(qsi0)
msaida[2,3]<-  sin(theta0)*sin(qsi0)
msaida[3,1]<- sin(theta0)*cos(phi0)
msaida[3,2]<- sin(theta0)*sin(phi0)
msaida[3,3]<-cos(theta0)

return(msaida)
}

for(imc in 1:nrep)
{




##############################################
#                                            #
#Gera-se n-1 observacoes de uma Watson       #
#                                            #
##############################################
k=0.01
x = y = vector()
x=y=z=array(0,dim=c(n1,1))
u1=matrix(0,3,n1)
for(j in 1:n1)
{
rho=4*k/(2*k+3+sqrt((2*k+3)^2-16*k))
r=(3*rho/2*k)^3*exp(-3+(2*k)/rho)
bandeira=0

while(bandeira==0){
un=runif(3)
S=un[1]^2/(1-rho*(1-un[1]^2))
W=k*S
V=(r*un[2]^2)/(1-rho*S)^3
if(V>exp(2*W))
{
bandeira=0
}
else{
bandeira=1
}
```

```
}
theta=acos(sqrt(S))
if(un[3]<0.5){
theta = pi-theta
phi=4*pi*un[3]
}else{

phi=2*pi*(2*un[3]-1)
}


x[j] = sin(theta)*cos(phi)
y[j] = sin(theta)*sin(phi)
z[j] = cos(theta)

u1[,j] = c(x[j],y[j],z[j])# cada coluna Ã© formada por cada vetor
#de 3 observacoes


} # aqui termina o for



u<-u1

#########################################################################
#             Determinando os valores dos k estimados                   #
#########################################################################


k_estimado = matrix(0,n,1)

k_estimado = function(a,c,autoval){
(c-a)/(1-autoval)+1-a+((a-1)*(a-c-1)*(1-autoval))/
(c-a)
}

k_estimado1<-function(a,c,r1)
{
(c*r1-a)/(r1*(1-r1))+r1/(2*c*(1-r1))
}

k_estimado2 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+(1-c/c-a))
}

k_estimado3 = function(a,c,r1){
(r1*c-a)/(2*r1*(1-r1))*(1+sqrt(1+(4*(c+1)*r1*(1-r1))/(a*(c-a))))
}

k_estimado4 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+r/a)
}


#####################################################################
#       Determinando os autovalores e autovetores de S              #
#####################################################################
soma<-array(0,dim=c(3,3))
for(j in 1:n){
soma = soma + u[,j]%*%t(u[,j])

}

S = (1/n)*soma
saida1 = eigen(S) #autovalores de S
saida1$vectors # autovetores
```

```
saida1$values # autovalores



#####################################################################
#Encontrando meu theta_chap (composto por 4 valores da matriz S)  #
#####################################################################
a = 1/2
c = 3/2
r1 =saida1$values[1]  ## pois r1 tende para 1 (maior autovalor de S)


k_matrizS<-k_estimado1(a,c,r1)
theta_chap = cbind(t(saida1$vectors[,1]),k_estimado1(a,c,r1))##armazena esses valores em
#uma coluna
t(theta_chap) ## transposto do vetor theta

#print("K")
#print(k)
#print("K estimado")
#print(k_matrizS)


#################################################################
#                                                               #
#Encontrando theta sem a i-esima observacao (theta_chap1)   #
#                                                               #
#################################################################
somai<-array(0,dim=c(3,3,n))

for(j in 1:n){
somai[,,j]=soma-u[,j]%*%t(u[,j])
}
somai = (1/(n-1))*somai ## fornece 20 matrizes 3 por 3


theta_chap1<-array(0,dim=c(n,4))

for(j in 1:n)
{

saida = eigen(somai[,,j]) #autovalores de somai[,,1]
saida$vectors # autovetores
saida$values # autovalores
r1_autovec = saida$vectors[,1]
r1_autoval<-saida$values[1]



theta_chap1[j,] = cbind(t(r1_autovec),k_estimado1(a,c,r1_autoval)) # matriz 20

}


#########################################################################
#              Valor de M(a,b,k_estimado1(a,c,r1))                #
#########################################################################

mabk<-function(a,b,k_matrizS)
{
sum = 0
erro=1
i=0

while((erro > 0.01)&(i<90)){
```

```
sum1=sum
sum  = sum + (gamma(a+i)*gamma(b)*(k_matrizS^{i}))/
(gamma(a)*gamma(b+i)*factorial(i))
i=i+1
erro = abs(sum-sum1)

}

return(sum)
}

###################################
#                                 #
#Valor numerico das constantes    #
#                                 #
###################################
a<-0.5
b<-1.5
result_mabk<-mabk(1/2,3/2,k_matrizS)
result_ma1b1k<-mabk(3/2,5/2,k_matrizS)
result_ma2b2k<-mabk(5/2,7/2,k_matrizS)

#result_ma1b1k<-ma1b1k(3/2,5/2,k_matrizS)
#result_ma2b2k<-ma2b2k(5/2,7/2,k_matrizS)


#########################################################################
#Calculando a matriz de informacao de fisher observada (J(theta_chap))#
#########################################################################


IF = matrix(0,4,4) ## matriz de zeros 4 por 4
IF[1,1] = 2*(k_matrizS)*(S[1,1])
IF[1,2] = (k_matrizS)*(S[2,1]+S[1,2])
IF[1,3] = (k_matrizS)*(S[3,1]+S[1,3])
IF[1,4] = 2*(S[1,1])+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[2,1] = (k_matrizS)*(S[2,1]+S[1,2])
IF[2,2] = 2*(k_matrizS)*(S[2,2])
IF[2,3] = (k_matrizS)*(S[3,2]+S[2,3])
IF[2,4] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[3,1] = (k_matrizS)*(S[3,1]+S[1,3])
IF[3,2] = (k_matrizS)*(S[3,2]+S[2,3])
IF[3,3] =  2*(k_matrizS)*(S[3,3])
IF[3,4] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,1] = (2*(S[1,1]))+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[4,2] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[4,3] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,4] = ((a^{2})*(b+1)*((result_ma1b1k)^{2}))-((a)*(a+1)*(b)*(result_mabk)*(result_ma2b2k))/
((b^{2})*(b+1)*((result_mabk)^{2}))
#Inf_obs = solve(IF) # inversa da matriz de informacao de Fisher observada




#########################################################################
#                   Calculando a distancia de Cook                      #
#########################################################################

dcook=array(0,n,1)
for(j in 1:n){
dif_transp = theta_chap1[j,]-theta_chap
# diferenca entre theta_chap1 e theta_chap
#(vetor com menos uma observacao - vetor com todas as observacoes)

t(dif_transp) # transposto do vetor dif_transp
```

```
dcook[j] = (dif_transp)%*%(IF)%*%(t(dif_transp))## alocando os vetores de theta_chap1 da matriz S1 em uma nova matriz
}


#########################
#                       #
#Distancia Geodesica    #
#                       #
#########################
geodesica<-function(v1,v2)
{
return(acos(t(v1)%*%v2))
}

distancia<-array(0,dim=c(n,1))
tu<-array(0,dim=c(3,n))

for(i in 1:n)
{
if(u[3,i]<0)
{
tu[,i]<- -u[,i]
}else{
tu[,i]<-u[,i]
}


#distancia[i]<-geodesica(tu[,i],saida1$vectors[,1])
distancia[i]<-min(geodesica(u[,i],saida1$vectors[,1]),
geodesica(-u[,i],saida1$vectors[,1]))
}



#########################################################################
#    1 Criterio: Ponto de corte para as distancias                      #
#########################################################################
ddcook  = sort(dcook) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = ddcook[round(n+1-f)]
FL = ddcook[round(f)]

for(j in 1:n)
{
if(dcook[j]>FU+(6*(FU-FL)))
{
erro1<-erro1+1
break
}
}


#########################################################################
#    2 Criterio: Teste outlier para discordancia                        #
#########################################################################

dados = array(0,dim=c(3,n))
dados1 = u # fornece os 20 vetores da matria u com todas as observacoes
soma_1 = 0
for(j in 1:n){
soma_1 = soma_1 + dados1[,j]%*%t(dados1[,j]) # fornece as 20 matrizes 3 por 3
# com todas as observacoes
}
# soma_1 = Tn # matriz de orientacao com todas as observacoes
soma_mat = array(0,dim=c(3,3,n))
for(i in 1:n){
```

```
dadost = dados1[,-i]
soma_mat1 = 0
for(j in 1:n-1){
soma_mat1 = soma_mat1 + dadost[,j]%*%t(dadost[,j])
}
soma_mat[,,i] = soma_mat1 # fornece as 20 matrizes 3 por 3 com n-1 observacoes

}


saida2 = eigen(soma_mat[,,i]) # autovalores de soma_mat[,,i]
saida2$vectors # autovetores
saida2$values # autovalores

saida2=vector("list", n) ## guardando os autovalores e autovetores de soma_mat
# em uma lista
for(j in 1:n){
saida2[[j]] = eigen(soma_mat[,,j]) #autovalores de soma_mat[,,j]

}


tau_chap = eigen(soma_1)$values[1]# autovalores de soma_1 # maior autovalor
#da matriz soma_1

tau_chap1 = matrix(0,n,1)
for(i in 1:n){
tau_chap1[i] = eigen(soma_mat[,,i])$values[1] # maior autovalor da matriz
#soma_mat
}


##########################################
#         Estatistica de teste(H)        #
##########################################

H = matrix(0,n,1)
for(i in 1:n){
H[i] = ((n-2)*(1+tau_chap1[i,]-tau_chap))/(n-1-tau_chap1[i,]) # estatistica de teste para
#detectar se o ponto seria um outlier (consideramos outlier se H for muito
#grande!)
}

SH  = sort(H) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = SH[round(n+1-f)]
FL = SH[round(f)]
for(j in 1:n)
{
if(H[j]>FU+(6*(FU-FL)))
{
erro2<-erro2+1
break
}
}

#########################################################################
#   3 Criterio: Quantil de uma qui-quadrada                             #
#########################################################################

for(j in 1: n)
{
if(dcook[j]>qchisq(0.95,3))
{
erro3<-erro3+1
break
}
```

```
}

print(imc)
#print(dcook)


#######################################
#                                     #
#  4 Criterio: Distancia Geodesica    #
#                                     #
#######################################

Sd  = sort(distancia) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = Sd[round(n+1-f)]
FL = Sd[round(f)]
#||distancia[j]<FL-(6*(FU-FL))
for(j in 1:n)
{
if(distancia[j]>FU+(6*(FU-FL)))
{
erro4<-erro4+1
break
}
}


} #fim do loop de geracao de amostras

print("erro do metodo ponto de corte")
print(erro1/nrep)
print("erro teste de discordancia")
print(erro2/nrep)
print("erro com a quiquadrado")
print(erro3/nrep)
print("erro com a distancia geodesica")
print(erro4/nrep)

resultadofinal = rbind(erro1/nrep,erro2/nrep,erro3/nrep,erro4/nrep)
write.table(resultadofinal,file='taxaerroC6.txt',col.names=TRUE)
```

A.3 Dados Reais: 50 observacoes

```
#############################################################################
#     Dados Reais  de uma distribuicao watson para uma amostra n = 50       #
#############################################################################

rm(list=ls())
set.seed(1234)
#k=1
n=50
dados=matrix(NA,n,2)


theta = c(-26.4, -32.2, -73.1, -80.2, -71.1, -58.7, -40.8, -14.9, -66.1, -1.8,
-52.1, -77.3, -68.8, -68.4, -29.2, -78.5, -65.4, -49.0, -67.0, -56.7, -80.5,
-77.7, -6.9, -59.4, -5.6, -62.6, -74.7, -65.3, -71.6, -23.3, -74.3, -81.0,
-12.7, -75.4, -85.9, -84.8, -7.4, -29.8, -85.2, -53.1, -38.3, -72.7, -60.2,
-63.4, -17.2, -81.6, -40.4, -53.6, -56.2, -75.1)

phi = c(324.0, 163.7, 51.9, 140.5, 267.2, 32.0, 28.1, 266.3, 144.3, 256.2,
83.2, 182.1, 110.4, 142.2, 246.3, 222.6, 247.7, 65.6, 282.6, 56.2, 108.4,
```

```
266.0, 19.1, 281.7, 107.4, 105.3, 120.2, 286.6, 106.4, 96.5, 90.2, 170.9,
199.4, 118.6, 63.7, 74.9, 93.8, 72.8, 113.2, 51.5, 146.8, 103.1, 33.2,
154.8, 89.9, 295.6, 41.0, 59.1, 35.6, 70.7)

x = y = vector()
x=y=z=array(0,dim=c(n,1))
u1=matrix(0,3,n)
for(i in 1:n)
{

x[i] = sin(theta[i])*cos(phi[i])
y[i] = sin(theta[i])*sin(phi[i])
z[i] = cos(theta[i])

u1[,i] = c(x[i],y[i],z[i])# cada coluna é formada por cada vetor

} # aqui termina o for
u<-u1




#####################################################################
#             Determinando os valores dos k estimados               #
#####################################################################

k_estimado = matrix(0,n,1)

k_estimado = function(a,c,autoval){
(c-a)/(1-autoval)+1-a+((a-1)*(a-c-1)*(1-autoval))/
(c-a)
}

k_estimado1<-function(a,c,r1)
{
(c*r1-a)/(r1*(1-r1))+r1/(2*c*(1-r1))
}

k_estimado2 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+(1-c/c-a))
}

k_estimado3 = function(a,c,r1){
(r1*c-a)/(2*r1*(1-r1))*(1+sqrt(1+(4*(c+1)*r1*(1-r1))/(a*(c-a))))
}

k_estimado4 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+r/a)
}


#################################################################
#       Determinando os autovalores e autovetores de S          #
#################################################################
soma<-array(0,dim=c(3,3))
for(j in 1:n){
soma = soma + u[,j]%*%t(u[,j])

}

S = (1/n)*soma
saida1 = eigen(S) #autovalores de S
saida1$vectors # autovetores
saida1$values # autovalores
```

```
#####################################################################
#Encontrando meu theta_chap (composto por 4 valores da matriz S)  #
#####################################################################
a = 1/2
c = 3/2
r1 =saida1$values[1]  ## pois r1 tende para 1 (maior autovalor de S)


k_matrizS<-k_estimado1(a,c,r1)

k<-k_matrizS

theta_chap = cbind(t(saida1$vectors[,1]),k_estimado1(a,c,r1))##armazena esses valores em
#uma coluna
t(theta_chap) ## transposto do vetor theta

#print("K")
#print(k)
#print("K estimado")
#print(k_matrizS)







#################################################################
#                                                               #
#Encontrando theta sem a i-esima observacao (theta_chap1)   #
#                                                               #
#################################################################
somai<-array(0,dim=c(3,3,n))

for(j in 1:n){
somai[,,j]=soma-u[,j]%*%t(u[,j])
}
somai = (1/(n-1))*somai ## fornece 20 matrizes 3 por 3


theta_chap1<-array(0,dim=c(n,4))

for(j in 1:n)
{

saida = eigen(somai[,,j]) #autovalores de somai[,,1]
saida$vectors # autovetores
saida$values # autovalores
r1_autovec = saida$vectors[,1]
r1_autoval<-saida$values[1]



theta_chap1[j,] = cbind(t(r1_autovec),k_estimado1(a,c,r1_autoval)) # matriz 20

}


#####################################################################
#              Valor de M(a,b,k_estimado1(a,c,r1))                #
#####################################################################

mabk<-function(a,b,k_matrizS)
{
sum = 0
erro=1
i=0
```

```
while((erro > 0.01)&(i<90)){

sum1=sum
sum  = sum + (gamma(a+i)*gamma(b)*(k_matrizS^{i}))/
(gamma(a)*gamma(b+i)*factorial(i))
i=i+1
erro = abs(sum-sum1)

}

return(sum)
}
```

```
###################################
#                                 #
#Valor numerico das constantes    #
#                                 #
###################################
a<-0.5
b<-1.5
result_mabk<-mabk(1/2,3/2,k_matrizS)
result_ma1b1k<-mabk(3/2,5/2,k_matrizS)
result_ma2b2k<-mabk(5/2,7/2,k_matrizS)

#result_ma1b1k<-ma1b1k(3/2,5/2,k_matrizS)
#result_ma2b2k<-ma2b2k(5/2,7/2,k_matrizS)


#########################################################################
#Calculando a matriz de informacao de fisher observada (J(theta_chap))#
#########################################################################


IF = matrix(0,4,4) ## matriz de zeros 4 por 4
IF[1,1] = 2*(k_matrizS)*(S[1,1])
IF[1,2] = (k_matrizS)*(S[2,1]+S[1,2])
IF[1,3] = (k_matrizS)*(S[3,1]+S[1,3])
IF[1,4] = 2*(S[1,1])+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[2,1] = (k_matrizS)*(S[2,1]+S[1,2])
IF[2,2] = 2*(k_matrizS)*(S[2,2])
IF[2,3] = (k_matrizS)*(S[3,2]+S[2,3])
IF[2,4] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[3,1] = (k_matrizS)*(S[3,1]+S[1,3])
IF[3,2] = (k_matrizS)*(S[3,2]+S[2,3])
IF[3,3] =  2*(k_matrizS)*(S[3,3])
IF[3,4] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,1] = (2*(S[1,1]))+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[4,2] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[4,3] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,4] = ((a^{2})*(b+1)*((result_ma1b1k)^{2}))-((a)*(a+1)*(b)*(result_mabk)*(result_ma2b2k))/
((b^{2})*(b+1)*((result_mabk)^{2}))
#Inf_obs = solve(IF) # inversa da matriz de informacao de Fisher observada
```

```
#####################################################################
#                  Calculando a distancia de Cook                   #
#####################################################################
```

```
dcook=array(0,n,1)
for(j in 1:n){
dif_transp = theta_chap1[j,]-theta_chap
# diferenca entre theta_chap1 e theta_chap
#(vetor com menos uma observacao - vetor com todas as observacoes)

t(dif_transp) # transposto do vetor dif_transp

dcook[j] = (dif_transp)%*%(IF)%*%(t(dif_transp))## alocando os vetores de theta_chap1 da matriz S1 em uma nova matriz


}




########################
#                      #
#Distancia Geodesica   #
#                      #
########################
geodesica<-function(v1,v2)
{
return(acos(t(v1)%*%v2))
}

distancia<-array(0,dim=c(n,1))
tu<-array(0,dim=c(3,n))

for(i in 1:n)
{
if(u[3,i]<0)
{
tu[,i]<- -u[,i]
}else{
tu[,i]<-u[,i]
}


#distancia[i]<-geodesica(tu[,i],saida1$vectors[,1])
distancia[i]<-min(geodesica(u[,i],saida1$vectors[,1]),geodesica(-u[,i],saida1$vectors[,1]))
}

#########################################################################
#    Criterios para compararar as distancias                            #
#########################################################################

#########################################################################
#    1 Criterio: Ponto de corte para as distâncias                      #
#########################################################################


ddcook  = sort(dcook) # reordena as distancias em ordem crescente
sucesso1 = matrix(0,n,1)
n = 50
f = (1/2)*(trunc((n+3)/2))
FUdcook = ddcook[round(n+1-f)]
FLdcook = ddcook[round(f)]
for(i in 1:n){
if(dcook[i]>FUdcook+(3*(FUdcook-FLdcook)))## variar o valor de k = 1.5, 3,
#4, e 6
{
sucesso1[i] = 1
}
```

```
else{
sucesso1[i] = 0
}
}


########################################################################
#    2 Criterio: Teste outlier para discordancia                       #
########################################################################

dados = array(0,dim=c(3,n))
dados1 = u1 # fornece os 50 vetores da matria u com todas as observacoes
soma_1 = 0
for(j in 1:n){
soma_1 = soma_1 + dados1[,j]%*%t(dados1[,j]) # fornece as 50 matrizes 3 por 3
# com todas as observacoes
}
# soma_1 = Tn # matriz de orientacao com todas as observacoes
soma_mat = array(0,dim=c(3,3,n))
for(i in 1:n){
dadost = dados1[,-i]
soma_mat1 = 0
for(j in 1:n-1){
soma_mat1 = soma_mat1 + dadost[,j]%*%t(dadost[,j])
}
soma_mat[,,i] = soma_mat1 # fornece as 50 matrizes 3 por 3 com n-1 observacoes

}


saida2 = eigen(soma_mat[,,i]) # autovalores de soma_mat[,,i]
saida2$vectors # autovetores
saida2$values # autovalores

saida2=vector("list", n) ## guardando os autovalores e autovetores de soma_mat
# em uma lista
for(k in 1:n){
saida2[[k]] = eigen(soma_mat[,,k]) #autovalores de soma_mat[,,k]

}

tau_chap = eigen(soma_1)$values[1]# autovalores de soma_1 # maior autovalor
#da matriz soma_1

tau_chap1 = matrix(0,n,1)
for(i in 1:n){
tau_chap1[i] = eigen(soma_mat[,,i])$values[1] # maior autovalor da matriz
#soma_mat
}


##########################################
#         Estatistica de teste(H)        #
##########################################

H = matrix(0,n,1)
for(i in 1:n){
H[i] = ((n-2)*(1+tau_chap1[i,]-tau_chap))/(n-1-tau_chap1[i,]) # estatistica de teste para
#detectar se o ponto seria um outlier (consideramos outlier se H for muito
#grande!)
}


sucesso2 = matrix(0,n,1)
SH  = sort(H) # reordena as distancias em ordem crescente
n = 50
f = (1/2)*(trunc((n+3)/2))
```

```
FUH = SH[round(n+1-f)]
FLH = SH[round(f)]
for(j in 1:n){
if(H[j]>FUH+(3*(FUH-FLH)))## variar o valor de k = 1.5, 3, 4, e 6
{
sucesso2[j]<-1
}
else{
sucesso2[j]<-0
}
}
#plot(H) # grafico de H




#########################################################################
#    3 Criterio: Quantil de uma qui-quadrada                            #
#########################################################################

sucesso3 = matrix(0,n,1)
teste_qq = qchisq(0.95,3)
for(i in 1:n){
if(dcook[i]>teste_qq){
sucesso3[i] = 1
}
else {
sucesso3[i] = 0
}
}

########################################
#                                      #
# 4 Criterio: Distancia Geodesica      #
#                                      #
########################################

sucesso4 = matrix(0,n,1)
Sd  = sort(distancia) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = Sd[round(n+1-f)]
FL = Sd[round(f)]
for(j in 1:n)
{
if(distancia[j]>FU+(3*(FU-FL)))
{
sucesso4[i] = 1
}
else{
sucesso4[i] = 0
}
}

##############
#            #
#Resultado   #
#            #
##############

Saida<-cbind(dcook,H,distancia)
print(Saida)

print("dcook com ponto de corte")
print(FUdcook+(3*(FUdcook-FLdcook)))
print("Criterio H - Outlier Discordancia")
```

```
print(FUH+(3*(FUH-FLH)))
print("Criterio Dcook")
print(teste_qq)
print("Criterio Geodesica")
print(FU+(3*(FU-FL)))
```

A.4 Dados Reais: 72 observacoes

```
##############################################################################
#     Dados Reais  de uma distribuicao watson para uma amostra n = 72        #
##############################################################################

rm(list=ls())
set.seed(1234)
#k=1
n=72
dados=matrix(NA,n,2)

dip = theta = c(80, 72, 63, 51, 62, 53, 53, 52, 48, 45, 44, 44, 34, 37, 38,
40, 25, 15, 22, 63, 35, 28, 28, 22, 33, 37, 32, 27, 24, 8, 6, 8, 11, 8, 6,
6, 8, 20, 21, 18, 25, 28, 32, 32, 32, 34, 38, 37, 44, 45, 48, 42, 47, 45,
43, 45, 50, 70, 59, 66, 65, 70, 66, 67, 83, 66, 69, 69, 72, 67, 69, 82)

dip_dir = phi = c(122, 132, 141, 145, 128, 133, 130, 129, 124, 120, 137, 141,
151, 138, 135, 135, 156, 156, 130, 112, 116, 113, 117, 110, 106, 106, 98, 84,
77, 111, 122, 140, 48, 279, 19, 28, 28, 310, 310, 331, 326, 332, 3, 324, 308,
304, 304, 299, 293, 293, 306, 310, 313, 319, 320, 320, 330, 327, 312, 317,
314, 312, 311, 307, 311, 310, 310, 305, 305, 301, 301, 300)

x = y = vector()
x=y=z=array(0,dim=c(n,1))
u1=matrix(0,3,n)
for(i in 1:n)
{

x[i] = sin(theta[i])*cos(phi[i])
y[i] = sin(theta[i])*sin(phi[i])
z[i] = cos(theta[i])

u1[,i] = c(x[i],y[i],z[i])# cada coluna é formada por cada vetor

} # aqui termina o for
u<-u1

#####################################################################
#             Determinando os valores dos k estimados               #
#####################################################################

k_estimado = matrix(0,n,1)

k_estimado = function(a,c,autoval){
(c-a)/(1-autoval)+1-a+((a-1)*(a-c-1)*(1-autoval))/
(c-a)
}

k_estimado1<-function(a,c,r1)
{
(c*r1-a)/(r1*(1-r1))+r1/(2*c*(1-r1))
}

k_estimado2 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+(1-c/c-a))
```

```
}

k_estimado3 = function(a,c,r1){
(r1*c-a)/(2*r1*(1-r1))*(1+sqrt(1+(4*(c+1)*r1*(1-r1))/(a*(c-a))))
}

k_estimado4 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+r/a)
}


#####################################################################
#       Determinando os autovalores e autovetores de S              #
#####################################################################
soma<-array(0,dim=c(3,3))
for(j in 1:n){
soma = soma + u[,j]%*%t(u[,j])

}

S = (1/n)*soma
saida1 = eigen(S) #autovalores de S
saida1$vectors # autovetores
saida1$values # autovalores







#########################################################################
#Encontrando meu theta_chap (composto por 4 valores da matriz S)  #
#########################################################################
a = 1/2
c = 3/2
r1 =saida1$values[1]  ## pois r1 tende para 1 (maior autovalor de S)


k_matrizS<-k_estimado1(a,c,r1)

k<-k_matrizS

theta_chap = cbind(t(saida1$vectors[,1]),k_estimado1(a,c,r1))##armazena esses valores em
#uma coluna
t(theta_chap) ## transposto do vetor theta

#print("K")
#print(k)
#print("K estimado")
#print(k_matrizS)


#################################################################
#                                                               #
#Encontrando theta sem a i-esima observacao (theta_chap1)   #
#                                                               #
#################################################################
somai<-array(0,dim=c(3,3,n))

for(j in 1:n){
somai[,,j]=soma-u[,j]%*%t(u[,j])
}
somai = (1/(n-1))*somai ## fornece 20 matrizes 3 por 3


theta_chap1<-array(0,dim=c(n,4))
```

```
for(j in 1:n)
{

saida = eigen(somai[,,j]) #autovalores de somai[,,1]
saida$vectors # autovetores
saida$values # autovalores
r1_autovec = saida$vectors[,1]
r1_autoval<-saida$values[1]



theta_chap1[j,] = cbind(t(r1_autovec),k_estimado1(a,c,r1_autoval)) # matriz 20

}


#####################################################################
#              Valor de M(a,b,k_estimado1(a,c,r1))                  #
#####################################################################

mabk<-function(a,b,k_matrizS)
{
sum = 0
erro=1
i=0

while((erro > 0.01)&(i<90)){

sum1=sum
sum  = sum + (gamma(a+i)*gamma(b)*(k_matrizS^{i}))/
(gamma(a)*gamma(b+i)*factorial(i))
i=i+1
erro = abs(sum-sum1)

}

return(sum)
}

###################################
#                                 #
#Valor numerico das constantes    #
#                                 #
###################################
a<-0.5
b<-1.5
result_mabk<-mabk(1/2,3/2,k_matrizS)
result_ma1b1k<-mabk(3/2,5/2,k_matrizS)
result_ma2b2k<-mabk(5/2,7/2,k_matrizS)

#result_ma1b1k<-ma1b1k(3/2,5/2,k_matrizS)
#result_ma2b2k<-ma2b2k(5/2,7/2,k_matrizS)


#####################################################################
#Calculando a matriz de informacao de fisher observada (J(theta_chap))#
#####################################################################


IF = matrix(0,4,4) ## matriz de zeros 4 por 4
IF[1,1] = 2*(k_matrizS)*(S[1,1])
IF[1,2] = (k_matrizS)*(S[2,1]+S[1,2])
IF[1,3] = (k_matrizS)*(S[3,1]+S[1,3])
IF[1,4] = 2*(S[1,1])+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[2,1] = (k_matrizS)*(S[2,1]+S[1,2])
IF[2,2] = 2*(k_matrizS)*(S[2,2])
```

```
IF[2,3] = (k_matrizS)*(S[3,2]+S[2,3])
IF[2,4] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[3,1] = (k_matrizS)*(S[3,1]+S[1,3])
IF[3,2] = (k_matrizS)*(S[3,2]+S[2,3])
IF[3,3] =  2*(k_matrizS)*(S[3,3])
IF[3,4] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,1] = (2*(S[1,1]))+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[4,2] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[4,3] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,4] = ((a^{2})*(b+1)*((result_ma1b1k)^{2}))-((a)*(a+1)*(b)*(result_mabk)*(result_ma2b2k))/
((b^{2})*(b+1)*((result_mabk)^{2}))
#Inf_obs = solve(IF) # inversa da matriz de informacao de Fisher observada




######################################################################
#                   Calculando a distancia de Cook                   #
######################################################################

dcook=array(0,n,1)
for(j in 1:n){
dif_transp = theta_chap1[j,]-theta_chap
# diferenca entre theta_chap1 e theta_chap
#(vetor com menos uma observacao - vetor com todas as observacoes)

t(dif_transp) # transposto do vetor dif_transp

dcook[j] = (dif_transp)%*%(IF)%*%(t(dif_transp))## alocando os vetores de theta_chap1 da matriz S1 em uma nova matriz

}

########################
#                      #
#Distancia Geodesica   #
#                      #
########################
geodesica<-function(v1,v2)
{
return(acos(t(v1)%*%v2))
}

distancia<-array(0,dim=c(n,1))
tu<-array(0,dim=c(3,n))

for(i in 1:n)
{
if(u[3,i]<0)
{
tu[,i]<- -u[,i]
}else{
tu[,i]<-u[,i]
}

#distancia[i]<-geodesica(tu[,i],saida1$vectors[,1])
distancia[i]<-min(geodesica(u[,i],saida1$vectors[,1]),geodesica(-u[,i],saida1$vectors[,1]))
}

#########################################################################
#    Criterios para compararar as distancias                            #
#########################################################################

#####################################################################
#    1 Criterio: Ponto de corte para as distâncias                  #
#####################################################################
```

```
ddcook  = sort(dcook) # reordena as distancias em ordem crescente
sucesso1 = matrix(0,n,1)
n = 72
f = (1/2)*(trunc((n+3)/2))
FUdcook = ddcook[round(n+1-f)]
FLdcook = ddcook[round(f)]
for(i in 1:n){
if(dcook[i]>FUdcook+(3*(FUdcook-FLdcook)))## variar o valor de k = 1.5, 3,
#4, e 6
{
sucesso1[i] = 1
}
else{
sucesso1[i] = 0
}
}


#########################################################################
#    2 Criterio: Teste outlier para discordancia                        #
#########################################################################

dados = array(0,dim=c(3,n))
dados1 = u1 # fornece os 50 vetores da matria u com todas as observacoes
soma_1 = 0
for(j in 1:n){
soma_1 = soma_1 + dados1[,j]%*%t(dados1[,j]) # fornece as 50 matrizes 3 por 3
# com todas as observacoes
}
# soma_1 = Tn # matriz de orientacao com todas as observacoes
soma_mat = array(0,dim=c(3,3,n))
for(i in 1:n){
dadost = dados1[,-i]
soma_mat1 = 0
for(j in 1:n-1){
soma_mat1 = soma_mat1 + dadost[,j]%*%t(dadost[,j])
}
soma_mat[,,i] = soma_mat1 # fornece as 50 matrizes 3 por 3 com n-1 observacoes

}


saida2 = eigen(soma_mat[,,i]) # autovalores de soma_mat[,,i]
saida2$vectors # autovetores
saida2$values # autovalores

saida2=vector("list", n) ## guardando os autovalores e autovetores de soma_mat
# em uma lista
for(k in 1:n){
saida2[[k]] = eigen(soma_mat[,,k]) #autovalores de soma_mat[,,k]

}

tau_chap = eigen(soma_1)$values[1]# autovalores de soma_1 # maior autovalor
#da matriz soma_1

tau_chap1 = matrix(0,n,1)
for(i in 1:n){
tau_chap1[i] = eigen(soma_mat[,,i])$values[1] # maior autovalor da matriz
#soma_mat
}


##########################################
#         Estatistica de teste(H)        #
##########################################
```

```
H = matrix(0,n,1)
for(i in 1:n){
H[i] = ((n-2)*(1+tau_chap1[i,]-tau_chap))/(n-1-tau_chap1[i,]) # estatistica de teste para
#detectar se o ponto seria um outlier (consideramos outlier se H for muito
#grande!)
}


sucesso2 = matrix(0,n,1)
SH  = sort(H) # reordena as distancias em ordem crescente
n = 72
f = (1/2)*(trunc((n+3)/2))
FUH = SH[round(n+1-f)]
FLH = SH[round(f)]
for(j in 1:n){
if(H[j]>FUH+(3*(FUH-FLH)))## variar o valor de k = 1.5, 3, 4, e 6
{
sucesso2[j]<-1
}
else{
sucesso2[j]<-0
}
}
#plot(H) # grafico de H


#########################################################################
#    3 Criterio: Quantil de uma qui-quadrada                            #
#########################################################################

sucesso3 = matrix(0,n,1)
teste_qq = qchisq(0.95,3)
for(i in 1:n){
if(dcook[i]>teste_qq){
sucesso3[i] = 1
}
else {
sucesso3[i] = 0
}
}







########################################
#                                      #
# 4 Criterio: Distancia Geodesica      #
#                                      #
########################################

sucesso4 = matrix(0,n,1)
Sd  = sort(distancia) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = Sd[round(n+1-f)]
FL = Sd[round(f)]
for(j in 1:n)
{
if(distancia[j]>FU+(3*(FU-FL)))
{
sucesso4[i] = 1
}
else{
sucesso4[i] = 0
}
```

```
}

###############
#             #
#Resultado    #
#             #
###############

Saida<-cbind(dcook,H,distancia)
print(Saida)

print("dcook com ponto de corte")
print(FUdcook+(3*(FUdcook-FLdcook)))
print("Criterio H - Outlier Discordancia")
print(FUH+(3*(FUH-FLH)))
print("Criterio Dcook")
print(teste_qq)
print("Criterio Geodesica")
print(FU+(3*(FU-FL)))
```

A.5 Dados Reais: 75 observacoes

```
#############################################################################
#     Dados Reais  de uma distribuicao watson para uma amostra n = 75       #
#############################################################################

rm(list=ls())
set.seed(1234)
#k=1
n=75
dados=matrix(NA,n,2)

dip = theta = c(50, 53, 85, 82, 82, 66, 75, 85, 87, 85, 82, 88, 86, 82, 83,
86, 80, 78, 85, 89, 85, 85, 86, 67, 87, 86, 81, 85, 79, 86, 88, 84, 87, 88,
83, 82, 89, 82, 82, 67, 85, 87, 82, 82, 82, 75, 68, 89, 81, 87, 63, 86, 81,
81, 89, 62, 81, 88, 70, 80, 77, 85, 74, 90, 90, 90, 90, 90, 90, 90, 90, 90,
90, 90, 90)

dip_dir = phi = c(65, 75, 233, 39, 53, 58, 50, 231, 220, 30, 59, 44, 54, 251,
233, 52, 26, 40, 266, 67, 61, 72, 54, 32, 238, 84, 230, 228, 230, 231, 40,
233, 234, 225, 234, 222, 230, 51, 46, 207, 221, 58, 48, 222, 10, 52, 49, 36,
225, 221, 216, 194, 228, 27, 226, 58, 35, 37, 235, 38, 227, 34, 225, 53, 57,
66, 45, 47, 54, 45, 60, 51, 42, 52, 63)


x = y = vector()
x=y=z=array(0,dim=c(n,1))
u1=matrix(0,3,n)
for(i in 1:n)
{

x[i] = sin(theta[i])*cos(phi[i])
y[i] = sin(theta[i])*sin(phi[i])
z[i] = cos(theta[i])

u1[,i] = c(x[i],y[i],z[i])# cada coluna é formada por cada vetor

} # aqui termina o for
u<-u1
```

```
#####################################################################
#             Determinando os valores dos k estimados               #
#####################################################################

k_estimado = matrix(0,n,1)

k_estimado = function(a,c,autoval){
(c-a)/(1-autoval)+1-a+((a-1)*(a-c-1)*(1-autoval))/
(c-a)
}

k_estimado1<-function(a,c,r1)
{
(c*r1-a)/(r1*(1-r1))+r1/(2*c*(1-r1))
}

k_estimado2 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+(1-c/c-a))
}

k_estimado3 = function(a,c,r1){
(r1*c-a)/(2*r1*(1-r1))*(1+sqrt(1+(4*(c+1)*r1*(1-r1))/(a*(c-a))))
}

k_estimado4 = function(a,c,r1){
(r1*c-a/r1*(1-r1))*(1+r/a)
}


#####################################################################
#       Determinando os autovalores e autovetores de S              #
#####################################################################
soma<-array(0,dim=c(3,3))
for(j in 1:n){
soma = soma + u[,j]%*%t(u[,j])

}

S = (1/n)*soma
saida1 = eigen(S) #autovalores de S
saida1$vectors # autovetores
saida1$values # autovalores


#########################################################################
#Encontrando meu theta_chap (composto por 4 valores da matriz S)  #
#########################################################################
a = 1/2
c = 3/2
r1 =saida1$values[1]  ## pois r1 tende para 1 (maior autovalor de S)


k_matrizS<-k_estimado1(a,c,r1)

k<-k_matrizS

theta_chap = cbind(t(saida1$vectors[,1]),k_estimado1(a,c,r1))##armazena esses valores em
#uma coluna
t(theta_chap) ## transposto do vetor theta

#print("K")
#print(k)
#print("K estimado")
#print(k_matrizS)
```

```
#############################################################
#                                                           #
#Encontrando theta sem a i-esima observacao (theta_chap1)   #
#                                                           #
#############################################################
somai<-array(0,dim=c(3,3,n))

for(j in 1:n){
somai[,,j]=soma-u[,j]%*%t(u[,j])
}
somai = (1/(n-1))*somai ## fornece 20 matrizes 3 por 3


theta_chap1<-array(0,dim=c(n,4))

for(j in 1:n)
{

saida = eigen(somai[,,j]) #autovalores de somai[,,1]
saida$vectors # autovetores
saida$values # autovalores
r1_autovec = saida$vectors[,1]
r1_autoval<-saida$values[1]



theta_chap1[j,] = cbind(t(r1_autovec),k_estimado1(a,c,r1_autoval)) # matriz 20

}


#####################################################################
#              Valor de M(a,b,k_estimado1(a,c,r1))                  #
#####################################################################

mabk<-function(a,b,k_matrizS)
{
sum = 0
erro=1
i=0

while((erro > 0.01)&(i<90)){

sum1=sum
sum  = sum + (gamma(a+i)*gamma(b)*(k_matrizS^{i}))/
(gamma(a)*gamma(b+i)*factorial(i))
i=i+1
erro = abs(sum-sum1)

}

return(sum)
}

###################################
#                                 #
#Valor numerico das constantes    #
#                                 #
###################################
a<-0.5
b<-1.5
result_mabk<-mabk(1/2,3/2,k_matrizS)
result_ma1b1k<-mabk(3/2,5/2,k_matrizS)
result_ma2b2k<-mabk(5/2,7/2,k_matrizS)
```

```
#result_ma1b1k<-ma1b1k(3/2,5/2,k_matrizS)
#result_ma2b2k<-ma2b2k(5/2,7/2,k_matrizS)


#########################################################################
#Calculando a matriz de informacao de fisher observada (J(theta_chap))#
#########################################################################


IF = matrix(0,4,4) ## matriz de zeros 4 por 4
IF[1,1] = 2*(k_matrizS)*(S[1,1])
IF[1,2] = (k_matrizS)*(S[2,1]+S[1,2])
IF[1,3] = (k_matrizS)*(S[3,1]+S[1,3])
IF[1,4] = 2*(S[1,1])+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[2,1] = (k_matrizS)*(S[2,1]+S[1,2])
IF[2,2] = 2*(k_matrizS)*(S[2,2])
IF[2,3] = (k_matrizS)*(S[3,2]+S[2,3])
IF[2,4] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[3,1] = (k_matrizS)*(S[3,1]+S[1,3])
IF[3,2] = (k_matrizS)*(S[3,2]+S[2,3])
IF[3,3] =  2*(k_matrizS)*(S[3,3])
IF[3,4] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,1] = (2*(S[1,1]))+S[2,1]+S[3,1]+S[1,2]+S[1,3]
IF[4,2] = S[2,1]+S[1,2]+(2*(S[2,2]))+S[3,2]+S[2,3]
IF[4,3] = S[3,1]+S[3,2]+S[1,3]+S[2,3]+(2*(S[3,3]))
IF[4,4] = ((a^{2})*(b+1)*((result_ma1b1k)^{2}))-((a)*(a+1)*(b)*(result_mabk)*(result_ma2b2k))/
((b^{2})*(b+1)*((result_mabk)^{2}))
#Inf_obs = solve(IF) # inversa da matriz de informacao de Fisher observada




#########################################################################
#                   Calculando a distancia de Cook                      #
#########################################################################

dcook=array(0,n,1)
for(j in 1:n){
dif_transp = theta_chap1[j,]-theta_chap
# diferenca entre theta_chap1 e theta_chap
#(vetor com menos uma observacao - vetor com todas as observacoes)

t(dif_transp) # transposto do vetor dif_transp

dcook[j] = (dif_transp)%*%(IF)%*%(t(dif_transp))## alocando os vetores de theta_chap1 da matriz S1 em uma nova matriz

}

########################
#                      #
#Distancia Geodesica   #
#                      #
########################
geodesica<-function(v1,v2)
{
return(acos(t(v1)%*%v2))
}

distancia<-array(0,dim=c(n,1))
tu<-array(0,dim=c(3,n))

for(i in 1:n)
{
if(u[3,i]<0)
{
```

```
tu[,i]<- -u[,i]
}else{
tu[,i]<-u[,i]
}

#distancia[i]<-geodesica(tu[,i],saida1$vectors[,1])
distancia[i]<-min(geodesica(u[,i],saida1$vectors[,1]),geodesica(-u[,i],saida1$vectors[,1]))
}

#########################################################################
#    Criterios para compararar as distancias                            #
#########################################################################

#########################################################################
#    1 Criterio: Ponto de corte para as distâncias                      #
#########################################################################


ddcook  = sort(dcook) # reordena as distancias em ordem crescente
sucesso1 = matrix(0,n,1)
n = 75
f = (1/2)*(trunc((n+3)/2))
FUdcook = ddcook[round(n+1-f)]
FLdcook = ddcook[round(f)]
for(i in 1:n){
if(dcook[i]>FUdcook+(3*(FUdcook-FLdcook)))## variar o valor de k = 1.5, 3,
#4, e 6
{
sucesso1[i] = 1
}
else{
sucesso1[i] = 0
}
}


#########################################################################
#    2 Criterio: Teste outlier para discordancia                        #
#########################################################################

dados = array(0,dim=c(3,n))
dados1 = u1 # fornece os 50 vetores da matria u com todas as observacoes
soma_1 = 0
for(j in 1:n){
soma_1 = soma_1 + dados1[,j]%*%t(dados1[,j]) # fornece as 50 matrizes 3 por 3
# com todas as observacoes
}
# soma_1 = Tn # matriz de orientacao com todas as observacoes
soma_mat = array(0,dim=c(3,3,n))
for(i in 1:n){
dadost = dados1[,-i]
soma_mat1 = 0
for(j in 1:n-1){
soma_mat1 = soma_mat1 + dadost[,j]%*%t(dadost[,j])
}
soma_mat[,,i] = soma_mat1 # fornece as 50 matrizes 3 por 3 com n-1 observacoes

}


saida2 = eigen(soma_mat[,,i]) # autovalores de soma_mat[,,i]
saida2$vectors # autovetores
saida2$values # autovalores

saida2=vector("list", n) ## guardando os autovalores e autovetores de soma_mat
# em uma lista
for(k in 1:n){
```

```
saida2[[k]] = eigen(soma_mat[,,k]) #autovalores de soma_mat[,,k]

}

tau_chap = eigen(soma_1)$values[1]# autovalores de soma_1 # maior autovalor
#da matriz soma_1

tau_chap1 = matrix(0,n,1)
for(i in 1:n){
tau_chap1[i] = eigen(soma_mat[,,i])$values[1] # maior autovalor da matriz
#soma_mat
}


##########################################
#         Estatistica de teste(H)        #
##########################################

H = matrix(0,n,1)
for(i in 1:n){
H[i] = ((n-2)*(1+tau_chap1[i,]-tau_chap))/(n-1-tau_chap1[i,]) # estatistica de teste para
#detectar se o ponto seria um outlier (consideramos outlier se H for muito
#grande!)
}


sucesso2 = matrix(0,n,1)
SH  = sort(H) # reordena as distancias em ordem crescente
n = 75
f = (1/2)*(trunc((n+3)/2))
FUH = SH[round(n+1-f)]
FLH = SH[round(f)]
for(j in 1:n){
if(H[j]>FUH+(3*(FUH-FLH)))## variar o valor de k = 1.5, 3, 4, e 6
{
sucesso2[j]<-1
}
else{
sucesso2[j]<-0
}
}
#plot(H) # grafico de H


#########################################################################
#    3 Criterio: Quantil de uma qui-quadrada                            #
#########################################################################

sucesso3 = matrix(0,n,1)
teste_qq = qchisq(0.95,3)
for(i in 1:n){
if(dcook[i]>teste_qq){
sucesso3[i] = 1
}
else {
sucesso3[i] = 0
}
}




#######################################
#                                     #
# 4 Criterio: Distancia Geodesica     #
```

```
#                                      #
########################################

sucesso4 = matrix(0,n,1)
Sd  = sort(distancia) # reordena as distancias em ordem crescente
f = (1/2)*(trunc((n+3)/2))
FU = Sd[round(n+1-f)]
FL = Sd[round(f)]
for(j in 1:n)
{
if(distancia[j]>FU+(3*(FU-FL)))
{
sucesso4[i] = 1
}
else{
sucesso4[i] = 0
}
}

###############
#             #
#Resultado    #
#             #
###############

Saida<-cbind(dcook,H,distancia)
print(Saida)

print("dcook com ponto de corte")
print(FUdcook+(3*(FUdcook-FLdcook)))
print("Criterio H - Outlier Discordancia")
print(FUH+(3*(FUH-FLH)))
print("Criterio Dcook")
print(teste_qq)
print("Criterio Geodesica")
print(FU+(3*(FU-FL)))
```