IDENTIFICAÇÃO DE PONTOS INFLUENTES EM UMA AMOSTRA ALEATÓRIA DE PRÉ-FORMAS DA DISTRIBUIÇÃO BINGHAM COMPLEXA (DISTÂNCIA DE COOK E MÉTODOS DE BOOTSTRAP)

OLGA PATRICIA REYES FLÓREZ

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, fevereiro de 2009

Flórez, Olga Patricia Reyes

Identificação de pontos influentes em uma amostra aleatória de pré-formas da distribuição Bingham complexa (distância de Cook e métodos de bootstrap) / Olga Patricia Reyes Flórez - Recife : O Autor, 2009.

viii, 56 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Estatística, 2009.

Inclui bibliografia e apêndice.

1. Análise multivariada. 2. Bootstrap. I. Título.

519.535 CDD (22.ed.) MEI2009-022

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

18 de fevereiro de 2009
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Olga Patrícia Reyes

intitulada

"Identificando Pontos Influentes em uma Amostra Aleatória de Pré-Formas (Distância de Cook e Métodos de Bootstrap)"

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

Coordenador da Pós-Graduação em Estatística

Prof. Klaus Leite P. Vasconcellos
Coordenador da Pós Graduação em Estatística
UFPE

Banca Examinadora:

Getúlio José Amorim do Amaral orientador

Ana Lúcia Bezerra Candeias (UFPE)

Francisco José de Azevedo Cysneiros

Este documento será anexado à versão final da dissertação.

À minha mãe Marina.
Em memória de meu pai Julio,
e meus tios Marina e Manuel.

Agradecimentos

A Deus, por me dar esta ótima oportunidade.
A minha mãe Marina, pelo apoio, amor e os múltiplos envios de comida colombiana.
A meus irmãos Carlos, Helda e Leo, a meus sobrinhos, Angelica, Juan Sebastian, Paula e Tomás, por sua constante voz de ânimo, por ser o motivo para seguir sempre em frente.
A meus tios, Claudia Helda, Mario, Humberto, Esperanza, Gustavo, Luis, Chelita e Cecilia.
A meus primos em especial a Tati, Paola, Manuel, Olga, Patricia, Yolanda, Jorge e Jualia, pelo apoio e boas energias.
A meus amigos em Colômbia Lenka, Carolina, Luisa, Gloria, Milena, Andrea, Sandra, Johanny, Alexandra, Holman, Edgar, Camilo, Euro, Daniel, Edwin, Sandra Alejo, Anita, Diego Panelito, Nano, Jas, Linda.
A minha familia colombiana en Brasil Hans, Cristiane, Miguel Zorro, Carlos, Walter, Alejandra, Alejandro, Miguel Angel, Piña, Janeth, Carolina, Luis Arturo, Ana Julia, Diego, Rebeca, Alexandra e Edwin, pelos momentos de diversão que tornaram mais amena minha permanência em Recife.
A minha familia Brasileira Alice, Rafaella, Claudia, Lidia (titia I), Ligia (titia II), Valmir, Juliana, Marcelitooooo, Raphael (éééééé), Hemilio, Kalil, Rafael, Luiz, Iva Iva e Izabel, que seria de minha vida sem voces?.
A meus amigos e colegas de mestrado Andrea, Wilton, Manoel, Jimmy, Josimar, Manoel, Lutemberg, Jeremias, Elton, Tarciana, Tatiene, Fabio, Abraão, Fabio V., Larissa, Carlos, Ricardo (RiRi), Iuri, Bruno, Ivan, Amanda, Alinne, Solange e Marcelo, pela amizade e apoio.
A meus professores em Colômbia Luis Alberto López, Humberto Mayorga e Piedad Urdinola, pela confiança e apoio.
A meu orientador Getúlio Amaral, pela orientação e a paciência.
À banca examinadora Francisco Cysneiros e Ana Lúcia Bezerra, pelas correções e sugestões.
A Valéria por ser tão gente boa, fofa e pelo carinho.
À CAPES pelo apoio financeiro.

Resumo

O objetivo desta dissertação é avaliar e aplicar métodos de análise de influência na análise estatística de formas. A partir do modelo de deleção de casos (CDM) obtem-se uma medida da distância de Cook quando o conjunto de dados tem distribuição Bingham complexa. Mediante simulações de Monte Carlo e o método bootstrap tem-se a estimação da região de confiança para a distância de Cook em diferentes conjuntos de dados com distribuição Bingham complexa. Além disso, é mostrado nesta dissertação que outros métodos para análise de influência podem funcionar em análise de formas. A eficácia da distância de Cook frente aos métodos apresentados é avaliada.

Palavras-chave: Bingham complexa; distância de Cook; bootstrap; análise de formas.

Abstract

The objective of this dissertation is the evaluation and application of influence analysis methods in statistical shape analysis. Through the model of deleting cases (CDM) is obtained a measure of the Cook distance when the data set has the complex Bingham distribution. Through simulations of Monte Carlo and the bootstrap method the estimation of region of confidence for the distance of Cook is gotten. By the Monte Carlo simulation and the Bootstrap methods the confidence interval is estimated for the Cook distance in different data sets with complex Bingham distribution. Beyond that, it is shown that other influence analysis methods can work on shape analysis. The efficacy of the Cook distance in the presented methods is evaluated.
Keywords: Bingham complex; Cook distance; bootstrap; shape analysis.

Conteúdo

Lista de Tabelas

Lista de Figuras

CAPÍTULO 1

Introdução

1.1 Introdução

A análise estatística de formas tem sido requerida em vários campos de investigação. Atualmente técnicas científicas e tecnológicas são desenvolvidas com o fim de estudar mais a fundo as formas dos objetos. O uso do computador e os avanços da tecnologia proporcionaram ferramentas úteis para a manipulação e captura de imagens de objetos em 2 e 3 dimensões. Medir, descrever e comparar as formas dos objetos é de grande importância prática em muitas aréas, tais como: biologia, medicina, computação visual, antropologia e arqueologia.

A análise estatística de formas centra-se na metodologia para a análise de formas na presença de aleatoridade. Portanto, um dos objetivos principais da análise de formas é estimar a forma média populacional utilizando amostras e, assim, estudar a estrutura da população, sua variabilidade e realizar inferências. O campo de análise de formas envolve métodos para o estudo das formas de objetos, nos quais informação de locação, rotação e escala podem ser removidos.

A forma mais eficaz para analisar formas de objetos é por meio da obtenção de locações geométricas ou pontos chaves, chamados de *landmarks* (marcos anatômicos). As coordenadas numéricas dos marcos anatômicos são usadas para representar um objeto e estas pertencem a um espaço chamado de espaço de marcos anatômicos.

A partir das coordenadas de um objeto no espaço de marcos anatômicos, um novo conjunto de coordenadas de objetos, que será chamado de coordenadas de pré-forma, pode ser obtido. Com o fim de remover os efeitos de escala e locação, transformações apropriadas são utilizadas. Após a remoção dos efeitos de locação e escala, um novo sistema de coordenadas é considerado, em um novo

espaço que será chamado de espaço de pré-formas. Baseado nas coordenadas pré-forma, a matriz produto (*SSP*) pode ser calculada conjuntamente com a forma média de uma amostra aleatória. A matriz produto representa a variação das coordenadas pré-forma e a forma média é definida como o autovetor associado ao maior autovalor dessa matriz.

Com o objetivo de que a informação de rotação das coordenadas pré-forma de um objeto sejam removidas, o objeto é rotacionado até ele ficar tão próximo quanto possível do "objeto médio".

Considerando que o espaço de pré-formas e o espaço de formas são espaços não-euclideanos, é difícil desenvolver uma análise padrão nesses espaços. É possível definir uma aproximação linear ao espaço euclideano, para evitar dificuldades nos espaços não-euclideanos. O espaço tangente é a versão linear do espaço de formas na proximidade de um ponto particular de espaço de formas. Para uma amostra aleatória de objetos, as coordenadas pré-formas desses objetos podem ser projetadas no espaço tangente da forma média amostral. As novas coordenadas são chamadas coordenadas tangentes.

Métodos inferenciais em análise de formas são freqüentemente realizados no espaço tangente. Tais métodos funcionam melhor quando os dados estão muito concentrados. No espaço tangente, dispõem-se de muitos procedimentos comumente usados em análise multivariada padrão.

É comum encontrar em um conjunto de dados observações que se afastam demasiado das restantes, parecendo que foram geradas por mecanismos diferentes. O estudo destas observações é importante dado que uma etapa importante em qualquer análise estatística de dados é estudar a qualidade das observações. Uma medida importante para a detecção de observações influentes é a proposta por Cook (1977). Nesta dissertação esta medida será aplicada na análise estatística de formas, quando o conjunto de dados têm distribuição Bingham complexa, e esta será comparada com outros métodos utilizados para detectar observações influentes em um conjunto de dados.

1.2 Revisão da Literatura

Em 1977 David Kendall publica *"The diffusion of shape"*, um breve resumo no qual introduz uma nova representação de formas de objetos em espaços complexos projetados. Neste resumo Kendall introduz a difusão da forma como uma evolução aleatória de órbitas (sob a ação de grupos gerados por rotação, translação e escalonamento) de um conjunto de partículas de difusão independentes de acordo com um movimento Browniano no espaço complexo projetado ou espaço euclidiano. Kendall (1984) introduz um sistemas de coordenadas o que se conhece por sistema de coordenadas de Kendall. Neste sistema de coordenadas a locação de um conjunto de formas é removida usando uma matriz especial, a matriz de Helmert. Muitos estatísticos que sabiam pouco sobre espaços complexos projetados e que não trabalhavam em processos de difusão não viram aplicações imediatas para os

seus próprios trabalhos. No entanto, Kendall em uma seqüência de palestras continuo explicando sobre a sua teoria, apresentando conjuntamente aplicações e problemas na arqueologia, desta forma, as teorias de Kendall foram ganhando seguidores.

Em Kendall (1984) um sistema de coordenadas é introduzido: mais tarde, chamaremos isso de sistema de coordenadas de Kendall. Uma contribuição importante de Kendall (1984) foi a definição matemática de forma, onde ele define um espaço matemático para representar a forma de um conjunto rotulado de k pontos em m dimensões.

Por outro lado, em Bookstein (1984) e Bookstein (1986) uma base matemática para o estudo de formas biológicas é apresentado. Introduz também o sistema de coordenadas de Bookstein, que remove os efeitos de translação, rotação e escala manipulando dois dos marcos anatômicos de tal maneira que eles estejam em posições fixas.

Quando Kendall foi convidado para debater o artigo de Bookstein (1986) estabeleceu uma conexão entre as duas teorias. Kendall e Bookstein se basearam no mesmo sentido: as formas poderiam ser representadas como variedades. O sistema de coordenada utilizado por Bookstein é diferente do sistema de coordenada tratado por Kendall. Os marcos anatômicos de Bookstein correspondem ao conjunto de etiquetas de k pontos em m dimensões de Kendall. Uma das diferenças mais importantes foi na aplicação que cada um deles fez. As aplicações de Kendall foram na arqueologia incluindo problemas na organização de antigos túmulos egípcios em ordem cronológica, com base na variação de estilo de artefatos que continham. Enquanto que as aplicações de Bookstein foram na ciências biológicas e medicina. Bookstein reuniu pesquisadores tradicionais tais como D´Arcy Wentworth Thompson, Julian Huxley, e mais tarde reúne pesquisadores em alometria e morfometria multivariada.

D´Arcy Wentworth Thompson pioneiro da matemática biológica e lembrado principalmente como o autor do livro *On Growth and Form* (1917), o livro foi qualificado como uns dos melhores trabalhos da literatura científica que tenha sido registrado na língua Inglesa. Thompson encontrou uma correlação entre formas e fenômenos mecânicos ou forma de uma estrutura biológica e sua função. Suas observações da filotaxia (relações numéricas entre estruturas espirais das plantas) e sua relação com a seqüência de Fibonacci foram básicas com o tempo.

Muito da teoria estatística tem sido dedicada à estimação dos parâmetros de locação e escala. A teoria estatística de formas concentra-se em aspectos relacionados com dados nos quais informação de locação e escala foram removidos. Em 1934 Fisher, R. A. introduz o conceito de *configuração* em amostras univariadas.

O princípio da análise de Procrustes pode ser devido a Mosier e, em seguida, ao trabalho de Sibson e Gower. Na comparação de diferenças de forma entre dois conjuntos de dados, a análise Procrustes procede transformando um conjunto de dados para tentar encaixar em outro. As transformações

nos permitem uma análise padrão incluindo mudanças na locação, escala e rotação. Quando uma transformação de um dos conjuntos de dados foi encontrada para ficar o mais próximo do outro conjunto, a soma das diferenças ao quadrado das coordenadas entre eles é chamada de distância Procrustes entre os dois conjuntos de dados.

Um modelo probabilístico importante para análise estatística de formas foi apresentado por Kent (1994). Este modelo é a distribuição Bingham complexa, uma versão complexa da distribuição Bingham real. Uma propriedade importante da distribuição Bingham complexa é a simetria complexa. Esta simetria complexa significa que um vetor e qualquer versão rotacionada desse vetor terão a mesma distribuição. Esta propriedade é útil porque a análise de formas pode ser desenvolvida enquanto se trabalha com pré-formas. A distribuição Bingham complexa na esfera é tão importante quanto a distribuição normal na reta.

Mardia e Dryden (1998) aproveitaram as observações de Kendall e começaram a trabalhar para enfocar as idéias na prática estatística. O modelo probabilístico apresentado por Kent resultou estar estreitamente relacionado com Mardia e o labor de décadas anteriores relacionado com a distribuição de dados direcionais. Dryden e Mardia a partir do final da década de 1980 publicaram uma grande série de trabalhos originais. Por exemplo, a análise estatística de *landmarks* usando a teoria da distribuição conjunta do tamanho e da forma de configurações planas gaussianas. Este trabalho foi uma extensão de trabalhos anteriores que consideraram análise de forma marginal. Casos especiais da distribuição de tamanho e forma são examinados e o modelo gaussiano isotrópico também.

1.3 Organização da Dissertação

No Capítulo 2 se apresentam conceitos básicos da análise estatística de formas. Serão revisados conceitos relacionados à representação matemática de formas, tais como matrizes de configuração e espaço de formas, diferentes sistemas de coordenadas e análise de Procrustes. Distribuições comumente utilizadas na análise estatística de formas, e o algoritmo para gerar a distribuição Bingham complexa serão apresentadas neste capítulo. No Capítulo 3, serão apresentados conceitos de influência tais como distância de Cook, influência local e influência em dados normais multivariados. No Capítulo 4 serão vistos o teste de discordância para dados distribuídos na esfera, os conceitos de influência são aplicados na teoria da análise estatística de formas, e será apresentada a distância de Cook para formas com distribuição Bingham complexa. No Capítulo 5 apresenta-se um exemplo dos diferentes métodos para avaliar influência em um conjunto de dados com distribuição Bingham complexa. São apresentados resultados numéricos, onde a eficiência da distância de Cook para formas com distribuição Bingham complexa será comparado com outros testes de influência. No Capítulo 6 são apresentadas as conclusões obtidas na dissertação. Finalmete no Capítulo 7 são apresentadas

algumas sugestões para trabalhos futuros.

1.4 Plataforma Computacional

Os resultados numéricos apresentados nesta dissertação foram obtidos utilizando o ambiente de programação e análise de dados R em sua versão $2.6.0$ para os sistemas operacionais Microsoft Windows e Linux. O R se encontra disponível gratuitamente através do site `http://www.R-project.org`.

A presente dissertação de mestrado foi digitada utilizando o sistema de tipografia LaTeX, que consiste em uma série de macros ou rotinas do sistema TeX que facilitam o desenvolvimento da edição de texto. Detalhes sobre o sistema de tipografia LaTeX podem ser encontrados em De Castro Korgi (2003).

CAPÍTULO 2

Análise Estatística de Formas

2.1 Conceitos Básicos

Os avanços da tecnologia têm conduzido a uma obtenção sistemática de imagens em 2 ou 3 dimensões e o estudo da forma dos objetos é cada vez mais utilizado na solução de problemas reais.

Neste trabalho, o conceito de *forma* representa toda a informação geométrica que se mantém quando efeitos de localização, dimensão e rotação são retirados de um objeto. Portanto, a forma de um objeto é invariante sob os efeitos de translação, rotação e escala. Dois objetos têm a mesma forma se eles podem ser transladados, redimensionados e rotacionados uns aos outros para que eles correspondam exatamente, isto é, se os objetos são semelhantes.

Descreve-se a forma por um número finito de pontos de localização em cada espécime que serão chamados de *landmark*. Um *landmark* (marco anatômico) é um ponto de correspondência em cada objeto. O marco anatômico pode ser fixado pelo conhecimento do pesquisador e/ou por questões geométricas.

2.1.1 Representação Matemática de Formas

Uma *configuração* é um conjunto de marcas coordenadas sobre um determinado objeto. A *matriz de configuração* Y é uma matriz de dimensão $k \times m$ de coordenadas Cartesianas de k marcas em m dimensões. A matriz de configurações é dada por

$$\begin{pmatrix} y_{1,1} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots \\ y_{k,1} & \cdots & y_{k,m} \end{pmatrix} \tag{2.1}$$

O *espaço de configuração* é o espaço de todas as possíveis marcas coordenadas. Removendo-se a informação sobre escala, locação e rotação a forma da matriz de configuração é obtida.

O espaço de marcos anatômicos é um espaço real $\mathbb{R}^m$ onde são representadas as coordenadas cartesianas de um marco. Por exemplo, para objetos bidimensionais ($m = 2$), o espaço de marcos anatômicos é $\mathbb{R}^2$. Nesta dissertação, só será tratado o caso $m = 2$.

Algumas operações precisam ser feitas na matriz Y para eliminar os efeitos de locação, escala e rotação. A matriz de configuração pode ser escrita como um vetor complexo quando $m = 2$. Defina um vetor complexo $k \times 1$

$$z^0 = (y_{1,1} + iy_{1,2}, \ldots, y_{k,1} + iy_{k,2})^T = (z^0_{(1)}, \ldots, z^0_{(k)})^T \tag{2.2}$$

cujos elementos correspondem as coordenadas complexas dos marcos anatômicos. Quando a configuração conserva os efeitos de locação, escala e rotação, o super índice ^{0} é usado. Os detalhes de cada transformação para o caso $m = 2$ serão dados a seguir.

O primeiro passo é remover a locação. Isso pode ser feito de várias maneiras, dependendo do sistema de coordenadas. Aqui serão usadas as coordenadas de Kendall. Para o sistema de coordenadas de Kendall, serão necessários detalhes sobre a matriz de Helmert e sobre a sub-matriz de Helmert. A sub-matriz de Helmert proporciona uma transformação linear particular que remove a locação pré-multiplicando z^0 [veja Small (1996, p. 130) e Dryden & Mardia (1998, p. 34)].

A sub-matriz de Helmert H é uma matriz de dimensão $(k-1) \times k$, que é a matriz de Helmert H^F com a primeira linha removida. A matriz de Helmert completa H^F é uma matriz ortogonal $k \times k$, cuja primeira linha tem todos os elementos iguais a $1/\sqrt{k}$, e tem a $(j+1)$-ésima linha, para $j \geq 1$, dada por

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

com $j = 1, \ldots, k-1$, onde o número de zeros na $(j+1)$-ésima linha é igual a $k - j - 1$. Por exemplo, se o número de marcos anatômicos é 4, a matriz de Helmert completa é dada por

$$H^F = \begin{pmatrix} 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ -1/\sqrt{6} & -1/\sqrt{6} & 2/\sqrt{6} & 0 \\ -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & 3/\sqrt{12} \end{pmatrix},$$

e a sub-matriz de Helmert é

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

A configuração helmertizada é dada por

$$\omega = Hz^0 \tag{2.3}$$

Diz-se que uma configuração é centrada se $1_k^T z^0 = 0$, onde 1_k é um vetor de uns de dimensão $k \times 1$. Configurações helmertizadas estão conectadas a configurações centradas pela seguinte propriedade da matriz de Helmert [veja Dryden & Mardia (1998, p. 54)]:

$$H^T H = I_k - \frac{1}{k} 1_k 1_k^T,$$

onde I_k é a matriz identidade de ordem $k \times k$. Além disso, uma vez que H^F é ortogonal, tem-se $H^T H = I_{k-1}$. Desse modo, sendo o vetor $z^0 = (z^0_{(1)}, \ldots, z^0_{(k)})^T$ uma configuração complexa, temos

$$(I_k - \frac{1}{k} 1_k 1_k^T) z^0 = z^0 - \bar{z}^0 1_k$$

onde $\bar{z}^0 = k^{-1} \sum_{i=1}^{k} z^0_{(i)}$. Portanto, uma vez que $z^0 - \bar{z}^0 1_k$ é uma configuração centrada, tem-se que as configurações centradas são iguais às configurações helmertizadas multiplicadas por H^T. Aplicar a matriz de Helmert é equivalente a centrar as configurações.

O efeito escala pode ser removido da configuração helmertizada ω usando

$$z = \frac{\omega}{\sqrt{\omega^* \omega}} = \frac{Hz^0}{\sqrt{(Hz^0)^* Hz^0}}, \tag{2.4}$$

onde ω^* é o transposto conjugado complexo de ω. O vetor z é chamado de pré-forma da configuração complexa z^0. Este nome foi usado por Kendall (1984). Note que a pré-forma é uma forma com a informação de rotação conservada.

O conceito de espaço de pré-formas será apresentado porque tem um papel muito importante no domínio da análise de formas [veja Dryden & Mardia (1998, p. 59) e Small (1996, p. 9)]. O espaço de todos os possíveis vetores complexos $(k-1) \times 1$ que não possuem informação de translação nem escala, é chamado espaço de pré-formas. Desse modo, o espaço das pré-formas (S_2^k) consiste em uma hiperesfera unitária complexa em $(k-1)$ dimensões; isto é

$$\mathbb{C}S^{k-2} = \{z : z^* z = 1, z \in \mathbb{C}^{k-1}\}, \tag{2.5}$$

onde $\mathbb{C}^{k-1}$ é o espaço complexo $(k-1)$-dimensional, e $\mathbb{C}S^{k-2}$ é igual à esfera real de raio um na dimensão $2k-2$, S^{2k-3}.

O espaço de formas pode ser pensado como o espaço de pré-formas com a informação de rotação removida. O espaço de formas é o conjunto de todas as possíveis formas. A dimensão do espaço de formas associado a objetos com k marcos anatômicos em m dimensões é dada por

$$km - m - 1 - \frac{m(m-1)}{2}.$$

O termo km é a dimensão total da matriz de configuração Y e se subtrai m, 1 e $m(m-1)/2$ como conseqüência da remoção de locação, escala e rotação, respectivamente [Dryden & Mardia (1998, p.56)]. A seguinte classe de equivalência pode eliminar a informação de rotação no vetor de pré-forma z:

$$[z^0] = \{e^{i\theta} z : \theta \in [0, 2\pi)\}, \tag{2.6}$$

onde por $[z^0]$ se identifica qualquer uma de suas versões rotacionadas. Kendall (1984) ressalta que o espaço de formas quando $m = 2$ é o espaço complexo projetado $\mathbb{C}P^{k-2}$, o espaço de linhas complexas passando pela origem. Assim, o espaço para k pontos em duas dimensões é definido por:

$$\Sigma_2^k = \mathbb{C}P^{k-2}.$$

2.1.2 Forma Média

Na análise de formas enfrentamos muitas situações em que precisamos obter uma estimativa de uma forma média. Considera-se agora um método para estimar a forma média de uma determinada população, o que proporciona uma adequada noção da forma média. Seja $z_1^0, \ldots, z_n^0$ uma amostra aleatória de configurações complexas originária de uma população de objetos Π, onde cada z_i^0 esta definido em (2.2).

Seja $z_1, \ldots, z_n$ as pré-formas de $z_1^0, \ldots, z_n^0$, onde z_i está defina em (2.4) e $z_i \in \mathbb{C}S^{k-2}$. O autovetor correspondente ao maior autovalor da matriz complexa de somas de quadrados e produtos *SSP*, equivale à forma média Procrustes completa $\hat{\mu}$ e é definida por [veja Kent(1994)]

$$\hat{S} = \sum_{i=1}^{n} z_i z_i^*. \tag{2.7}$$

A matriz complexa $\hat{S}$ é hermitiana, pois satisfaz a condição $\hat{S} = \hat{S}^*$ [Axler (1997, p. 128)]. A matrix complexa de somas de quadrados e produtos $\hat{S}$ tem posto completo com probabilidade 1 se a distribuição das pré-formas tem densidade com respeito à distribuição uniforme na esfera das pré-formas e $n \geq k-1$. Aplicando o teorema da decomposição espectral para matrizes hermitianas, dado em Mirsky (1955, p. 388), $\hat{S}$ é escrito como

$$\hat{S} = \sum_{j=1}^{k-1} \hat{\lambda}_j \hat{\mu}_j \hat{\mu}_j^*, \tag{2.8}$$

onde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{k-1} \geq 0$ são os autovalores e $\hat{\mu}_1, \ldots, \hat{\mu}_{k-1}$ são os autovetores correspondentes de $\hat{S}$.

Dado que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_{k-1} \geq 0$, que será comumente o caso na prática, a forma média $\hat{\mu}$ é definida por

$$\hat{\mu} = \hat{\mu}_1. \tag{2.9}$$

2.2 Sistemas de Cooordenadas

A fim de descrever a forma de um objeto é útil especificar um sistema de coordenadas. Considera-se neste trabalho alguns dos sistemas de coordenadas mais usados e importantes em análise estatística de formas. A escolha adequada do sistema de coordenadas para formas será invariante sob locação, escala e rotação da configuração. Serão considerados aqui: coordenadas Procrustes completas, coordenadas tangentes e coordenadas polares de Kent.

2.2.1 Coordenadas Procrustes Completas

A análise de Procrustes é uma metodologia utilizada para *encaixar* dois ou mais objetos. O objetivo é fazer transformações tais que duas configurações complexas sejam tão próximos uma da outra quanto possível em termos da norma Euclidiana. Isto é feito usando as pré-formas desses objetos porque estas têm a mesma translação e escala.

Considere-se duas configurações complexas $y = (y_1, \ldots\ldots, y_k)$ e $w = (w_1, \ldots, w_k)$, tais que $y, w \in \mathbb{C}_k$, com $y^* 1_k = 0 = w^* 1_k$, onde y^* denota o transposto do complexo conjugado de y e w^* denota o transposto do complexo conjugado de w. A fim de comparar configurações precisamos estabelecer uma medida de distância entre as duas formas.

Um procedimento adequado para encaixar w em y é usando transformações de similaridade. As diferenças observadas entre y ajustado e y observado, indicam a magnitude das diferenças de forma entre w e y. Considere o modelo de regressão complexo

$$\begin{aligned} y &= (a + ib)1_k + \beta e^{i\theta} w + \epsilon \\ &= [1_k, w]A + \epsilon \\ &= X_D A + \epsilon \end{aligned} \tag{2.10}$$

onde $A = (A_1, A_2)^T = (a + ib, \beta e^{i\theta})$ são parâmetros complexos 2×1 com translação $a + ib$, escala $\beta > 0$ e rotação $0 \leq \theta < 2\pi$; ϵ é um vetor complexo de erros; e $X_D = [1_k, w]$ é a matriz de planejamento $k \times 2$.

Para obter as estimativas dos parâmetros desta regressão complexa, é necessário minimizar a soma de quadrados dos erros que é dada por

$$D^2(y,w) = \epsilon^*\epsilon = (y - X_D A)^*(y - X_D A).$$

A superimposição Procrustes completa de w em y é obtida por estimar A com $\hat{A}$, onde

$$\hat{A} = (\hat{a} + i\hat{b}, \hat{\beta}e^{i\hat{\theta}})^T = \text{arginf } \epsilon^*\epsilon = \text{arginf } (y - X_D A)^*(y - X_D A).$$

O ajuste Procrustes completo (superimposição) de w em y é dado por

$$w^P = X_D\hat{A} = (\hat{a} + i\hat{b})1_k + \hat{\beta}e^{i\hat{\theta}}w,$$

onde o vetor (β, θ, a, b) é escolhido de tal maneira que minimize

$$D^2(y,w) = ||y - w\beta e^{i\theta} - (a + ib)1_k||^2.$$

A soma de quadrados $D^2(y,w)$ será minimizada ao escolher os seguintes valores

$$\hat{a} + i\hat{b} = 0, \tag{2.11}$$

$$\hat{\theta} = \arg(w^*y) = -\arg(y^*w), \tag{2.12}$$

$$\hat{\beta} = \frac{(w^*yy^*w)^2}{w^*w} \tag{2.13}$$

Detalhes sobre a prova destes resultados podem ser encontrados em Dryden & Mardia (1998).

Esta é a solução de mínimos quadrados ordinários, mas com variáveis complexas. A solução pode ser escrita da seguinte maneira:

$$\hat{A} = (\hat{A}_1, \hat{A}_2)^T = (X_D^*X_D)^{-1}X_D^*y \Rightarrow \hat{A}_1 = 0, \quad \hat{A}_2 = \frac{w^*y}{(w^*w)}. \tag{2.14}$$

Note que o ajuste Procrustes completo de w em y é dado explicitamente por

$$w^P = X_D\hat{A} = \hat{\beta}e^{i\hat{\theta}}w = \frac{w^*yw}{(w^*w)}. \tag{2.15}$$

O vetor de resíduos $r = y - X_D\hat{A}$ é dado por

$$r = [I_k - X_D(X_D^*X_D)^{-1}X_D^*]y = (I_k - H_c)y,$$

onde H_c é a *matriz chapéu* de X_D. Isto é

$$H_c = X_D(X_D^*X_D)^{-1}X_D^*.$$

O valor minimizado da função D^2 é

$$D^2(r,0) = r^*r = y^*y - \frac{(y^*ww^*y)}{(w^*w)}. \tag{2.16}$$

A expressão (2.16) não é simétrica em y e w a menos que $y^*y = w^*w$. Uma padronização conveniente é tomar as configurações como unitárias, ou seja,

$$\sqrt{y^*y} = \sqrt{w^*w} = 1.$$

Se a padronização for considerada, então temos uma medida adequada de distância entre formas.

A *distância Procrustes completa* entre duas configurações complexas w e y é definida por:

$$\begin{aligned} d_F(w,y) &= \inf_{\beta,\theta,a,b} \left\| \frac{y}{||y||} - \frac{w}{||w||}\beta e^{i\theta} - a - ib \right\| \\ &= \left\{ 1 - \frac{y^*ww^*y}{w^*wy^*y}^{1/2} \right\}. \end{aligned} \tag{2.17}$$

Seja $z_1, \ldots, z_n$ uma amostra aleatória de pré-formas e sejam $\omega_1, \ldots, \omega_n$ as correspondentes configurações helmertizadas.

As configurações têm uma rotação arbitraria [veja Dryden & Mardia (1998, pp. $44-45$)]. Assim, antes de proceder com a análise estatística de formas, é necessário rotacionar todas as configurações de tal maneira que estejam o mais próximo possível da forma media amostral. O cálculo é o mesmo de (2.15):

$$\omega_i^P = \frac{\omega_i^* \hat{\mu} \omega_i}{\omega_i^* \omega_i}, \quad i = 1, \ldots, n, \tag{2.18}$$

desta forma $\omega_1^P, \ldots, \omega_n^P$ são as coordenadas Procrustes completas.Quando as pré-formas são escritas como $z_i = Hz_i^0/\sqrt{(Hz_i^0)^*Hz_i^0} = \omega_i/\|\omega_i\|$, onde $\|\omega_i\| = \sqrt{\omega_i^*\omega_i}$, as coordenadas Procrustes também podem ser calculadas assim:

$$\omega_i^P = z_i^* \hat{\mu} z_i, \quad i = 1, \ldots, n.$$

2.2.2 Coordenadas Tangentes

O espaço tangente é a versão linear do espaço de formas na proximidade de um ponto particular do espaço de formas (o pólo da projeção tangente). O pólo é habitualmente escolhido para ser a forma média, obtida a partir do conjunto de dados de interesse, e daí a escolha das coordenadas depende do conjunto de dados em estudo. Suponha uma projeção tangente para a esfera de pré-formas que não depende da rotação da figura original e portanto, um sistema adequado de coordenadas tangentes para formas.

A distância Euclidiana no espaço tangente ao espaço de formas é uma boa aproximação para a distância Procrustes (2.17) no espaço de formas nas imediações do pólo. Na análise de formas vemos que o espaço tangente é extremamente importante e útil.

Existem diferentes tipos de coordenadas no espaço tangente. Aqui, usaremos as coordenadas Procrustes parciais, que são dadas por

$$t_i = e^{i\hat{\theta}}[I_{k-1} - \hat{\mu}\hat{\mu^*}]z_i, \quad i = 1, \ldots, n, \tag{2.19}$$

onde z_i é um vetor pré-forma definido em (2.4) e $\hat{\theta} = \arg(-\hat{\mu}^* z_i)$ minimiza $\|\hat{\mu} - z e^{i\hat{\theta}}\|^2$ e $\|z\| = \sqrt{z^* z}$. Coordenadas tangentes Procrustes parciais envolvem apenas rotação (e não escalonamento) para encaixar com a pré-forma.

Suponha que $z_1, \ldots, z_n$ é uma amostra aleatória de pré-formas e $t_1, \ldots, t_n$ são suas coordenadas tangentes, onde cada z_i e t_i são calculadas usando (2.4) e (2.19), respectivamente. Seja v_i um vetor de tamanho $(2k-2) \times 1$ obtido empilhando as partes real e imaginária das coordenadas da cada t_i. Se $t_i = x_i + iy_i$, essa operação é representada por $cvec$ e define-se

$$v_i = cvec(t_i) = (x_i^T, y_i^T), \tag{2.20}$$

onde $x_i = \Re(t_i)$ é a parte real de t_i e $y_i = \Im(t_i)$ é a parte imaginária de t_i. Se o número de marcos anatômicos é k, um vetor pré-forma z_i tem dimensão $(k-1)$ e seu correspondente vetor de coordenadas tangentes v_i, onde v_i é dado em (2.20), tem dimensão $(2k-2)$.

Métodos padrão de análise multivariada podem ser aplicados a coordenadas tangentes reais v_i. Quando os dados estão altamente concentrados, métodos baseados na distribuição normal multivariada podem ser aplicados às coordenadas reais tangentes v_i [veja Dryden & Mardia (1998, p. 151)].

2.2.3 Coordenadas Polares de Kent

Kent (1994) propõe um sistema de coordenadas polares na esfera. Seja $(z_1, \ldots, z_{k-1})^T$ um ponto em $\mathbb{C}S^{k-2}$. Este é transformado em $(s_1, \ldots, s_{k-2}, \theta_1, \ldots, \theta_{k-1})$ onde

$$\Re(z_j) = s_j^{1/2} \cos(\theta_j), \;\; \Im(z_j) = s_j^{1/2} \sin(\theta_j), \tag{2.21}$$

para $j = 1, \ldots, k-1$, $s_j \geq 0$, $0 \leq \theta_j \leq 2\pi$ e $s_{k-1} = 1 - s_1 - \ldots - s_{k-2}$. Mais detalhes sobre este sistema de coordenadas podem ser encontrados em Shelupshy (1962). As coordenadas $s_1, \ldots, s_{k-2}$ pertencem ao *simplex* unitário de dimensão $k-2$, S_{k-2}. Ao identificar a esfera complexa de pré-formas com o produto cartesiano $S_{k-2} \times [0, 2\pi)^{k-1}$, temos a medida de volume de $\mathbb{C}S^{k-2}$ por

$$2^{2-k} \mathrm{d}s_1 \ldots \mathrm{d}s_{k-2} \mathrm{d}\theta_1 \ldots \mathrm{d}\theta_{k-1}. \tag{2.22}$$

O volume total é

$$\frac{2\pi^{k-1}}{(k-2)!},$$

uma vez que o volume do j-ésimo *simplex* é $1/j!$, para $j = 1, 2, \ldots$.

Rotacionando z a um eixo fixado, coordenadas de forma podem ser obtidas. Considere a informação de rotação da figura original em θ_{k-1}; então, as coordenadas de forma (de dimensão $2k-4$) são

$$(s_1, \ldots, s_{k-2}, \phi_1, \ldots, \phi_{k-2}), \tag{2.23}$$

onde $\phi_j = \theta_j - \theta_{k-1}$, $j = 1, \ldots, k-2$. Então a medida de volume no espaço de formas é

$$2^{2-k}\mathrm{d}s_1 \ldots \mathrm{d}s_{k-2}\mathrm{d}\phi_1 \ldots \mathrm{d}\phi_{k-2}, \tag{2.24}$$

e o volume total é

$$\frac{\pi^{k-2}}{(k-2)!}.$$

2.3 Algumas Distribuições de Probabilidade em Análise de Formas

São de fundamental interesse as distribuições de probabilidade no espaço de formas, pois estas fornecem modelos para a análise estatística de formas. Considerando que o espaço de formas é não-euclidiano, as distribuições têm que ser tratadas com um cuidado especial. Neste trabalho trataremos com quatro importantes distribuições de probabilidade: a distribuição uniforme, a distribuição Bingham complexa, distribuição a normal complexa e distribuição Watson complexa.

2.3.1 Distribuição Uniforme

Usando coordenadas polares de Kent, foi mostrada uma medida de volume no espaço de pré-formas na equação (2.22) e uma medida de volume no espaço de formas na equação (2.24). Neste mesmo espaço, ao normalizar a medida de volume no espaço de formas, obtém-se a medida uniforme $\mathrm{d}\gamma$. Considere a configuração z^0; as coordenadas polares de Kent (2.21) na esfera de pré-formas são obtidas da pré-forma $z = (z_1, \ldots, z_{k-1})^T = Hz^0/\|Hz^0\|$ por

$$Re(z_j) = s_j^{1/2}\cos(\theta_j), \quad \Im(z_j) = s_j^{1/2}\sin(\theta_j)$$

para $j = 1, \ldots, k-1$, $s_j \geq 0$, $0 \leq \theta_j \leq 2\pi$.

A *medida de forma uniforme* é dada por

$$\mathrm{d}\gamma = \frac{(k-2)!}{(2\pi)^{k-2}}\mathrm{d}s_1 \ldots \mathrm{d}s_{k-2}\mathrm{d}\phi_1 \ldots \mathrm{d}\phi_{k-2},$$

com $\int \mathrm{d}\phi = 1$.

Transformando $U^K = (U_3^K, \ldots, U_k^K, V_3^K, \ldots, V_k^K)^T$ para coordenadas de Kendall, temos a medida uniforme sobre o espaço de formas dada por

$$\mathrm{d}\gamma = f_\infty(u^K)\mathrm{d}U_3^K \ldots \mathrm{d}U_k^K\mathrm{d}V_3^K \ldots \mathrm{d}V_k^K$$

onde

$$f_\infty(u) = \frac{(k-2)!\pi}{\{\pi(1+u^Tu)\}^{k-1}}.$$

Para prova deste resutado veja Dryden & Mardia (1998).

2.3.2 Distribuição Bingham Complexa

Vamos considerar o caso de uma distribuição de probabilidade na esfera de pré-formas $\mathbb{C}S^{k-2}$, onde $\mathbb{C}S^{k-2}$ é a esfera unitária complexa em $k-1$ dimensões definida em (2.5). No caso de análise de formas, considere k marcos anatômicos em $m=2$ dimensões com coordenadas complexas escritas z^0 (2.2), isto é, vetores complexos $k \times 1$. Pré-multiplicando z^0 pela sub-matriz de Helmert (de dimensão $(k-1) \times k$), obtém-se $k-1$ configurações helmertizadas ω (2.3). Normalizando por $\|\omega\|$, segue que a pré-forma é dada por

$$z = (z_1, \ldots, z_{k-1})^T = \frac{\omega}{\|\omega\|} \quad \in \mathbb{C}S^{k-2} \tag{2.25}$$

A distribuição de Bingham complexa em $\mathbb{C}S^{k-2}$, denotada $\mathbb{C}B_{k-2}(A)$, tem função de probabilidade

$$f(z) = c(A)^{-1} \exp(z^* A z), \qquad z \in \mathbb{C}S^{k-2}, \tag{2.26}$$

onde z^* denota o transposto conjugado complexo de z, A é uma matriz hermitiana $(A = A^*)$ de tamanho $(k-1) \times (k-1)$ e $c(A)$ é uma constante normalizadora dada por:

$$c(A) = 2\pi^{k-1} \sum_{j=1}^{k-1} a_j \exp(\lambda_j), \qquad a_j^{-1} = \prod_{i \neq j} (\lambda_j - \lambda_i), \tag{2.27}$$

em que $\lambda_1 < \lambda_2 < \ldots < \lambda_{k-1} = 0$ representam os autovalores de A. Note que $c(A) = c(\Lambda)$ depende apenas dos autovalores de A e $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{k-1})$. A prova deste resultado pode ser encontrada em Dryden & Mardia (1998, p. 113).

A distribuição de Bingham complexa foi obtida a partir da distribuição Bingham real, que é utilizada em dados esféricos.

A distribuição tem a seguinte propriedade:

$$f(e^{i\theta} z) = f(z)$$

e, portanto, é invariante em relação a rotações da pré-forma z. Então, se um objeto é rotacionado, ele tem a mesma densidade e tanto este quanto o objeto original participarão identicamente da inferência. Esta propriedade faz da Bingham complexa uma distribuição adequada para análise de formas.

Uma vez que $z^* z = 1$ para $z \in \mathbb{C}S^{k-2}$, pode-se ver que as matrizes parâmetro A e $A + \alpha I$ definem a mesma distribuição de Bingham complexa, com $c(A + \alpha I) = c(A) e^{\alpha}$, onde α é um número complexo. Para mais detalhes sobre as propriedades da distribuição de Bingham complexa veja Kent (1994).

Relação com a Distribuição Bingham Real

A distribuição Bingham complexa de dimensão $k-2$ pode ser considerada um caso especial da distribuição Bingham real de dimensão $2k-3$ da seguinte forma. Se o $j-$ésimo elemento de z é $(z)_j = x_j + iy_j$, define $u = (x_1, y_1, \ldots, x_{k-1}, y_{k-1})^T$ um vetor de dimensão $(2k-2)$, ou seja, dividir cada número em suas partes real e imaginaria. Além disso, se $A = (a_{hj})$ possui entradas $a_{hj} = \alpha_{hj}\exp(i\theta_{hj})$ com $\theta_{jh} = -\theta_{hj}$, $-\pi \leq \theta_{jh} \leq \pi$, defina-se a matriz B de dimensão $(2k-2)\times(2k-2)$ composta por $(k-1)^2$ blocos de tamanho (2×2) dada por

$$B_{hj} = \alpha_{hj}\begin{pmatrix} \cos\theta_{hj} & -\sin\theta_{hj} \\ \sin\theta_{hj} & \cos\theta_{hj} \end{pmatrix}$$

Então $z^*Az = u^TBu$ de modo que a distribuição Bingham complexa para z é equivalente a distribuição Bingham real para u.

Propriedades

Considere $\gamma_1, \ldots, \gamma_{k-1}$ denotam os autovetores padronizados de A tais que $\gamma_j^*\gamma_j = 1$, $\gamma_i^*\gamma_j = 0$, $i \neq j$ e $A\gamma_j = \lambda_j\gamma_j$, $j = 1, \ldots, k-1$. Cada γ_j é definido apenas devido a rotação por um escalar complexo unitário. Se os $\lambda_1, \ldots, \lambda_{k-2}$ estão afastados de 0 (zero), a distribuição das pré-formas tem *alta concentração*, ou seja, a variabilidade das pré-formas na esfera complexa é baixa. Entretanto quando os λ_j para todo j, a distribuição Bingham terá *baixa concentração*, isto é, a variabilidade das pré-formas na esfera complexa é alta, neste caso a distribuição das pré-formas tenderá para a distribuição uniforme em $\mathbb{C}S^{k-2}$.

Estimadores de Máxima Verossimilhança

Seja $z_1, \ldots, z_n$ uma amostra de uma população modelada pela distribuição Bingham complexa, com $n \geq k-1$. Seja

$$S = \sum_{i=1}^{n} z_i z_i^*$$

a matriz complexa de somas de quadrados e produtos de tamanho $(k-1)\times(k-1)$. Suponha que os autovalores de S são positivos e distintos, $0 < l_1 < \ldots < l_{k-1}$ e seja $g_1, \ldots, g_{k-1}$ seus correspondentes autovetores. Note que $\sum_j l_j = n$.

Assim, sob a distribuição Bingham complexa os estimadores de máxima verossimilhança (*MLE*) para $\gamma_1, \ldots, \gamma_{k-1}$, e $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{k-1})$ são dados por

$$\hat{\gamma}_j = g_j, \quad j = 1, \ldots, k-1,$$

e a solução de

$$\frac{\partial \log c(\Lambda)}{\partial \lambda_j} = \frac{1}{n} l_j, \quad j = 1, \ldots, k-2,$$

sob alta concentração das pré-formas

$$\hat{\lambda}_j \cong -n/l_j, \quad j = 1, \ldots, k-2.$$

A prova destes resultados será apresentada agora e pode ser encontrada em Dryden & Mardia (1998). Para os dados a log-verossimilhança reduz a

$$\begin{aligned} L &= \sum_{i=1}^{n} z_i^* A z_i - n \log c(A) = \text{traço} \left(\sum_{i=1}^{n} z_i z_i^* A \right) - \log c(A) \\ &= \text{traço}(SA) - n \log c(\Lambda) = \sum_j \lambda_j \gamma_j^* S \gamma_j - n \log c(\Lambda). \end{aligned}$$

Considerando $\lambda_1 < \ldots < \lambda_{k-2} < \lambda_{k-1} = 0$ constantes, pode-se constatar que

$$\hat{\gamma_j} = g_j, \quad j = 1, \ldots, k-1,$$

e então

$$L = \sum_{j=1}^{k-1} l_j \lambda_j - n \log c(\Lambda). \tag{2.28}$$

Os estimadores de máxima verossimilhança dos autovalores são encontrados através da resolução de

$$\frac{\partial \log c(\Lambda)}{\partial \lambda_j} = \frac{1}{n} l_j, \quad j = 1, \ldots, k-2.$$

Sob alta concentração

$$\log c(\Lambda) \simeq \text{const} - \sum_{j=1}^{k-2} \log(-\lambda_j)$$

dando $\hat{\lambda}_j \cong -n/l_j$, $j = 1, \ldots, k-2$.

O autovetor dominante $\hat{\gamma}_{k-1}$ pode ser considerado como o eixo médio dos dados, ou seja, uma estimativa da forma média.

2.3.3 Distribuição Normal Complexa

Seja $z_j = x_j + iy_j$ uma variável aleatória com distribuição conjunta normal complexa com media $\xi_j = \mu_j + i\nu_j$, $j = 1, \ldots, p$ e $\Sigma = \Sigma_1 + i\Sigma_2$ é uma matriz de covariância $p \times p$ hermitiana. Se $\boldsymbol{x} = (x_1, \ldots, x_p, y_1, \ldots, y_p)^T$ e $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p, \nu_1, \ldots, \nu_p)^T$, tem-se que

$$\boldsymbol{x} \sim N_{2p} \left(\boldsymbol{\mu}, \frac{1}{2} \begin{bmatrix} \Sigma_1 & \Sigma_2 \\ -\Sigma_2 & \Sigma_1 \end{bmatrix} \right), \tag{2.29}$$

onde $\Sigma_2 = -\Sigma_2^T$ é anti-simétrica e Σ_1 é simétrica positiva definida. Em particular, var$(x_j) =$ var(y_j) e cov$(x_j, y_j) = 0$, e em cada ponto a estrutura de covariância é isotrópica. Dados $z = (z_1, \ldots, z_p)^T$ e $\xi = (\xi_1, \ldots, \xi_p)$, a função densidade de probabilidade da distribuição normal complexa é

$$f(z) = \frac{1}{\pi^p |\Sigma|} e^{-(z-\xi)^* \Sigma^{-1} (z-\xi)}. \tag{2.30}$$

A notação $z \sim \mathbb{C}N(\xi, \Sigma)$ denota que z tem distribuição normal complexa com vetor de médias ξ e matriz de covariância Σ.

A distribuição de Bingham complexa pode ser obtida condicionando uma distribuição normal complexa multivariada com média zero a ter norma 1. Ou seja, se $w \sim \mathbb{C}N_{k-1}(0, \Sigma)$, tem-se

$$w|\{\|w\| = 1\} \sim \mathbb{C}B_{k-2}(-\Sigma^{-1}).$$

Então, uma interpretação da matriz hermitiana $-A$ é que esta é a inversa da matriz de covariância de uma variável aleatória normal complexa com média zero, que é condicionada a ter norma 1 para se obter a distribuição de Bingham complexa. Ora, a distribuição Bingham complexa é um exemplo de aproximação condicionada.

Distribuição Watson Real

Dados relacionados à posição angular de linhas aleatórias que não têm uma orientação natural associada à elas, ou em que nenhum final pode ser identificado como o ponto de partida, são medidos em termos de ângulos, em radianos (graus), com um intervalo de valores possíveis $[0, \pi)$. Serão chamados de dados axiais.

Um dos mais simples modelos de dados axiais é o modelo (Dimroh-Scheidegger-Watson), que tem densidade

$$f(\boldsymbol{x}) = C_W \exp\{\kappa(\boldsymbol{\mu x})^2\}, \tag{2.31}$$

onde

$$C_W = 1 \Big/ \left(4\pi \int_0^1 \exp \kappa u^2 \mathrm{d}u\right). \tag{2.32}$$

A distribuição $W(\boldsymbol{\mu}, \kappa)$, definida pela equação (2.31) é rotacionalmente simétrica ao redor de $\boldsymbol{\mu}$. Para $\kappa > 0$, a densidade apresenta máximo em $\pm\boldsymbol{\mu}$, e por isso a distribuição é bipolar. Quando κ aumenta, a distribuição torna-se mais concentrada sobre $\pm\boldsymbol{\mu}$.

2.3.4 Distribuição Watson Complexa

A função de distribuição Watson complexa definida na esfera de pré-formas $\mathbb{C}S^{k-2}$, denotada por $\mathbb{C}W_{k-2}(\mu, \xi)$, tem função de probabilidade dada por

$$f(z) = c_1(\xi)^{-1} \exp\{\xi |z^* \mu|^2\}, \tag{2.33}$$

onde $c_1(\xi)$ é a constante de integração, e se define

$$c_1(\xi) = 2\pi^{k-1}(\xi)^{2-k}\left\{e^{\xi} - \sum_{r=0}^{k-3}\frac{(\xi)^r}{r!}\right\} = \frac{2\pi^{k-1}}{(k-2)!}{}_1F_1(1;k-1;\xi),$$

onde

$${}_1F_1(a;b;x) = 1 + \frac{a}{b}\frac{x}{1!} + \frac{a(a+1)}{b(b+1)}\frac{x^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{x^3}{3!} + \ldots \tag{2.34}$$

é a função hipergeométrica confluente.

O parâmetro de concentração ξ pode ser tomado como negativo, embora em análise de formas o interesse principal é $\xi > 0$. A distribuição Watson complexa é um importante caso particular da distribuição Bingham complexa, quando há apenas dois autovalores distintos na matriz A (o maior autovalor é diferentes dos outros autovalores que serão todos iguais).

2.4 Gerador de Amostras com Distribuição Bingham Complexa

A distribuição Bingham complexa é relevante na análise de formas de objetos. O problema de simular dados a partir desta distribuição reduz-se a simulação de uma distribuição exponencial truncada multivariada [Kent, Constable & Er $(2004$, p. $53)$].

Para $k \geq 3$, seja $\mathbb{C}S^{k-2} = \{z : z^*z = 1, z \in \mathbb{C}^{k-1}\}$ denotando a esfera unitária em $\mathbb{C}^{k-1}$. A distribuição Bingham complexa é definida pela função de probabilidade pela equação (2.26). Sem perda de generalidade, podemos mudar os autovalores de A de modo que eles sejam não positivos com o maior deles igual a 0. Seja $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1} = 0$ os autovalores de $-A$. Considere-se os autovetores padronizados pelas colunas da matriz unitária $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{k-1})$, com $\Gamma^*\Gamma = I_{k-1}$. Escreva $\lambda = (\lambda_1, \ldots, \lambda_{k-2})$ como o vetor dos primeiros $k-2$ autovalores. Os $\{\lambda_j\}$ podem ser pensados como parâmetros de concentração.

Para simular a partir de $\mathbb{C}B_{k-2}(A)$, é conveniente rotacionar o eixo principal, com componentes $w_j = \gamma_j z$, $j = 1, \ldots, k-1$. Usando coordenadas polares, as componentes de $w = (w_1, \ldots, w_{k-1})$ podem ser expressas como $w_j = s^{1/2}e^{i\theta}$, para $j = 1, \ldots, k-1$, $s_j \geq 0$, $0 \leq \theta \leq 2\pi$ e $s_{k-1} = 1 - s_1 - \ldots - s_{k-2}$. As coordenadas $s_1, \ldots, s_{k-2}$ pertencem ao *simplex* unitário de dimensão $k-2$, S_{k-2}.

Sob a distribuição Bingham complexa $\theta_1, \ldots, \theta_{k-1}$ são uniformemente distribuídos em $[0, 2\pi]$ independentemente uns dos outros e os $s = (s_1, \ldots, s_{k-2})^T$ têm "distribuição exponencial truncada multivariada", com função de probabilidade

$$f(s) = d(\lambda)^{-1}\exp(\sum_{j=1}^{k-2} -\lambda_j s_j), \qquad s \in S_{k-2}, \tag{2.35}$$

onde as constantes normalizadoras em (2.26) e (2.35) estão relacionadas por $c(A) = 2\pi^{k-1}d(\lambda)$.

Seja $y = \sum_{j=1}^{k-2} s_j$ e escreva $s_j = yr_j$, $j = 1, \ldots, k-3$ de modo que $y \in [0,1]$ representa o "*tamanho*"de s e $r = (r_1, \ldots, r_{k-3}) \in S_{k-3}$ representa o "perfil"de s. Note que a dimensão de r é menor que a de s em uma unidade. Após da mudança de variáveis, (2.35) torna-se

$$f(y, r) \propto y^{k-3} e^{\lambda y}, \quad r \in S_{k-3}, \ 0 \leq y \leq 1 \tag{2.36}$$

correspondendo a distribuição uniforme sobre o simplex S_{k-3} para r, independente de uma distribuição gama truncada $\Gamma(k-1, \lambda)$ para $0 \leq y \leq 1$.

Uma vez w é simulado, podemos definir $z = \Gamma w$ para obter os valores simulados de z. Assim, a simulação de z essencialmente vem da simulação de s.

2.4.1 Método de Simulação

Suponha uma variável aleatória contínua $X \in \mathbb{R}$ com função de distribuição acumulada $F(x)$. Queremos gerar uma amostra da variável aleatória X a partir de uma amostra com distribuição uniforme em $[0,1]$. Uma alternativa é simular mediante o método da inversão. Se U tem distribuição uniforme no intervalo $[0,1]$, então $X = F^{-1}(U)$ tem a distribuição requerida. Considere a distribuição exponencial truncada no intervalo $[0,1]$. Esta distribuição, denotada TExp(λ), tem função de densidade $f(x) = (\lambda e^{-\lambda x})/(1 - e^{-\lambda})$ e função de distribuição $F(x) = (1 - e^{-\lambda x}/(1 - e^{-\lambda}))$, $0 \leq x \leq 1$. O método de inversão retorna

$$X = -\frac{1}{\lambda} \log(1 - U(1 - e^{-\lambda})).$$

2.4.2 Algoritmo

1. Simular variáveis aleatórias uniformes $U_j \sim U[0,1]$, $j = 1 \ldots, k-2$.
2. Calcular $S'_j = -(1/\lambda_j) \log(1 - U_j(1 - e^{-\lambda_j}))$, $S' = (S'_1, \ldots, S'_{k-2})$, tais que S'_j são variáveis aleatórias independentes TExp(λ_j).
   Este método usa $(k-2)$ exponenciais truncadas para gerar $(k-1)$ vetores com distribuição Bingham complexa. Suponha que os autovalores da matriz A são $\tilde{\lambda}_1 < \ldots < \tilde{\lambda}_{k-2} < \tilde{\lambda}_{k-1} = 0$, e escreva $\lambda_j = \tilde{\lambda}_{k-1} - \tilde{\lambda}_j$, $j = 1, \ldots, k-2$. O vetor de entrada é
   $$\tilde{\lambda} = (\lambda_1, \ldots, \lambda_{k-2}).$$
3. Se $\sum_{j=1}^{k-2} S'_j < 1$, escreva $S'_{k-1} = 1 - \sum_{j=1}^{k-2} S'_j$. Caso contrário, volte ao passo 1.
4. Gerar ângulos independentes $\theta_j \sim U[0, 2\pi)$, $j = 1, \ldots, k-1$.
5. Calcular $z_j = {S'_j}^{1/2} e^{i\theta_j}$, $j = 1, \ldots, k-1$.

O algoritmo fornece um vetor $z = (z_1, \ldots, z_{k-1})^T$, que possui uma distribuição Bingham complexa.

CAPÍTULO 3

Influência

Os métodos diagnósticos são úteis para encontrar inadequações em um modelo geral. A detecção de observações influentes é um objetivo importante na análise de diagnóstico, pois, pontos que apresentam um peso significativo na estimava dos parâmetros, proporcionam inferências erradas sob um modelo ajustado de forma incorreta.

Cook (1977) propõe uma medida para modelos de regressão linear de resposta normal, que avalia o impacto de retirar uma observação particular nas estimativas da regressão. Esta técnica foi desenvolvida para outros tipos de modelos. Roos (1987) discute a geometria da deleção em regressão não-linear, em quanto que, Cook, Peña e Weisberg (1988) comparam o afastamento da verossimilhança com medidas tradicionais de deleção de observações tais como Cook e o DFFITS$_i$.

Deixar de detectar pontos conjuntamente discordantes é um problema que pode surgir na deleção individual de observações. Neste caso, Cook (1986) sugere avaliar a influência conjunta das observações introduzindo uma pequena perturbação no modelo, ao invés de avaliar pela retirada de uma observação do conjunto de dados.

No caso de análise de formas o cálculo da forma média é de grande importância. Indivíduos influentes no conjunto de dados podem levar a estimativas inadequadas da forma média, e assim, testes utilizados na análise estatística de formas tais como T^2 de Hotelling e Goodall podem apresentar resultados errados [veja Dryden & Mardia (1998)]. Neste capítulo iremos introduzir medidas de influência relevantes para esta dissertação.

3.1 Modelo de Deleção de Casos

Detectar uma observação influente, é um passo natural o exame dos efeitos da exclusão da observação, no entanto, o problema de determinar qual é o ponto ou pontos para excluir pode ser uma tarefa difícil. A metodologia proposta por Cook (1977) para modelos lineares de resposta normal, avalia por meio de uma medida apropriada o efeito da exclusão de uma observação da análise nas estimativas dos parâmetros do modelo. Esta medida envolve a diferença do vetor $(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)})$, sendo $\hat{\boldsymbol{\theta}}_{(i)}$ a estimativa de máxima verossimilhança de $\boldsymbol{\theta}$ quando a i-ésima observação é excluída do modelo.

Assuma que a i-ésima resposta y_i é uma função conhecida g_i com vetor de parâmetros desconhecidos $\boldsymbol{\theta}$ e um erro não observável ε_i

$$y_i = g_i(\boldsymbol{\theta}, \varepsilon_i) \quad i = 1, \ldots, n. \tag{3.1}$$

Mediante a medida chamada afastamento da verossimilhança, se estuda a influencia da i-ésima observação na estimativa de $\boldsymbol{\theta}$. O afastamento da verossimilhança é definido da seguinte maneira:

$$LD(\hat{\boldsymbol{\theta}}_{(i)}) = 2[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_{(i)})], \tag{3.2}$$

onde $L(\boldsymbol{\theta})$ é o logaritmo da função de verossimilhança de $\boldsymbol{\theta}$. Considerando que para um n grande, este simples cálculo pode ser computacionalmente caro. Pode-se obter uma aproximação quadrática de $L_{(i)}(\boldsymbol{\theta})$, sendo esta última a log verossimilhança obtida após da exclusão da i-ésima observação. Desenvolvendo $L_{(i)}(\boldsymbol{\theta})$ em Taylor até segunda ordem em torno de $\hat{\boldsymbol{\theta}}$, obtemos

$$L_{(i)}(\boldsymbol{\theta}) \approx L_{(i)}(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \dot{\mathbf{L}}_{(i)}(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \ddot{\mathbf{L}}_{(i)}(\boldsymbol{\theta})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{3.3}$$

em que $\dot{\mathbf{L}}_{(i)}(\boldsymbol{\theta}) = \partial L_{(i)}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ e $\ddot{\mathbf{L}}_{(i)} = \partial^2 L_{(i)}(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$. Se $-\ddot{\mathbf{L}}_{(i)}(\boldsymbol{\theta})$ é positiva definida, a aproximação quadrática é maximizada em

$$\hat{\boldsymbol{\theta}}^{\mathbf{1}}_{(i)} = \hat{\boldsymbol{\theta}} - \left[(\ddot{\mathbf{L}}_{(i)}(\boldsymbol{\theta}))^{-1} \dot{\mathbf{L}}_{(i)}(\hat{\boldsymbol{\theta}}) \right] \tag{3.4}$$

O cálculo $\hat{\boldsymbol{\theta}}^{\mathbf{1}}_{(i)}$ refere-se aproximação a um passo de $\hat{\boldsymbol{\theta}}_{(i)}$, assim, se $\hat{\boldsymbol{\theta}}_{(i)}$ não é muito diferente de $\hat{\boldsymbol{\theta}}$ e $L_{(i)}(\boldsymbol{\theta})$ é localmente quadrática, a aproximação a um passo deveria estar próxima do valor de $\hat{\boldsymbol{\theta}}$ [Cook & Weisberg (1982, p. 183)].

3.1.1 Distância de Cook

Considere-se $L(\boldsymbol{\theta})$ a log verossimilhança baseada nos dados completos. O afastamento da verossimilhança $LD(\hat{\boldsymbol{\theta}}_{(i)})$ é definido como em (3.2) ou utilizando o estimador a um passo

$$LD(\hat{\boldsymbol{\theta}}^{\mathbf{1}}_{(i)}) = 2[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}^{\mathbf{1}}_{(i)})].$$

As medidas $LD(\hat{\boldsymbol{\theta}}_{(i)})$ e $LD(\hat{\boldsymbol{\theta}}^{\mathbf{1}}_{(i)})$ também podem ser interpretadas em termos da região de confiança assintótica

$$\{\boldsymbol{\theta} : 2[L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta})] \leq \chi^2(p, \alpha)\}.$$

Esta é região de confiança de $100(1-\alpha)\%$ para $\boldsymbol{\theta}$, onde p é a dimensão do vetor de parâmetros $\boldsymbol{\theta}$. Portanto, $LD(\hat{\boldsymbol{\theta}}_{(i)})$ pode ser ajustado por comparação com a distribuição $\chi^2(\alpha)$.

Supondo que $LD(\hat{\boldsymbol{\theta}}_{(i)})$ pode ser bem representada por uma função quadrática, $LD(\hat{\boldsymbol{\theta}}_{(i)})$ usualmente pode ser aproximada pela expansão de Taylor até segunda ordem de $L(\hat{\boldsymbol{\theta}}_{(i)})$ ao redor de $\hat{\boldsymbol{\theta}}$,

$$L(\hat{\boldsymbol{\theta}}_{(i)}) \cong L(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}_{(i)} - \hat{\boldsymbol{\theta}})^T \dot{\mathbf{L}}(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T [\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}). \tag{3.5}$$

Deste modo, substituindo (3.5) na equação (3.2) e, desde que $\dot{\mathbf{L}}(\hat{\theta}) = 0$, pode-se obter uma aproximação de $LD(\hat{\boldsymbol{\theta}}_{(i)})$. Esta aproximação, denominada $DG(\hat{\boldsymbol{\theta}}_{(i)})$, assume a seguinte expressão

$$DG(\hat{\boldsymbol{\theta}}_{(i)}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T [-\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}).$$

A expressão acima pode ser enunciada de uma maneira mais geral, a qual tem sido chamada de *Distância de Cook generalizada*

$$DG(\hat{\boldsymbol{\theta}}_{(i)}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T \mathbf{C} (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \tag{3.6}$$

onde $\mathbf{C}$ é uma matriz positiva definida e assintoticamente equivalente à matriz de informação esperada de $\boldsymbol{\theta}$. Se consideram dois casos para $\mathbf{C}$, $-\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})$ e $\mathbf{K}(\boldsymbol{\theta})$, as matrizes de informação observada e esperada de Fisher, respectivamente.

Da forma similar a Cook (1977), notamos que é possível avaliar a magnitude de $DG((\hat{\theta}_{(i)}))$ observando que, assintoticamente

$$\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T \mathbf{C} (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}) \leq \mathcal{X}^2_{p,\alpha}\} \tag{3.7}$$

é a region de confiança de $100(1-\alpha)\%$ para $\boldsymbol{\theta}$, em que p é a dimensão do vetor de parâmetros $\boldsymbol{\theta}$. A estatística $DG(\hat{\boldsymbol{\theta}}_{(i)})$ pode ser vista como uma base útil para a detecção de casos que devem ser cuidadosamente examinados para grande erros. A procura de um grande erro deve necessariamente forçar a remoção ou correção do processo correspondente, e estas ações podem provocar uma mudança substancial nos resultados de uma análise caso $DG(\hat{\boldsymbol{\theta}}_{(i)})$ seja grande [Cook (1986, p. 135)].

3.2 Influência Local

O método de influência local proposto por Cook (1986) consiste em estudar o comportamento de alguma medida particular de influência segundo pequenas perturbações nos dados ou no modelo. A

existência de pontos que sob pequenas modificações no modelo causam variações desproporcionais nos resultados [Paula $(2004,$ p . $40)$].

Para um conjunto de dados observados, seja $L(\boldsymbol{\theta})$ o logaritmo da função de verossimilhança de $\boldsymbol{\theta}$. Seja $L(\boldsymbol{\theta}|\boldsymbol{w})$ o logaritmo da função de verossimilhança perturbada, em que $\boldsymbol{w}$ é um vetor $q \times 1$, $\boldsymbol{w} \in \Omega \subset \mathbb{R}^q$. Existe $\boldsymbol{w}_0$ (vetor de não perturbação) tal que $L(\boldsymbol{\theta}|\boldsymbol{w}_0) = L(\boldsymbol{\theta})$. Geralmente, $\boldsymbol{w}_0$ pode ser considerado como qualquer esquema de perturbação bem definido, e portanto, não está restrito a ser uma coleção de pesos. Algumas das formas mas comuns de perturbação são:

1. Perturbação de casos: $L(\boldsymbol{\theta}|\boldsymbol{w}) = \sum_{i=1}^{n} w_i L(\boldsymbol{\theta})$, $0 \leq w_i \leq 1$.
2. Perturbação na resposta (alavancagem): $y_{iw} = y_i \sigma_{y_i} w_i$, $w_i \in \mathbb{R}$.
3. Perturbação em x_i (contínua): $x_{iw} = x_i + \sigma_{x_i} w_i$, $w \in \mathbb{R}$.
4. Perturbação na matriz de variância-covariância: $\Sigma_{iw} = w_i^{-1}\Sigma_i$, $w_i \in \mathbb{R} - \{0\}$.

Sejam $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{(w)}$ denotando os estimadores de máxima verossimilhança sob $L(\boldsymbol{\theta})$ e $L(\boldsymbol{\theta}|\boldsymbol{w})$, respectivamente, e assuma que $L(\boldsymbol{\theta}|\boldsymbol{w})$ é duas vezes diferenciável em $(\boldsymbol{\theta}^T, \boldsymbol{w}^T)$. Para avaliar o efeito das perturbações nas estimativas fornecidas pelo modelo o afastamento da verossimilhança é definido por

$$LD(\hat{\boldsymbol{\theta}}_{(w)}) = 2[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_{(w)}].$$

Note que $LD(\hat{\boldsymbol{\theta}}_{(w_0)}) = 0$. A idéia de influência local é estudar o comportamento da função $LD(\hat{\boldsymbol{\theta}}_{(w)})$ numa vizinhança de $\boldsymbol{w}_0$. Considere a superfície geométrica formada pelos valores do vetor

$$\boldsymbol{\alpha}(\boldsymbol{w}) = \begin{pmatrix} \boldsymbol{w} \\ LD(\boldsymbol{\theta}_{(w)}) \end{pmatrix}$$

O processo consiste em escolher um direção unitária $\mathbf{d}$, $\|\mathbf{d}\| = 1$, e assim, considerar o gráfico de $LD(\hat{\boldsymbol{\theta}}_{(w_0+t\mathbf{d})})$ e t, $t \in \mathbb{R}$. Este gráfico é chamado *linha projetada*. A curvatura normal da linha projetada, denotada por $C_{\mathbf{d}}$ é definida como sendo a curvatura de $(t, LD(\hat{\boldsymbol{\theta}}_{(w_0+t\mathbf{d})})$, em $t = 0$. Denomina-se $C_{\mathbf{d}}$ a curvatura normal da superfície $\boldsymbol{\alpha}(\boldsymbol{w})$ em $\boldsymbol{w}_0$ e na direção $\mathbf{d}$. Considere

$$C_d = LD(\hat{\boldsymbol{\theta}}_{(w_0+t\mathbf{d})})\Big|_{t=0} = \frac{\partial^2 LD(\hat{\boldsymbol{\theta}}_{(w_0+t\mathbf{d})})}{\partial t^2}\Big|_{t=0}. \tag{3.8}$$

Cook (1986) mostra que a curvatura normal na direção unitária $\mathbf{d}$ é dada por

$$C_{\mathbf{d}}(\hat{\boldsymbol{\theta}}) = 2\|\mathbf{d}^T \Delta^T \ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}}) \Delta \mathbf{d}\|, \tag{3.9}$$

em que $-\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}})$ é a matriz observada de Fisher e Δ é uma matriz $r \times q$ com elementos

$$\Delta_{ij} = \frac{\partial^2 L(\boldsymbol{\theta}|\boldsymbol{w})}{\partial \theta_i \partial w_j}$$

avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\boldsymbol{w} = \boldsymbol{w}_0$.

O objetivo está em achar as direções que produzem a maior influência local. A direção de maior curvatura normal, denotada por $\mathbf{d}_{\max}$, é o autovetor normalizado correspondente ao maior autovalor $C_{\mathbf{d}_{\max}}$ da matriz

$$A = \Delta^T(\ddot{\mathbf{L}}(\hat{\boldsymbol{\theta}}))^{-1}\Delta.$$

Os valores de $\mathbf{d}_{\max}$ contem a influência local das observações nessa direção particular. Logo, o gráfico de $|\mathbf{d}_{\max}|$ contra a ordem das observações pode revelar aqueles pontos com maior influência. Tais observações podem ser responsáveis por mudanças nas estimativas dos parâmetros sob pequenas perturbações no modelo.

CAPÍTULO 4

Influência em Análise Estatística de Formas

Na análise estatística de formas, as formas dos objetos são descritos por um número finito de pontos que são chamados de marcos anatômicos. Os marcos anatômicos são pontos de correspondência em cada objeto. O espaço de marcos anatômicos é um espaço no conjunto dos reais $\mathbb{R}^m$ onde são representadas as coordenadas cartesianas de um marco. Nesta dissertação se consideram objetos bidimensionais ($m = 2$), assim o espaço de marcos anatômicos é definido em $\mathbb{R}^2$. Quando $m = 2$ as coordenadas cartesianas podem ser escritas como vetores complexos, que chamamos de configurações complexas ou coordenadas complexas dos marcos anatômicos .

Transformações apropriadas são usadas para remover os efeitos de escala e rotação de um objeto. O primeiro passo é remover a locação. Para o sistema de coordenadas de Kendall, a locação das configurações complexas é removida multiplicando pela sub-matriz de Helmert. Ao produto das configurações complexas com a sub-matriz de Helmert chamaremos de configurações helmertizadas. A escala é removida da configuração helmertizada dividindo esta pela do mesmo. O vetor resultante é chamado de pré-forma da configuração complexa. Aqui a pré-forma é uma forma com a informação de rotação conservada. As pré-formas pertencem a um novo espaço chamado de espaço de pré-formas mostrado em $\mathbb{C}S^{k-2}$ definido em (2.5).

O espaço tangente é a versão linear do espaço de formas na proximidade de um ponto particular do espaço de formas. Suponha uma projeção tangente para a esfera de pré-formas que não depende da rotação da figura original e portanto, um sistema adequado de coordenadas tangentes para formas. Aqui, consideraremos as coordenadas tangentes Procrustes. Estas coordenadas tangentes da esfera de pré-formas pertencem ao espaço tangente $\mathbb{R}^{2k-2}$.

Serão indicadas três medidas de influência neste capítulo. Duas medidas no espaço de pré-formas $\mathbb{C}S^{k-2}$, na primeira medida, a distância de Cook é adaptada ao espaço de pré-formas, na segunda o teste de discordância é adaptado ao espaço de pré-formas, estas duas medidas são definidas para dados con distribuição Bingham complexa. A ultima medida é utilizada no espaço tangente, que é a medida de influência local em dados normais multivariados.

4.1 Estimação da Distância de Cook para a Distribuição Bingham Complexa

Seja a função de log verossimilhança da distribuição Bingham complexa como

$$L = \sum_{j=1}^{k-1} l_j \lambda_j - n \log c(\Lambda),$$

onde $c(\Lambda) = 2\pi^{k-1} \sum_{j=1}^{k-1} a_j e^{\lambda_j}$ e $a_j^{-1} = \prod_{i \neq j}(\lambda_j - \lambda_i)$. A primeira derivada da função de log verossimilhança com respeito a λ_m é dada por:

$$u_m = \frac{\partial L}{\partial \lambda_m} = l_m - \frac{n}{c(\Lambda)} c_m(\Lambda), \text{ onde } c_m = \frac{\partial c(\Lambda)}{\partial \lambda_m}$$

assim

$$c_m(\Lambda) = \frac{\partial c(\Lambda)}{\partial \lambda_m} = \frac{\partial}{\partial \lambda_m} \left[2\pi^{k-1} \sum_{j=1}^{k-1} a_j e^{\lambda_j} \right] = 2\pi^{k-1} \sum_{j=1}^{k-1} \left[\frac{\partial a_j e^{\lambda_j}}{\partial \lambda_m} \right].$$

Agora precisamos calcular $\partial a_j / \partial \lambda_m$, vamos dividir em dois casos, caso $m \neq j$

$$\frac{\partial a_j}{\partial \lambda_m} = \frac{1}{\left[\prod_{i \neq j} (\lambda_j - \lambda_i) \right]^2} \prod_{i \neq j, m} (\lambda_j - \lambda_i) = \frac{1}{(\lambda_j - \lambda m) \prod_{i \neq j} (\lambda_j - \lambda_i)} = \frac{a_j}{\lambda j - \lambda m}, \tag{4.1}$$

e caso $m = j$

$$\begin{aligned} \frac{\partial a_m}{\partial \lambda_m} &= -\frac{1}{\left[\prod_{i \neq m} (\lambda_m - \lambda_i) \right]^2} \prod_{i \neq m, 1} (\lambda_m - \lambda_i) + \prod_{i \neq m, 2} (\lambda_m - \lambda_i) + \ldots + \prod_{i \neq m, k-1} (\lambda_m - \lambda_i) \\ &= -\frac{1}{\prod_{i \neq m} (\lambda_m - \lambda i)} \sum_{i \neq m} \frac{1}{(\lambda_m - \lambda_i)} = -a_m b_m, \end{aligned} \tag{4.2}$$

onde $b_j = \sum_{i \neq j} \frac{1}{\lambda_j - \lambda_i}$.

Logo

$$c_m(\Lambda) = 2\pi^{k-1} \sum_{j=1}^{k-1} \left[\frac{\partial a_j e^{\lambda_j}}{\partial \lambda_m} \right] = 2\pi^{k-1} \left[\sum_{j \neq m}^{k-1} \frac{a_j e^{\lambda_j}}{(\lambda_j - \lambda_m)} + a_m e^{\lambda_m} - a_m b_m e^{\lambda_m} \right]. \tag{4.3}$$

Os termos da segunda derivada da log verossimilhança são dados por

$$u_{m,s} = \frac{n}{[c(\Lambda)]^2}\left\{c_m(\Lambda)c_s(\Lambda) - c_{m,s}(\Lambda)c(\Lambda)\right\}, \text{ onde } c_{m,s}(\Lambda) = \frac{\partial^2 c_m(\Lambda)}{\partial\lambda_m\partial\lambda_s},$$

as derivadas $c_m(\Lambda)$ e $c_s(\Lambda)$ são calculadas em (4.3), para as segundas derivadas é necessário só calcular a derivada $c_{m,s}(\Lambda)$. Primeiro vamos calcular a derivada $c_{m,s}(\Lambda)$ para o caso $s \neq m$.

$$\begin{aligned} c_{m,s}(\Lambda) &= \frac{\partial}{\partial\lambda_s}\left\{2\pi^{k-1}\left[\sum_{j\neq m}^{k-1}\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)} + a_m e^{\lambda_m} - a_m b_m e^{\lambda_m}\right]\right\} \\ &= 2\pi^{k-1}\left[\sum_{j\neq m}^{k-1}\frac{\partial}{\partial\lambda_s}\left[\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)}\right] + \frac{\partial a_m e^{\lambda_m}}{\partial\lambda_s} - \frac{\partial a_m b_m e^{\lambda_m}}{\partial\lambda_s}\right]. \end{aligned}$$

Aqui a derivada $c_{m,s}(\Lambda)$ é divida em três partes. Primeiro $\partial(a_j e^{\lambda_j})/\partial(\lambda_j - \lambda_m)$, e esta na vez será dividida em dois caso,
caso $s \neq j$

$$\begin{aligned} \frac{\partial}{\partial\lambda_s}\left[\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)}\right] &= \frac{e^{\lambda_j}}{(\lambda_j-\lambda_m)}\frac{\partial a_j}{\partial\lambda_s} = \frac{e^{\lambda_m}}{(\lambda_j-\lambda_m)}\frac{a_j}{(\lambda_j-\lambda_s)} \qquad \text{(devido a (4.1))} \\ &= \frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)(\lambda_j-\lambda_s)}, \end{aligned}$$

e caso $s = j$

$$\begin{aligned} \frac{\partial}{\partial\lambda_s}\left[\frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)}\right] &= \frac{e^{\lambda_s}}{(\lambda_s-\lambda_m)}\frac{\partial a_s}{\partial\lambda_s} + a_s e^{\lambda_s}\frac{\partial}{\partial\lambda_s}\left[\frac{1}{(\lambda_s-\lambda_m)}\right] + \frac{a_s}{(\lambda_s-\lambda_m)}\frac{\partial e^{\lambda_s}}{\partial\lambda_s} \\ &= \frac{e^{\lambda_s}}{(\lambda_s-\lambda_m)}(-a_s b_s) + a_s e^{\lambda_s}\left[-\frac{1}{(\lambda_s-\lambda_m)^2}\right] + \frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)} \qquad \text{(devido a (4.2))} \\ &= -\frac{a_s b_s e^{\lambda_s}}{(\lambda_s-\lambda_m)} - \frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)^2} + \frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)}. \end{aligned}$$

A segunda parte da derivada de $c_{m,s}(\Lambda)$, é dada por

$$\frac{\partial a_m e^{\lambda_m}}{\partial\lambda_s} = e^{\lambda_m}\frac{\partial a_m}{\partial\lambda_s} = \frac{a_m e^{\lambda_m}}{(\lambda_m-\lambda_s)}. \qquad \text{(devido a (4.1))}$$

E por ultimo a terceira parte da derivada de $c_{m,s}(\Lambda)$ é

$$\begin{aligned} \frac{\partial a_m b_m e^{\lambda_m}}{\partial\lambda_s} &= b_m e^{\lambda_m}\frac{\partial a_m}{\partial\lambda_s} + a_m e^{\lambda_m}\frac{\partial b_m}{\partial\lambda_s} = \frac{a_m b_m e^{\lambda_m}}{(\lambda_m-\lambda s)} + a_m e^{\lambda_m}\frac{\partial}{\partial\lambda_s}\left[\sum_{i\neq m}\frac{1}{(\lambda_m-\lambda_i)}\right] \\ &\qquad \text{(devido a (4.1))} \\ &= \frac{a_m b_m e^{\lambda_m}}{(\lambda_m-\lambda s)} + a_m e^{\lambda_m}\left[\frac{1}{(\lambda_m-\lambda_s)^2}\right] = \frac{a_m b_m e^{\lambda_m}}{(\lambda_m-\lambda s)} + \frac{a_m e^{\lambda_m}}{(\lambda_m-\lambda_s)^2}. \end{aligned}$$

Assim temos que para $s \neq m$:

$$
\begin{aligned}
c_{m,s}(\Lambda) &= \frac{\partial}{\partial \lambda_s}\left\{2\pi^{k-1}\left[\sum_{j\neq m}^{k-1}\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)} + a_m e^{\lambda_m} - a_m b_m e^{\lambda_m}\right]\right\} \\
&= 2\pi^{k-1}\Bigg\{\sum_{j\neq m,s}^{k-1}\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)(\lambda_j-\lambda_s)} - \frac{a_s b_s e^{\lambda_s}}{(\lambda_s-\lambda_m)} + \frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)} - \frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)^2} \\
&- \frac{a_m b_m e_m^{\lambda}}{(\lambda_m-\lambda_s)} + \frac{a_m e^{\lambda_m}}{(\lambda_m-\lambda_s)} - \frac{a_m e^{\lambda_m}}{(\lambda_m-\lambda_s)^2}\Bigg\} \\
&= 2\pi^{k-1}\Bigg\{\sum_{j\neq m,s}^{k-1}\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)(\lambda_j-\lambda_s)} - \left[\frac{a_s b_s e^{\lambda_s}}{(\lambda_s-\lambda_m)} + \frac{a_m b_m e_m^{\lambda}}{(\lambda_m-\lambda_s)}\right] \\
&+ \left[\frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)} + \frac{a_m e^{\lambda_m}}{(\lambda_m-\lambda_s)}\right] - \left[\frac{a_s e^{\lambda_s}}{(\lambda_s-\lambda_m)^2} + \frac{a_m e^{\lambda_m}}{(\lambda_m-\lambda_s)^2}\right]\Bigg\}. \qquad (4.4)
\end{aligned}
$$

A derivada $c_{m,s}(\Lambda)$ para o caso $m = s$ (denotado $c_{m,m}(\Lambda)$) é dada por:

$$
\begin{aligned}
c_{m,m}(\Lambda) &= \frac{\partial}{\partial \lambda_m}\left\{2\pi^{k-1}\left[\sum_{j\neq m}^{k-1}\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)} + a_m e^{\lambda_m} - a_m b_m e^{\lambda_m}\right]\right\} \\
&= 2\pi^{k-1}\left[\sum_{j\neq m}^{k-1}\frac{\partial}{\partial \lambda_m}\left[\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)}\right] + \frac{\partial a_m e^{\lambda_m}}{\partial \lambda_m} - \frac{\partial a_m b_m e^{\lambda_m}}{\partial \lambda_m}\right].
\end{aligned}
$$

A derivada $c_{m,m}(\Lambda)$ também será dividida em três partes. A primeira parte $\partial a_j e^{\lambda_j}/\partial(\lambda_j - \lambda_m)$ é

$$
\begin{aligned}
\frac{\partial}{\partial \lambda_m}\left[\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)}\right] &= \frac{e^{\lambda_j}}{(\lambda_j-\lambda_m)}\frac{\partial a_j}{\partial \lambda_m} + a_j e^{\lambda_j}\frac{\partial}{\partial \lambda_m}\left[\frac{1}{(\lambda_j-\lambda_m)}\right] \\
&= \frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)^2} + \frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)^2} = 2\left\{\frac{a_j e^{\lambda_j}}{(\lambda_j-\lambda_m)^2}\right\}, \qquad \text{(devido a 4.1)}
\end{aligned}
$$

a segunda parte é dada por

$$
\begin{aligned}
\frac{\partial a_m e^{\lambda_m}}{\partial \lambda_m} &= e^{\lambda_m}\frac{\partial a_j}{\partial \lambda_m} + a_j\frac{\partial e^{\lambda_m}}{\partial \lambda_m} \\
&= -a_m b_m e^{\lambda_m} + a_m e^{\lambda_m}, \qquad \text{(devido a (4.2))}
\end{aligned}
$$

e finalmente a terceira parte é

$$
\begin{aligned}
\frac{\partial a_m b_m e^{\lambda_m}}{\partial \lambda_m} &= b_m e^{\lambda_m}\frac{\partial a_m}{\partial \lambda_m} + a_m e^{\lambda_m}\frac{\partial b_m}{\partial \lambda_m} + a_m b_m\frac{\partial e^{\lambda_m}}{\partial \lambda_m} \\
&= -a_m b_m^2 e^{\lambda_m} + a_m e^{\lambda_m}\sum_{j\neq m}\left[\frac{-1}{(\lambda_m-\lambda_j)^2}\right] + a_m b_m e^{\lambda_m} \qquad \text{(devido a (4.2))} \\
&= -a_m b_m^2 e^{\lambda_m} + a_m e^{\lambda_m}\left[-\sum_{j\neq m}\frac{1}{(\lambda_m-\lambda_j)^2}\right] + a_m b_m e^{\lambda_m} \\
&= -a_m b_m^2 e^{\lambda_m} - a_m g_m e^{\lambda_m} + a_m b_m e^{\lambda_m},
\end{aligned}
$$

onde $g_m = \sum_{j \neq m} \frac{1}{(\lambda_m - \lambda_j)^2}$.
Assim temos

$$
\begin{aligned}
c_{m,m}(\Lambda) &= \frac{\partial}{\partial \lambda_m} \left\{ 2\pi^{k-1} \left[\sum_{j \neq m}^{k-1} \frac{a_j e^{\lambda_j}}{(\lambda_j - \lambda_m)} + a_m e^{\lambda_m} - a_m b_m e^{\lambda_m} \right] \right\} \\
&= 2\pi^{k-1} \left\{ 2 \sum_{j \neq m}^{k-1} \frac{a_j e^{\lambda_j}}{(\lambda_j - \lambda_m)^2} - a_m b_m e^{\lambda_m} + a_m e^{\lambda_m} + a_m b_m^2 e^{\lambda_m} + a_m g_m e^{\lambda_m} - a_m b_m e^{\lambda_m} \right\} \\
&= 2\pi^{k-1} \left\{ 2 \sum_{j \neq m}^{k-1} \frac{a_j e^{\lambda_j}}{(\lambda_j - \lambda_m)^2} - 2a_m b_m e^{\lambda_m} + a_m e^{\lambda_m} + a_m b_m^2 e^{\lambda_m} + a_m d_m e^{\lambda_m} \right\}.
\end{aligned}
\tag{4.5}
$$

Assim a matriz de informação de Fisher $\ddot{\mathbf{L}}$ para distancia de Cook quando os dados tem distribuição Bingham complexa é dado por:

$$
\mathfrak{Lb} = \begin{cases} u_{m,s} = \frac{n}{[c(\Lambda)]^2} \left\{ c_m(\Lambda) c_s(\Lambda) - c_{m,s}(\Lambda) c(\Lambda) \right\} & \text{se } m \neq s \\ u_{m,s} = \frac{n}{[c(\Lambda)]^2} \left\{ (c_m(\Lambda))^2 - c_{m,m}(\Lambda) c(\Lambda) \right\} & \text{se } m = s, \end{cases}
\tag{4.6}
$$

onde $c_m(\Lambda)$, $c_{m,s}(\Lambda)$ e $c_{m,m}(\Lambda)$ estão definidas em (4.3), (4.4) e (4.5) respectivamente.

O cálculo da distância de Cook para dados com distribuição Bingham complexa será

$$
DG(\hat{\boldsymbol{\theta}}_{(i)}) = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T [-\mathfrak{Lb}(\hat{\boldsymbol{\theta}})] (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}).
\tag{4.7}
$$

onde $\hat{\boldsymbol{\theta}}_{(i)}$ é o vetor estimado de parâmetros de tamanho $(k-1) \times 1$ da distribuição Bingham complexa obtido após da exclusão da $i-$ésima observação.

4.2 Método de Bootstrap para Identificar Pontos Influentes Usando a Distância de Cook

O termo bootstrap refere-se a uma classe de procedimentos para estimação de parâmetros em geral e taxas de erro em particular. Foi introduzido por Efron (1979) e desde então tem recebido bastante atenção dos estatísticos [Efron & Tibshirani (1993) e Davison & Hinkley (1997)]. Suponha que dispomos de uma amostra de treinamento $z = (z_1, \ldots, z_n)$. A idéia básica é retirar aleatoriamente amostras com reposição, cada amostra com tamanho igual ao do conjunto de treinamento. Este procedimento é repetido B vezes (B suficientemente grande, isto é, B maior a 200), produzindo B amostras bootstrap.

Desejamos encontrar o percentil $100(1-\alpha)$ da distância de Cook $DG(\boldsymbol{\theta}_{(i)})$ calculada em nosso conjunto padrão. Retiramos B amostras bootstrap $Z^{*(1)}, \ldots, Z^{*(B)}$ cada uma com o mesmo

tamanho do conjunto de treinamento original e calculamos o valor da distância de Cook em cada uma das B amostras. Os valores da distância de Cook para $DG(\boldsymbol{\theta}_{(\boldsymbol{i_1})}), \ldots, DG(\boldsymbol{\theta}_{(\boldsymbol{i_B})})$ são utilizados para encontrar o percentil $100(1-\alpha)\%$ da distância de Cook $DG(\boldsymbol{\theta}_{(\boldsymbol{i})})$, sendo a media dos percentiles $100(1-\alpha)\%$ de $DG(\boldsymbol{\theta}_{(\boldsymbol{i_1})}), \ldots, DG(\boldsymbol{\theta}_{(\boldsymbol{i_B})})$.

4.3 Teste para Discordância

As observações que na opinião do investigador, encontra-se afastadas da maior parte dos dados são chamados de *outliers*. A identificação de outliers consiste na detecção, com métodos subjetivos, das observações potencialmente anormais. A identificação é feita, geralmente, por análise gráfica ou, no caso de o número de dados ser pequeno, por observação direta dos mesmos.

Em muitos casos as razões da existência de *outliers* 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. Na literatura é comum encontrar a palavra discordante como um sinônimo de *outlier*. Um método apropriado para identificar outliers é fazendo um teste de discordância.

Métodos utilizados para eixos reais podem ser adaptados para eixos complexos em pré-formas. Uma referência importante para este fato é Amaral, G., Dryden, I. & Wood, A., (2007). Aqui o teste para discordância em dados com distribuição Watson real é adaptado ser um teste para discordância em dados com distribuição Bingham complexa [Fisher, N. I., Lewis, T. & Embleton, B. J. (2004)].

Seja $z_1, \ldots, z_n$ as pré-formas de $z_1^0, \ldots, z_n^0$ com distribuição Bingham complexa, onde z_i esta definida em (2.4) e $z_i \in \mathbb{C}B_{k-2}(A)$. Denota-se S_n a matriz complexa de somas de quadrado e produtos *SSP* definida em (2.7) calculada usando a amostra de pré-formas completa, e S_{m-} a matriz *SSP* calculada omitindo a pré-forma z_m, assim

$$S_{m-} = S_n - z_m z_m^*.$$

Seja $\hat{\tau}_{1,(n)}$ e $\hat{\tau}_{1,(m-)}$ os menores autovalores de $S_{(n)}$ e $S_{(m-)}$ respectivamente, e calcula-se:

$$J_m = (n-(k-1))\frac{(\hat{\tau}_{1,(n)} - \hat{\tau}_{1,(m-)})}{\hat{\tau}_{1,(m-)}} \tag{4.8}$$

onde k é o número de marcos anatômicos. Assim, o individuo é julgado como discordante se J_m é muito grande.

4.4 Influência Local em Dados Normais Multivariados

Um outlier é uma observação que é numericamente distante do restante dos dados. Inferências derivadas de conjuntos de dados que incluem *outliers* podem levar a conclusões erradas sobre os

dados. A distância de Mahalanobis definida como

$$(\boldsymbol{x}_i - T(\boldsymbol{X}))^T C(\boldsymbol{X})^{-1}(\boldsymbol{x}_i - T(\boldsymbol{X}))$$

é calculada para cada observação $\boldsymbol{x}_i$ e é utilizada para detecção de *outliers*. Aqui, $T(\boldsymbol{X})$ é vetor de médias amostral, e $C(\boldsymbol{X})$ é a matriz de covariância amostral. O método baseado na distância de Mahalanobis funciona bem na detecção de um único *outlier*, mas existem casos nos quais alguns *outliers* não têm necessariamente uma grande distância de Mahalanobis. O método de influência local (veja sessão $\mathbf{3.2}$) introduzido por Cook (1986) funciona como um método para avaliar a influência de uma mínima perturbação em um modelo. Kim (1996), adapta este método de influência local para dados com distribuição normal multivariada. O método de influência local permite fazer perturbações simultâneas que afetam todas as observações. Sabe-se que a matriz de covariância é mais sensível a outliers, portanto, a perturbação é escolhida para colocar peso sobre a matriz de covariância para cada observação.

4.4.1 Influência Local

Considere $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ uma amostra aleatória de tamanho n extraída de uma distribuição normal p-variada $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (modelo não perturbado), onde $\boldsymbol{\mu}$ é o vetor de médias, e $\boldsymbol{\Sigma}$ é a matriz de covariância desconhecida. Seja $\boldsymbol{w} = \{w_1, \ldots, w_n\}$ o vetor de pertubações de tamanho $n \times 1$. Considera-se o modelo perturbado em que a j-ésima observação $\boldsymbol{x}_j$ é perturbada de acordo com

$$\boldsymbol{x}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/w_j)$$

para cada j. Quando o termo w_j é igual a 1, o modelo perturbado se reduz ao modelo não perturbado.

A matriz $\boldsymbol{\Sigma}$ será reparametrizada tal que $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$, onde $\mathbf{A} = a_{ij}$ é uma matriz triangular inferior com os elementos da diagonal positivos. Seja $\mathbf{B}^T = \mathbf{A}^{-1}$, então, $\mathbf{B} = (b_{ij})$ é uma matriz triangular superior e $\boldsymbol{\Sigma}^{-1} = \mathbf{B}\mathbf{B}^T$. Defina-se $\boldsymbol{\theta}$ o vetor $[p(p+3)/2] \times 1$ de parâmetros formados pelo empilhamento de $\boldsymbol{\mu}$ e os elementos diferentes de zero das colunas $\mathbf{b}_i$ de $\mathbf{B}$ desde a primeira coluna até a p-ésima coluna. Sob o modelo não perturbado, o **MLE** de $\boldsymbol{\mu}$ é $\bar{\boldsymbol{x}}$, e o **MLE** de $\boldsymbol{\Sigma}$ é a matriz de covariâncias amostral **S**. O princípio de invariância de **MLE** garante que o **MLE** de $\mathbf{A}$ é a raiz de Cholesky de $\mathbf{S}$, ou seja, $\mathbf{S} = \hat{\mathbf{A}}\hat{\mathbf{A}}^T$. Além disso, têm-se que $\mathbf{S}^{-1} = \hat{\mathbf{B}}\hat{\mathbf{B}}^T$ e $\hat{\mathbf{B}}^T = \hat{\mathbf{A}}^{-1}$.

A log-verossimilhança para os modelos não perturbado e perturbado são $L(\boldsymbol{\theta})$ e $L(\boldsymbol{\theta}|\mathbf{w})$ respectivamente. O afastamento da verossimilhança $LD(\boldsymbol{\theta}_{\boldsymbol{w}})$ é definido como $2[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_w)]$, onde $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_w$ são os **MLEs** de $\boldsymbol{\theta}$ sob os modelos não perturbado e perturbado respectivamente. Seja $\mathbf{1}_m$ o vetor $m \times 1$ com todos os elementos iguais a 1. Defina a matriz de dimensão $[p(p+3)/2] \times n$

$$\Delta = \frac{\partial^2 L(\boldsymbol{\theta}|\mathbf{w})}{\partial \boldsymbol{\theta} \partial \mathbf{w}^T}$$

avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ e $\mathbf{w} = \mathbf{1}_n$, e a matriz $[p(p+3)/2] \times [p(p+3)/2]$

$$\ddot{\mathbf{L}} = \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \tag{4.9}$$

avaliada em $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Seja

$$\ddot{\mathbf{F}} = \Delta^T (\ddot{\mathbf{L}})^{-1} \Delta, \tag{4.10}$$

uma matriz de tamanho $n \times n$.

Para um vetor $\mathbf{q}$, denotamos por $\mathbf{q}_{(i+)}$ ao vetor que consiste nos primeiros i elementos do vetor $\mathbf{q}$. A matriz de dados de tamanho $p \times n$ é denotada como $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, com $\boldsymbol{x}_i$ com distribuição normal p-variada $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Seja $\mathbf{D}_i$ uma matriz diagonal tal que os elementos da diagonal são os elementos de $\hat{\mathbf{b}}_i^T(\boldsymbol{X} - \hat{\boldsymbol{\mu}}\mathbf{1}_n)$. A matriz $\ddot{\mathbf{F}}$ pode ser calculada da seguinte maneira

$$\ddot{\mathbf{F}} = -\frac{1}{n}\left[(\boldsymbol{X} - \hat{\boldsymbol{\mu}}\mathbf{1}_n^T)^T \hat{\mathbf{B}}\hat{\mathbf{B}}^T(\boldsymbol{X} - \hat{\boldsymbol{\mu}}\mathbf{1}_n^T) + \sum_{i=1}^{p} \mathbf{D}_i(\boldsymbol{X} - \hat{\boldsymbol{\mu}}\mathbf{1}_n^T)^T \mathbf{Q}_i (\boldsymbol{X} - \hat{\boldsymbol{\mu}}\mathbf{1}_n^T)\mathbf{D}_i\right] \tag{4.11}$$

onde $\mathbf{Q}_i$ é uma matriz $p \times p$ dada por

$$\mathbf{Q}_i = \begin{pmatrix} \hat{\mathbf{B}}_i \hat{\mathbf{B}}_i^T & 0 \\ 0 & 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} \hat{\mathbf{b}}_{i(i+)} \hat{\mathbf{b}}_{i(i+)}^T & 0 \\ 0 & 0 \end{pmatrix}$$

Aqui, $\hat{\mathbf{B}}_i \hat{\mathbf{B}}_i^T$ é a inversa da submatriz principal de $\mathbf{S}$ de ordem i. Seja $\mathbf{e}_r = \boldsymbol{x}_r - \hat{\boldsymbol{\mu}}$, a $(r,s)-$ésimo elemento do primeiro termo na matriz $\ddot{\mathbf{F}}$ pode-se escrever

$$\mathbf{e}_r^T \hat{\mathbf{B}}\hat{\mathbf{B}}^T \mathbf{e}_s = \sum_{i=1}^{p} (\hat{\mathbf{b}}_i^T \mathbf{e}_r)(\hat{\mathbf{b}}_i^T \mathbf{e}_s),$$

e o segundo termo em $\ddot{\mathbf{F}}$ pode ser expresso

$$\sum_{i=1}^{p}\left\{(\hat{\mathbf{b}}_i^T \mathbf{e}_r)(\hat{\mathbf{b}}_i^T \mathbf{e}_s)\left[\sum_{k=1}^{i}(\hat{\mathbf{b}}_k^T \mathbf{e}_r)(\hat{\mathbf{b}}_k^T \mathbf{e}_s)\right]\right\} - \frac{1}{2}\sum_{i=1}^{p}(\hat{\mathbf{b}}_i^T \mathbf{e}_r)^2(\hat{\mathbf{b}}_i^T \mathbf{e}_s)^2 = \frac{1}{2}\left[\sum_{i=1}^{p}(\hat{\mathbf{b}}_i^T \mathbf{e}_r)(\hat{\mathbf{b}}_i^T \mathbf{e}_s)\right]^2.$$

Assim, o $(r,s)-$ésimo elemento da matriz $\ddot{\mathbf{F}}$ é equivalente a

$$-\frac{1}{n}\left[\mathbf{e}_r^T \mathbf{S}^{-1} \mathbf{e}_s + \frac{1}{2}(\mathbf{e_r^T S^{-1} e}_s)^2\right]. \tag{4.12}$$

Denota-se $\boldsymbol{\rho_1}$ e $\boldsymbol{\rho_2}$ os autovetores associados ao menor e segundo menor autovalor da matriz $\ddot{\mathbf{F}}$ respectivamente, e seja $\mathbf{1}_{(i)}$ um vetor $i \times 1$ com o $i-$ésimo elemento igual a 1 e os outros iguais a 0. Então, o menor autovalor em valor absoluto é a máxima curvatura da curva que equivale a porção da superfície cortada pelo plano compreendido pelos vetores $\mathbf{1}_{(n+1)}$ e $(\boldsymbol{\rho}_1^T, 0)$ [Cook (1986, p. $138-139$)]. Aqui, os vetores $\boldsymbol{\rho_1}$ e $\boldsymbol{\rho_2}$ fornecem indicações sobre possíveis outliers. Assim, o gráfico de $\boldsymbol{\rho_1}$ versus $\boldsymbol{\rho_2}$ é útil para obter informações sobre outliers [Kim (1996)].

CAPÍTULO 5

Avaliação Numérica

5.1 Introdução

A análise de influência é uma metodologia freqüentemente utilizada em modelos de regressão. A detecção de observações influentes em um conjunto de dados é de grande importância na análise de diagnóstico, pois observações que apresentam peso importante na estimativa dos parâmetros podem levar a inferências erradas. Do mesmo modo acontece na análise de formas, observações influentes no conjunto de dados podem proporcionar, por exemplo, estimativas erradas da forma média e, desta forma, conclusões incorretas referentes ao conjunto de dados. Uma possível solução neste caso é usar métodos que ajudem na detecção de observações influentes ou/e outliers no conjunto de dados.

Para a distribuição Bingham complexa (veja Seção $2.3.2$), estimar corretamente os parâmetros é de grande importância na prática. Observações influentes no conjunto de dados de interesse influem nas estimativas dos parâmetros, pois na distribuição Bingham complexa os parâmetros estimados dependem diretamente da forma média, quando o conjunto de dados possui alta concentração.

5.2 Metodologia

Através de simulações de Monte Carlo foram avaliados os desempenhos de diferentes métodos para detecção de observações aberrantes em conjuntos de dados com distribuição Bingham complexa. No estudo de simulação, foram comparados o número de vezes que a distância de Cook, influência local para dados normais multivariados e o teste de discordância, métodos descritos no Capítulo 3, identificaram uma observação influente sob diferentes cenários.

Os dados foram gerados da distribuição Bingham complexa $\mathbb{C}B_{k-2}(A)$, com vetor de parâmetros $\lambda_1 < \lambda_2 < \ldots < \lambda_{k-1} = 0$ os quais representam os autovalores de A. O parâmetro k que representa o número de marcos anatômicos em cada forma permaneceu constante ao longo do experimento.

Os valores considerados para os vetores de parâmetros são $\boldsymbol{\lambda_{(1)}} = (-21, -20, 0)$, $\boldsymbol{\lambda_{(2)}} = (-51, -50, 0)$, $\boldsymbol{\lambda_{(3)}} = (-101, -100, 0)$ e $\boldsymbol{\lambda_{(4)}} = (-1.5, -1, 0)$, o valor do número de marcos anatômicos foi fixado em $k = 4$ e os tamanhos amostrais utilizados foram $n = 15, 20, 30$ e 50. As amostras da distribuição Bingham complexa foram obtidas mediante o algoritmo descrito na Seção 2.4.

O processo de simulação é divido em três partes, a seguir:
Na primeira parte do processo de simulação um breve exemplo é apresentado. Neste exemplo, 29 observações da distribuição Bingham complexa com vetor de parâmetros $\boldsymbol{\lambda_{(3)}} = (-101, -100, 0)$, uma observação da distribuição Bingham complexa com vetor de parâmetros $\boldsymbol{\lambda_{(4)}} = (-1.5, -1, 0)$, são geradas. Finalmente obtém-se uma amostra de tamanho 30, onde a última observação gerada é considerada como influente. Para esta última amostra os métodos para detecção de observações influentes são implementados.

Na segunda parte, uma amostra da distribuição Bingham complexa de tamanho n e vetor de parâmetros $\boldsymbol{\lambda_{(i)}}$, $i = 1, 2, 3$, é gerada e o desempenho da distância de Cook é avaliado, quando a amostra gerada provem da mesma população. Pelo fato de não ter um conhecimento adequado da distribuição para distância de Cook (3.6), no caso das amostras distribuídas na esfera, e em particular em amostras com distribuição Bingham complexa, utilizamos o método de bootstrap para encontrar a região de confiança. O número de amostras bootstrap foi fixado em 200. Os resultados da simulação também são interpretados em termos da região de confiança para a distância de Cook descrita em (3.7).

Na terceira parte, para cada réplica de Monte Carlo é gerada uma amostra de tamanho $n-1$ originária da distribuição Bingham complexa com vetor de parâmetros $\boldsymbol{\lambda_{(i)}}$, $i = 1, 2, 3$, e adicionalmente é gerada uma observação da distribuição Bingham complexa com vetor de parâmetros igual a $\boldsymbol{\lambda_{(4)}}$. Considera-se a última observação na amostra como uma observação influente, pois esta foi gerada de uma distribuição diferente às primeiras $n-1$ observações. Como base nesta amostra final de tamanho n, o desempenho dos diferentes métodos para identificação de dados influentes é avaliado.

5.3 Exemplo

Na Tabela 5.1 é apresentado o resultado de um experimento de simulação que tem como objetivo ilustrar a identificação de um ponto influente. O experimento é feito com a geração de 29 observações

da distribuição Bingham complexa com vetor de parâmetros $\boldsymbol{\lambda_{(3)}} = (-101, 100, 0)$, este vetor de parâmetros produz uma amostra com alta concentração. Uma única observação com distribuição Bingham complexa com vetor de parâmetros $\boldsymbol{\lambda_{(4)}} = (-1.5, 1, 0)$ é gerada, sendo esta a última observação da amostra. Assim, obtemos uma amostra final de tamanho 30. Os métodos distância de Cook definido em (4.7) e o teste para discordância definido em (4.8) são utilizados na amostra obtida por simulação. A primeira e segunda coluna da Tabela (5.1) indicam os resultados da distância de Cook ($DG(\boldsymbol{\theta_{(i)}})$) e o teste para discordância (J_m), respectivamente. Cada uma destas medidas é calculada para cada uma das observações na amostra. Considerando que a amostra se encontra definida no espaço de pré-formas, ela foi projetada no espaço tangente, e o método de influência local para dados normais multivariados foi implementado. É perceptível que a observação de ordem 30 apresenta os valores da distância de Cook e da estatística J_m muito superiores aqueles das demais observações. Finalmente, as duas últimas colunas representam os dois menores autovetores da matriz $\ddot{\mathbf{F}}$ mencionada na equação (4.12). Pode-se observar que na coluna $\boldsymbol{\rho_1}$ o maior valor em valor absoluto se encontra na observação 30. Porém, neste método, o critério para avaliar uma observação como influente depende dos dois valores dos autovetores. Portanto, é melhor utilizar o método gráfico para identificar com precisão as observações influentes.

Na Figura 5.1 mostra-se o gráfico de $\boldsymbol{\rho_1}$ *versus* $\boldsymbol{\rho_2}$ para os dados simulados. Esta figura proporciona uma informação mais clara que os resultados apresentados na Tabela 5.1. Nota-se que a observação 30 é influente ao longo da direção do vetor $\boldsymbol{\rho_1}$. Esta observação encontra-se significativamente afastada do resto do grupo de observações, e pode ser considerado como uma dado influente dentro da população observada.

Na Figura 5.2 apresentam-se os valores da distância de Cook para cada uma das observações. Claramente podemos perceber um alto valor da distância de Cook para a observação 30, enquanto que os valores da distância de Cook no restante das observação está perto de 0.

Os valores do teste para discordância são observados na Figura 5.3. Neste gráfico percebe-se que o valor do teste para discordância na observação 30 é alto com respeito ao restante das observações. Observamos que tanto a distância de Cook como o teste de discordância apresentam resultados similares e de fácil interpretação. Aqui a informação fornecida pela distância de Cook e o teste de discordância podem ser equivalentes.

5.4 Estudo de Simulação

Na Tabela 5.2 apresentamos os resultados de um estudo de simulação com a finalidade de comparar o desempenho através da porcentagem de vezes que a distância de Cook identifica observações influentes em distintas regiões de confiança:

Tabela 5.1: Resultados da distância de Cook, influência local e teste para discordância para uma amostra da distribuição Bingham complexa.

Obs	$DG(\boldsymbol{\theta}_{(i)})$	J_m	$\boldsymbol{\rho_1}$	$\boldsymbol{\rho_2}$
1	0.04	0.33	−0.00	0.08
2	0.02	0.66	0.01	0.20
3	0.05	0.18	−0.00	0.01
4	0.02	1.21	0.00	0.13
5	0.05	1.62	0.01	0.08
6	0.03	0.35	0.01	0.07
7	0.04	1.44	0.01	0.11
8	0.11	2.59	0.04	0.17
9	0.05	0.04	0.00	0.02
10	0.08	2.04	−0.00	0.13
11	0.03	0.28	0.02	0.02
12	0.04	1.42	−0.00	0.12
13	0.02	0.17	0.00	0.20
14	0.01	0.53	0.05	0.02
15	0.06	1.78	0.00	0.14
16	0.03	0.34	0.02	0.02
17	0.01	1.45	0.00	0.29
18	0.37	3.92	0.00	0.37
19	0.02	0.21	0.02	0.10
20	0.05	0.06	0.00	0.05
21	0.44	4.36	0.00	0.48
22	0.06	0.12	−0.00	0.00
23	0.04	0.73	0.00	0.03
24	0.03	0.67	0.00	0.04
25	0.02	1.42	0.02	0.40
26	0.02	0.56	0.00	0.14
27	0.03	0.87	0.01	0.06
28	0.04	0.09	0.01	0.03
29	0.02	1.57	0.00	0.37
30	24.04	15.08	1.00	−0.03

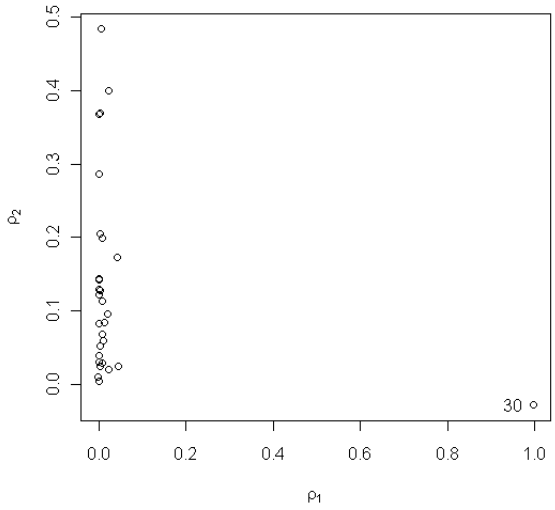

Figura 5.1: Gráfico de $\boldsymbol{\rho_1}$ *versus* $\boldsymbol{\rho_2}$ para a amostra da distribuição Bingham complexa.

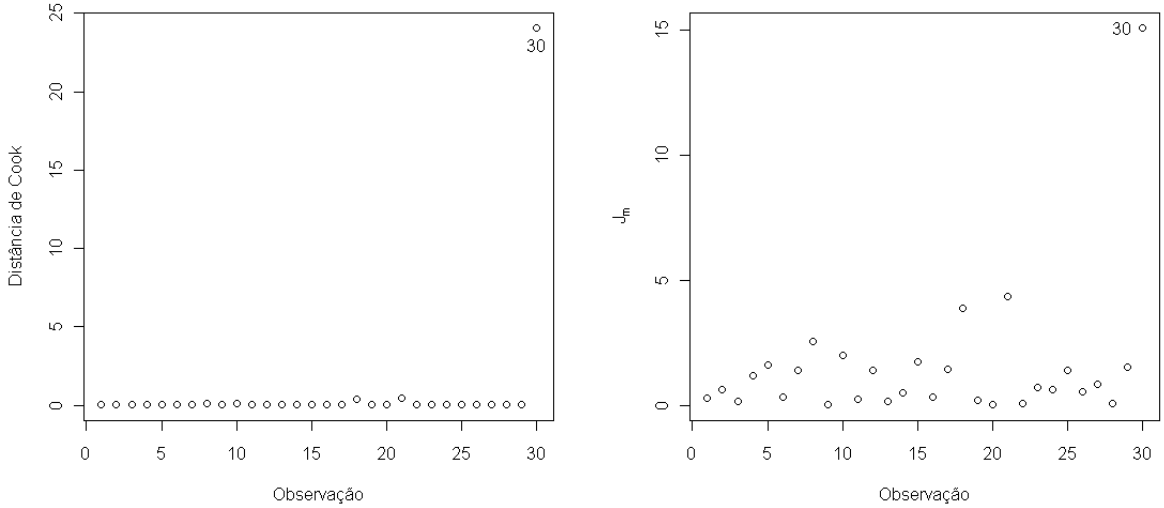

Figura 5.2: Gráfico da Distância de Cook para a amostra da distribuição Bingham complexa.

Figura 5.3: Gráfico do Teste para Discordância para a amostra da distribuição Bingham complexa.

- $\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T \mathbf{C} (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}) \leq \mathcal{X}^2_{p,\alpha}\}$ (veja Vanegas, L., H. (2005)).
- Região de confiança bootstrap (veja Seção 4.2).

O desempenho da distância de Cook nas duas regiões de confiança mencionadas anteriormente foi avaliado em 12 cenários de simulação distintos, baseados em observações da distribuição Bingham complexa, onde uma amostra de tamanho n é gerada com vetor de parâmetros $\boldsymbol{\lambda}_{(i)}$ $i = 1, 2, 3$. Foram utilizadas 1000 réplicas de Monte Carlo, amostras de treinamento de tamanhos 15, 20, 30 e 50 e o número de amostras bootstrap em cada réplica de Monte Carlo foi de tamanho 200. Em cada cenário calcula-se a porcentagem de vezes que a distância de Cook identifica observações influentes em cada réplica de Monte Carlo com um nível significância $\alpha = 0.05$. Podemos observar que para a região de confiança em $DG(\boldsymbol{\theta_{(i)}})$ o tamanho do teste está muito longe do nível nominal ($\alpha = 0.05$), ou seja, as porcentagem de identificação de observações influentes são significativamente menores que o nível nominal. Este resultado parece ser independente do tamanho de amostra n e do valor do vetor de parâmetros $\boldsymbol{\lambda}$. Entretanto, para a região de confiança $DG(\boldsymbol{\theta_{(i_B)}})$ podemos observar que para este caso o tamanho do teste está próximo do nível nominal. Notamos que na maioria dos casos quando o tamanho de amostra vai aumentando o tamanho do teste vai ficando mais próximo do nível nominal. Este comportamento também é observado quando a concentração das observações aumenta. A conclusão principal deste estudo de simulação é que a região de confiança $DG(\boldsymbol{\theta_{(i)}})$ não é a mais adequado para este tipo de estudo, este fato pode atribuir-se a que a distribuição das observações é na esfera complexa $\mathbb{C}S^{k-2}$. Propomos usar a região de confiança $DG(\boldsymbol{\theta_{(i_B)}})$ (região de confiança bootstrap), pois esta mostrou possuir um bom desempenho apesar de ter um alto "custo"computacional.

Considere-se agora uma amostra de tamanho n, onde $n-1$ das observações são geradas a partir da distribuição Bingham com vetor de parâmetros $\boldsymbol{\lambda_{(i)}}$, e a última observação é gerada a partir da distribuição Bingham complexa com vetor de parâmetros $\boldsymbol{\lambda_{(4)}}$. Assim, a última observação do conjunto de dados foi gerada para ser influente.

O objetivo desta simulação é avaliar e comparar o desempenho da distância de Cook frente aos métodos teste para discordância e influência local em dados normais multivariados, na detecção da observação aberrante no conjunto de dados. Cada um dos métodos foi avaliado em 12 cenários.

A distância de Cook será avaliada nas regiões de confiança descritas anteriormente. A região de confiança para o teste para discordância J_m também foi originada mediante o método bootstrap. No método de influência local em dados normais multivariados, considera-se $\rho_1^{(i)}$ e $\rho_2^{(i)}$ como a $i-$ésima componente do vetor $\boldsymbol{\rho_1}$ e a $i-$ésima componente do vetor $\boldsymbol{\rho_2}$, respectivamente, e assim o vetor $\boldsymbol{\rho}^{(i)} = (\rho_1^{(i)}, \rho_2^{(i)})$ é definido. Aqui a norma do vetor $\boldsymbol{\rho}^{(i)}$ é calculada, e se considera a $i-$ésima observação como aberrante quando a norma do vetor $\boldsymbol{\rho}^{(i)}$ é a maior no conjunto de normas dos vetores.

Tabela 5.2: Comparação do tamanho do teste nas regiões de confiança da Distância de Cook.

n	$\boldsymbol{\lambda}$			$DG(\boldsymbol{\theta_{(i)}})$	$DG(\boldsymbol{\theta_{(i_B)}})$
15	-21	-20	0	0.001	0.084
	-51	-50	0	0.001	0.072
	-101	-100	0	0.001	0.062
20	-21	-20	0	0.002	0.080
	-51	-50	0	0.002	0.075
	-101	-100	0	0.002	0.058
30	-21	-20	0	0.002	0.067
	-51	-50	0	0.002	0.060
	-101	-100	0	0.003	0.053
50	-21	-20	0	0.001	0.063
	-51	-50	0	0.001	0.058
	-101	-100	0	0.001	0.051

Na Tabela 5.3 se ilustra o número de vezes que cada um dos métodos identificou a última observação no conjunto de dados como influente. Para a distância de Cook na região de confiança $\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T \mathbf{C} (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}) \leq \mathcal{X}^2_{p,\alpha}\}$, podemos observar uma estreita relação entre o tamanho de amostra n e o vetor de parâmetros $\boldsymbol{\lambda}$. Notamos que quando $\boldsymbol{\lambda} = (-101, -100, 0)$, o poder do teste cresce à medida que o tamanho de amostra aumenta.

Para a distância de Cook (bootstrap), o teste J_m e o método de influência local para dados normais multivariados, podemos observar resultados similares. Na medida que o tamanho de amostra aumenta o poder do teste cresce. Da mesma forma acontece quando as amostras são geradas com alta concentração. Ou seja, quando a concentração dos dados aumenta o poder do teste cresce. A distância de Cook (bootstrap) e o método de influência local apresentarão resultados satisfatórios com respeito ao poder do teste, no entanto o método da distância de Cook (botstrap) é mais "custoso" computacionalmente.

Tabela 5.3: Poder do teste para cada um dos métodos de indentificação de observações influentes.

n	$\boldsymbol{\lambda}$			$DG(\boldsymbol{\theta}_{(i)})$	$DG(\boldsymbol{\theta}_{(i_B)})$	J_m	$\boldsymbol{\rho_1}$ $\boldsymbol{\rho_2}$
15	-21	-20	0	0.371	0.889	0.518	0.753
	-51	-50	0	0.871	0.993	0.597	0.949
	-101	-100	0	0.925	0.997	0.662	0.988
20	-21	-20	0	0.263	0.895	0.484	0.815
	-51	-50	0	0.816	0.964	0.577	0.952
	-101	-100	0	0.950	0.992	0.614	0.987
30	-21	-20	0	0.097	0.920	0.586	0.800
	-51	-50	0	0.762	0.989	0.678	0.957
	-101	-100	0	0.935	0.996	0.714	0.991
50	-21	-20	0	0.011	0.883	0.545	0.829
	-51	-50	0	0.641	0.987	0.670	0.966
	-101	-100	0	0.939	0.995	0.714	0.993

CAPÍTULO 6

Conclusões

No presente trabalho se implementou e avaliou a Distancia de Cook na detecção de observações influentes na análise estatística de formas. Através do modelo de deleção de casos (CDM) obtivemos a medida da distância Cook quando o conjunto de dados tem distribuição Bingham complexa. Considerando os resultados obtidos através das simulações de Monte Carlo realizadas, pode-se concluir que a medida distância de Cook não tem um bom desempenho quando a região de confiança é definida como $\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^T \mathbf{C} (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}) \leq \mathcal{X}^2_{p,\alpha}\}$, entretanto, quando a região de confiança calculada pelo método bootstrap o desempenho obtido é bem melhor.

Considerando-se que na análise estatística de formas o tópico de influência é pouco estudado, varias técnicas utilizadas em outros âmbitos da estatística adptadas nesta dissertação. Neste trabalho adaptamos três métodos de análise de influência para a análise estatística de formas. Levando-se em conta o poder do teste, verificamos que a distância de Cook (bootstrap) apresentou resultados satisfatórios, onde ao aumentar o tamanho de amostra e a concentração das observações o poder do teste melhorava. Observamos que o teste para discordância apresentou um comportamento similar à distância de Cook, no entanto o poder do teste neste método apresenta valores baixos.

Uma boa alternativa para análise de influência em análise de formas é trabalhar no plano tangente. Aqui o método de influência local para dados normais multivariados apresenta um ótimo desempenho quando as pré-formas são altamente concentradas, o que era esperado de acordo com a literatura sobre análise estatística de formas. Este método apresentou resultados igualmente eficientes à distância de Cook, mas quando a concentração dos dados for baixa, a distância de Cook irá apresentar resultados superiores.

A distância de Cook e o teste para discordância são de fácil interpretação, enquanto que para o método de influência local é necessário fazer uma análise gráfica dos resultados para não obter conclusões incorretas.

CAPÍTULO 7

Trabalhos Futuros

Considerando que os resultados do teste de discordância J_m não foram bons, pois este é apropriada para a dados com distribuição de Watron Real, é necessário formular uma estatística para identificar pontos aberrantes da distribuição de Bingham complexa.

Considerando que a forma média é um autovetor, as medidas de influência que são utilizadas em componentes principais podem ser relevantes para análise estatística de formas. Como em Brooks (1994) que apresenta funções de influência para componentes principais, estas funções podem produzir resultados importantes na análise estatística de formas. Comparar estas funções com as medidas apresentadas nesta dissertação pode ser um tópico importante para pesquisas futuras.

Método de bootsrtap é pouco abordado na literatura sobre a identificação de pontos influentes e pontos aberrantes. Dessa forma, é necessário comparar a proposta desta dissertação com outros métodos disponíveis na literatura como por exemplo o método apresentado em Azzalini, A., Bowman. A., W. & Händle, W. (1989).

APÊNDICE A

Programas

Neste Apêndice apresentamos o código na linguagem R utilizado nas simulações de Monte Carlo.

O seguinte programa gera amostras com distribuição Bingham complexa.

```
 bingham<-function(bnsam,bk,blambda)
{
    expo<-array(dim=c(bk-2))
    bing<-array(dim=c(bk-1))
    out<-array(dim=c(bk-1))
    bingsam<-array(dim=c((bk-1),bnsam))
    expos<-array(dim=c((bk-2),bnsam))
    uni<-array(dim=c(bk-1))
    for(is in 1:(bnsam))
    {
        flag<-0
        while(flag==0)
        {
            uni<-runif(bk-2,0,1)
                for(i in 1:(bk-2))
                {
                    expo[i]<--(1/blambda[i])*log(1-uni[i]*(1-exp(-blambda[i])))
                }
                if(sum(expo)<1)
                {
                    flag<-1
```

```
                }
        }
        flag<-0
        for( i in 1:(bk-2))
        {
            bing[i]<-expo[i]
        }
        bing[bk-1]<-1-sum(expo)
        theta<-runif(bk-1,0,1)*2*pi
        for(i in 1:(bk-1))
        {
        out[i]<-(bing[i]^(1/2))*exp(1i*theta[i])
        }
            bingsam[,is]<-out
        }
    bingsam
}
```

Programa para calcular a matriz de informação quando os dados tem distribuição Bingham complexa.

```
a<-function(x)
# x  vetor de lambdas
{
   v=matrix(0,k-1,1)
   for(j in 1:(k-1)) v[j]=(prod((x[j]-x)[-j]))^(-1)
   return (v)
}
b<-function(x)
{
   v=matrix(0,k-1,1)
   for(j in 1:(k-1)) v[j]= sum(((x[j]-x)[-j])^(-1))
  return (v)
}
d<-function(x)
{
   v=matrix(0,k-1,1)
   for(j in 1:(k-1)) v[j]= sum(((x[j]-x)[-j])^(-2))
  return (v)
}
       ##########################################
```

```
       #    Primeiras e segundas derivadas      #
       ########################################
CA<-function(x)
{
ca = (2*pi^(k-1))*sum(a(x)*exp(x))
      return (ca)
}
CmA<-function(x)
{
      v = matrix(0,k-1,1)
      for(m in 1:(k-1)) v[m]= (2*(pi)^(k-1))*(sum(((a(x)*exp(x))[-m])*
      (((x-x[m])[c(-m)])^(-1)))+((a(x)*exp(x))[m]) -((a(x)*b(x)*exp(x))[m]))
return (v)
}
CmsA<-function(x) {
   v<-array(0, c((k-1), (k-1)))
   for(i in 1:(k-1))
   {
      for(j in 1:(k-1))
      {
       if(i !=j)
       {
          v[i,j]= (2*(pi)^(k-1))*
 (sum(((a(x)*exp(x))[c(-i,-j)])*(((x-x[i])[c(-i,-j)])^(-1))*(((x-x[j])[c(-i,-j)])^(-1)))-
                                  (((a(x)*b(x)*exp(x))[c(i)])*((x[i]-x[j])^(-1)))-
                                  (((a(x)*b(x)*exp(x))[c(j)])*((x[j]-x[i])^(-1)))-
                                  (((a(x)*exp(x))[c(i)])*((x[i]-x[j])^(-2)))-
                                  (((a(x)*exp(x))[c(j)])*((x[j]-x[i])^(-2)))+
                                  (((a(x)*exp(x))[c(i)])*((x[i]-x[j])^(-1)))+
                                  (((a(x)*exp(x))[c(j)])*((x[j]-x[i])^(-1))))
        }
        else
        {
       v[i,j]= (2*(pi)^(k-1))*(2*sum(((a(x)*exp(x))[c(-i)])*(((x-x[i])[c(-i)])^(-2)))-
                              2*((a(x)*b(x)*exp(x))[c(i)])+
                              ((a(x)*exp(x))[c(i)])+
                              ((a(x)*((b(x))^(2))*exp(x))[c(i)])+
                              ((a(x)*d(x)*exp(x))[c(i)]))
        }
    }
```

```
   }
return (v) }
        ######################################
        #      Matriz de informaçao          #
        ######################################
He<- function (x) {
    v<-array(0, c((k-1),(k-1)))
    for(i in 1:(k-1))
    {
        for(j in 1:(k-1))
        {
            if(i!=j)
            {
              v[i,j]= (n /((CA(x))^2))*((CmA(x)[i]*CmA(x)[j])-((CmsA(x)[i,j])*CA(x)))
            }
            else
            {
              v[i,j]= (n/((CA(x))^2))*(((CmA(x)[i])^2)-(((CmsA(x))[i,j])* CA(x)))
            }
        }
    }
return (v)}

################################################
#                                              #
# Calculo pre-formas e autovalores da matriz S #
#                                              #
################################################
zal<-function(z0,nzal)
{
    star<-function(v1)
    {
        return(t(Conj(v1)))
    }
    norma<-function(u)
    {
        nor<-sqrt(sum(diag(Conj(t(u))*u )))
        return (nor)
    }
    #pre-shape
    z<-array(0,c(k-1,nzal))
```

```
    for(i in 1:nzal)
    {
        z[,i]<-z0[,i]
    }
    zz<-array(0,c(k-1,k-1,nzal))
    for(i in 1:nzal)
    {
        zz[,,i]<-z[,i]*star(z[,i])
    }
#Matriz complexa de somas de quadrados e produtos.
    s<-array(0,c(k-1,k-1))
    for(i in 1:nzal)
    {
        s<-s+zz[,,i]
    }
#parte da versimilanza
 v = eigen(s)
 l = v$values
 m = v$vectors
 return(l)
}
```

Processo de Monte Carlo.

```
#-----------Parametros da simulacao-----------------#

NREPM = 1000
NREPBoo = 200
n = 30
k=4
H<-defh(k-1)

lambda<-array(k-1)
lambda[1]=-51
lambda[2]=-50
lambda[3]=0
lambda
lambda.1<-array(0,c(k-2))
lambda.1[1]<- lambda[3]-lambda[1]
lambda.1[2]<- lambda[3]-lambda[2]
lambda.1
```

```
lambda.in<-array(k-1)
lambda.in[1]=-1
lambda.in[2]=-0.5
lambda.in[3]=0
lambda.in
lambda.in1<-array(0,c(k-2))
lambda.in1[1]<- lambda.in[3]-lambda.in[1]
lambda.in1[2]<- lambda.in[3]-lambda.in[2]
lambda.in1


#-------------Log Verossimilhanca-----------------------#

loglik = function(lambda)
{
return (sum(l*lambda)-n*log(2*(pi)^(k-1))-n*log(sum(a(lambda)*exp(lambda))))
}



loglik.i = function(lambda)
{
return (sum(l*lambda)-(n-1)*log(2*(pi)^(k-1))-(n-1)*log(sum(a(lambda)*exp(lambda))))
}

#-------------Funciones-----------------------------------#
norma<-function(u)
{
nor<-sqrt(sum(diag(Conj(t(u))%*%u)))
return (nor)
}

star<-function(v1)
{
return(t(Conj(v1)))
}

#------------------Laco de Monte Carlo--------------------#
Quantil=array(0,c(NREPM,1))
cont=0
```

```
cont.1=0
cont.2=0
for(j in 1:NREPM){
#-------------Gerador de amostras-------------------------#

am = bingham(n-1,k,lambda.1)
influ = bingham(1,k,lambda.in1)
amostra = matrix(c(am,influ), nrow=(k-1), ncol=n)

#----------------Identifica outliers--- --------------------#

rho=zal(amostra,n)[k-1]
rho.i=array(0,c(n,1))
for(i in 1:n){
rho.i[i,]=zal(amostra[,-i],n-1)[k-1]
}

Jn=array(0,c(n,1))
for(i in 1:n){
Jn[i,]=(n-3)*((rho-rho.i[i,])/rho.i[i,])
}


#-------------------BFGS----------------------------------#

#Chute inicial
chi=lambda

#Calculo parametro l com base na amostra Original
l= zal(amostra,n)

#Estimacao de lambda
bfgs= optim(chi, loglik, hessian = FALSE,control=list(fnscale=-1))
thetas=bfgs$par

#------------BFGS sin obseravcao i-------------------------#
thetas.i=array(0, c(k-1,n))
for(i in 1:n)
{
#Calculo parametro i
```

```
l= zal(amostra[,-i],(n-1))

#Estimacao de lambdas
bfgs.i=optim(chi, loglik.i, hessian = FALSE,control=list(fnscale=-1))
thetas.i[,i]=bfgs.i$par
}
#---------------Distancia de CooK-------------------------#

Dg<- array(0,c(n,1))
for(i in 1:n)
{
    Dg[i,]=t(thetas.i[,i]-thetas)%*%(-He(thetas))%*%(thetas.i[,i]-thetas)

}


#-------------Proceso Boostrap----------------------------#
#------------Amostras Boostrap----------------------------#
indice<-1:n
amboo<-array(0,c(k-1,n,NREPBoo))

for(i in 1:NREPBoo)
{
ordem<-sample(indice,n,replace="TRUE")
amboo[,,i]=amostra[,ordem]
}

#--------------------BFGS---------------------------------#

thetas.boo=array(0,c(k-1,NREPBoo))
for(i in 1:NREPBoo){
#Chute inicial
chi=lambda

#Caluclo parametro l com base na amostra Bootstrap
l=zal(amboo[,,i],n)

#BFGS
bfgs.boo = optim(chi, loglik, hessian = FALSE,control=list(fnscale=-1))
```

```
#Estimacao dos lambdas
thetas.boo[,i]=bfgs.boo$par
conver = bfgs$convergence
}

#------------------BFGS sin obseravcion i------------------#

thetas.booi=array(0,c(k-1,n,NREPBoo))
for(w in 1:NREPBoo){
for(i in 1:n){
# Chute inicial
      chi = lambda

# Calculo do parametro l sim o idividuo w da amostra
l = zal(amboo[,-i,w],n-1)

      # BFGS
bfgs.booi = optim(chi, loglik.i, hessian = FALSE,control=list(fnscale=-1))

      # Parametros
thetas.booi[,i,w] = bfgs.booi$par
}}

#-----------------Distancia de Cook Boostrap----------------#

Dg.boo<- array(0,c(n,NREPBoo))
for(w in 1:NREPBoo){
for(i in 1:n){

    Dg.boo[i,w]=t(thetas.booi[,i,w]-thetas.boo[,w])%*%
    (-He(thetas.boo[,w]))%*%(thetas.booi[,i,w]-thetas.boo[,w])

}}

#-------------------------------------------------------------#

Dg.ordem=array(0,c(n,NREPBoo))
for(i in 1:NREPBoo){
Dg.ordem[,i]=sort(Dg.boo[,i])
}
```

```
Dboot=array(0,c(NREPBoo,1))
for(i in 1:NREPBoo){
Dboot[i,]= Dg.ordem[round(n*0.95),i]
}

Quantil[j,]= mean(Dboot)

#-----------------------------------------------------------------#

if(Dg[n,]>=7.814728){cont=cont+1}
if(Dg[n,]>=Quantil[j,]){cont.1=cont.1+1}

print(j)
}
```

Referências

[1] Amaral, G., Dryden, I. & Wood, A. (2007), "Pivotal boostrap methods for k-sample problems in directional statistics and shape analysis". *Journal of the Americam Statistical Association*, **102**, $695 - 707$.

[2] Azzalini, A., Bowman. A., W. & Händle, W. (1989), "On the use of nonparametric regression for model checking". *Biometrika* **76**, $1 - 11$.

[3] Brooks, S., P. (1994), "Diagnostic for principal components: Influence function as diagnostic tools". *The Statistician* **43**, $483 - 494$.

[4] Bookstein, F. L. (1984), "A statistical method for biological shape comparisions". *Journal of Theoretical Biology* **107**, $475 - 520$.

[5] Bookstein, F. L. (1986), "Size and shape spaces for landmark data in two dimensions (with discussion)". *Statistical Science* **1**, $181 - 242$.

[6] Cook, R. D. & Weisberg, S. (1982), *Residuals and influence in Regression*. New York: Chapman and Hall.

[7] Cook, D. (1986), "Assessment of Local Influence". *Journal of the Royal Statistical Society*, Serie **B**, **48**, $133 - 169$.

[8] Davison, A., C. & Hinkley, D., V. (1997), *Bootstrap Methods and Their Application*. New York: Cambridge University Press.

[9] De Castro, R. (2003), *El universo del LaTeX*. Bogotá: Universidad Nacional de Colombia.

[10] Dryden, I. & Mardia, K. (1998), *Statistical Shape Analysis*. Chichester: Wiley and Sons.

[11] Efron, B. & Tibshirani, R. (1993), *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[12] Fisher, N. I., Lewis, T. & Embleton, B. J. (2004), *Statistical Analysis of Spherical Data*. Cambridge: Press Syndicate of the University of Cambridge.

[13] Fisher, R. A. (1934), "Two new properties of mathematical likelihood".

[14] Jaques, F. V. (2008), Testes de permutação e bootstrap em análise estatística de formas: Aplicações à zoologia. Dissertação de mestrado, Universidade Federal de Pernambuco.

[15] Kendall, D. G. (1984), "Shape manifolds, procrustean metric and comlex projective spaces". *Bulletin of the London Mathematical Society* **16**, $81 - 121$.

[16] Kent, J. T. (1994), "The complex Bingham distribution and shape analysis". *Jornal of the Royal Statistical Socoety, Series* **56**, $285 - 299$.

[17] Kent, J. T., Constable, P. D. & Er, F. (2004), "Simulation for the complex Bingham distribution". *Statistics and Computing* **14**, $53 - 57$.

[18] Kim, M. G. (1996), "Local influence in multivariate normal data". *Journal of Applied Statistics*, **23**, $535 - 541$.

[19] Mardia, K. V. & Jupp, P. E. (2000), *Directional Statistics*. New York: John Wiley & Sons, LTD.

[20] Mirsky, L. (1955), *An introduction to linear Algebra*. London: Oxford University Press.

[21] Paula, G. A. (2004), *Modelos de regressão com com apoio computacional*. Sao Paulo: Instituto de Matemática e Estatística (USP).

[22] Shelupsky, D. (1962), "An introduction to spherical coordinates". *American Mathematical Monthly* **69**, $644 - 646$.

[23] Small, C. G. (1996), *The Statistical Theory of Shape*. New York: Springer-Verlag.

[24] Vanegas, L. H. (2005), Diagnóstico em modelos simétricos de regressão. Dissertação de mestrado, Universidade Federal de Pernambuco.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo