AUGUSTO MACIEL DA SILVA

MEDIDAS ANGULARES EM COMPONENTES PRINCIPAIS REPARAMETRIZADOS EM AMOSTRAS COM VALORES DISCREPANTES

LAVRAS – MG

2013

AUGUSTO MACIEL DA SILVA

MEDIDAS ANGULARES EM COMPONENTES PRINCIPAIS REPARAMETRIZADOS EM AMOSTRAS COM VALORES DISCREPANTES

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador

Dr. Augusto Ramalho de Morais

Coorientador

Dr. Marcelo Angelo Cirillo

LAVRAS – MG

2013

Ficha Catalográfica Elaborada pela Coordenadoria de Produtos e Serviços da Biblioteca Universitária da UFLA

Silva, Augusto Maciel da.
Medidas angulares em componentes principais reparametrizados em amostras com valores discrepantes / Augusto Maciel da Silva. – Lavras : UFLA, 2013.
110 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2013.
Orientador: Augusto Ramalho de Morais.
Bibliografia.

1. Estatística circular. 2. Componentes interpretáveis. 3. Normal contaminada. 4. Estruturas de correlação. I. Universidade Federal de Lavras. II. Título.

CDD – 519.535

AUGUSTO MACIEL DA SILVA

MEDIDAS ANGULARES EM COMPONENTES PRINCIPAIS REPARAMETRIZADOS EM AMOSTRAS COM VALORES DISCREPANTES

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 31 de julho de 2013.

Dra. Carla Regina Guimarães Brighenti	UFSJ
Dr. Ronaldo Rocha Bastos	UFJF
Dr. Fortunato Silva de Menezes	UFLA
Dr. Marcelo Ângelo Cirillo	UFLA
Dra. Thelma Sáfadi	UFLA

Dr. Augusto Ramalho de Morais
Orientador

LAVRAS – MG

2013

A minha amada esposa Liliam,

Pelo amor, amizade, apoio incondicional e paciência

DEDICO

Aos meu pais, Maria Isabel da Silva,

José Salvador da Silva, meu irmão André Maciel da Silva e tio Ismael,

OFEREÇO

AGRADECIMENTOS

A Deus pelo dom da vida e por conceder-me luz e sabedoria e a Nossa Senhora Aparecida pela proteção e companhia nesta caminhada.

A minha esposa Liliam, pelo amor, força e parceria ao longo destes anos.

Aos meus pais Bebel e Nenzinho pelas orações, carinho, conselhos e, principalmente, pela confiança em mim depositada.

Ao meu irmão André pela amizade e por compartilhar com todos a perseverança e desejo de vencer.

Ao meu cachorro Shelby pelos inúmeros momentos de alegria e sempre me receber com entusiasmo ao chegar estressado em casa.

Aos meus sogros Mari e Zezé, por todo apoio e tempo a mim dispensados.

Ao professor Augusto Ramalho de Morais, pela orientação e ensinamentos.

Ao professor Marcelo Ângelo Cirillo, pela coorientação e, principalmente, pela amizade cultivada ao longo destes anos.

Aos membros da banca pela disponibilidade e contribuições finais para este trabalho.

Aos professores e funcionários do Departamento de Ciências Exatas da Universidade Federal de Lavras.

Aos colegas do DEX, Ana Paula, Paulo, Crysttian, Moysés, Tânia, Edcarlos, Adriana, Diogo e Felipe pela convivência e amizade nestes anos.

Ao grande amigo Leandro Ferreira pelo grande companheirismo, paciência e conselhos dentro e fora do doutorado

Ao amigo e irmão, Carlos Eduardo, pela amizade verdadeira e, também, pelo suporte gráfico ao longo destes anos.

À amiga Ana Lúcia por tornar mais fácil esta caminhada.

Aos professores e funcionários do Departamento de Estatística da Universidade Federal de Santa Maria.

Aos amigos Fábio e Débora, pela amizade e excelente receptividade em terras gaúchas.

Aos novos amigos Maro e Eliane e ao futuro afilhado, Daniel.

Dave, Chris, Nate, Pat e Taylor, por embalar as noites de trabalho.

A CAPES pela bolsa de estudos, essencial para a realização deste trabalho.

A todos que de alguma forma contribuíram para a realização deste trabalho, meus sinceros agradecimentos.

RESUMO

Alguns tipos de dados, como as medidas angulares, requerem certas restrições na utilização de métodos estatísticos, sendo tratados pela estatística circular. Medidas que retratam ângulos são exemplos de dados circulares. Em se tratando de técnicas estatísticas multivariadas, medidas angulares estão relacionadas com os Componentes Principais e Interpretáveis. A análise de Componentes Principais é uma técnica de redução de dimensionalidade que identifica combinações lineares que expliquem a maior parte da variação dos dados. Os Componentes Interpretáveis utilizam restrições para que se tenha uma melhor interpretação dos coeficientes dessas combinações limitando os valores assumidos pelos coeficientes. Sua eficiência em relação ao Componente Principal é avaliada em relação ao ângulo formado entre os componentes, que deve ser mínimo. Assim, objetivou-se neste trabalho avaliar, por meio de simulação computacional, o efeito da presença de observações discrepantes na reparametrização dos componentes principais pelos componentes interpretáveis, utilizando de diferentes probabilidades de mistura, estruturas de correlação e coeficientes de correlação utilizada na geração das amostras. Foi proposta uma medida para identificação das distâncias circulares entre os valores médios angulares sob contaminação e sem contaminação. Os resultados obtidos por meio de simulação mostraram que as médias angulares dos componentes se diferem quanto ao coeficiente de correlação e estrutura de correlação utilizada e a medida de distância circular proposta identificou o efeito das observações discrepantes, por meio de pontos dissimilares.

Palavras-chave: Medidas Angulares. Estatística Circular. Componente Principal. Componente Interpretável. Observações Discrepantes.

ABSTRACT

Some types of data, such as angular measurements, require certain restrictions on the use of statistical methods, being treated by circular statistics. Angle measurements are examples of circular data. When considering multivariate statistical techniques, angular measurements are related to the Principal Components and Interpretable Components. Principal Component Analysis is a dimensionality reduction technique which identifies linear combinations that explain most data variations. Interpretable Components use restrictions in order to have a better interpretation of the coefficients of these combinations, restricting the values assumed by the coefficients. Their efficiency compared to the Principal Component is evaluated in relation to the angle formed between the components, which should be minimal. Thus, the objective of this study was to evaluate by computer simulation the effect of outliers in the reparameterization of the principal components by the interpretable components using different mixture probabilities, correlation structures and correlation coefficients, used to generate the samples. We proposed a measure to identify the circular distances between the expected angular values under contamination and without contamination. The results obtained through simulation showed that the angular means of the components differ in regard to the correlation coefficient and the correlation structure used, and that the circular distance measurement proposed identified the effect of outliers through dissimilar points.

Keywords: Angular measurements. Circular statistics. Principal Components. Interpretable Components. Outliers.

LISTA DE FIGURAS

ANEXO B

LISTA DE TABELAS

ANEXO A

SUMÁRIO

1 INTRODUÇÃO

A análise estatística de dados, por vezes, requer conhecimento sobre determinadas características dos mesmos. Determinadas técnicas, mesmo as mais usuais, podem apresentar algum tipo de restrição dependendo do conjunto de dados com o qual se está trabalhando.

Dessa forma, dados estatísticos podem ser classificados de acordo com a sua topologia distribucional. Dados lineares podem ser representados considerando uma reta. Por outro lado, a circunferência é apropriada para representar um conjunto de medidas angulares, que podem se referir a observações mensuradas, por exemplo, como ângulos, distribuídos geralmente em graus ou radianos (ABUZAID et al., 2012).

Medidas angulares ocorrem em vários campos do conhecimento, como biologia, meteorologia, medicina, análise de imagens, astronomia (MARDIA, 1972). Uma observação circular pode ser definida como um ponto em um círculo de raio unitário ou um vetor unitário indicando direção. Desde que uma direção inicial e uma orientação do círculo sejam definidas, cada observação circular pode ser especificada pelo ângulo formado entre a direção inicial do círculo e o ponto no círculo correspondente à observação.

A periodicidade relacionada a esse tipo de medida acarreta situações que não ocorrem em observações na reta. Sendo assim, existem técnicas estatísticas específicas para tratar esses tipos de dados, sendo necessárias definições de medidas de posição e dispersão, bem como modelos probabilísticos apropriados que são tratados pela estatística circular.

Medidas angulares estão sujeitas aos mesmos fenômenos que os dados lineares, como, por exemplo, ocorrência de *outliers*. Estudos sobre ocorrência de *outliers* em dados circulares são encontrados em Ibrahim et al. (2013) e Abuzaid et al. (2012).

Em se tratando de técnicas estatísticas, cita-se, como exemplo, algumas técnicas de análise multivariada em que podem ser obtidas de alguma forma, medidas angulares, como os Componentes Principais e Componentes Interpretáveis, que têm como medida resultante um ângulo formado entre as suas direções.

Um dos objetivos da estatística multivariada é a redução de dimensionalidade de um conjunto de dados, com perda mínima de informação, para que se possam executar análises de forma menos complexa e a técnica de análise de Componentes Principais pode ser utilizada para esse fim.

Segundo Johnson e Wichern (2007), a análise de Componentes Principais tem por característica explicar a estrutura de variância e covariância de um conjunto de variáveis, por meio de poucas combinações lineares destas variáveis, que promovem uma rotação no eixo do sistema. Assim, podem-se citar dois objetivos que são a redução do sistema de dados e a interpretação.

Apesar da facilidade de aplicação da técnica de Componentes Principais, dependendo do número de variáveis, da presença de *outliers* e do número de componentes retidos, estes podem apresentar situações que dificultam algum tipo de interpretação. Assim, Chipman e Gu (2005) introduziram algumas restrições aos componentes, para que sejam mais interpretáveis, no sentido de limitar os valores assumidos pelos seus coeficientes, restringindo-os. Desta forma surgiu uma nova reparametrização, denominada Componentes Interpretáveis (CI).

Essencialmente os CI são validados, por meio da obtenção do ângulo entre o Componente Interpretável e o Componente Principal, que deve apresentar mínima variação angular. Portanto, ao se fazer inferência na distribuição desses ângulos, torna-se necessária a utilização de uma inferência estatística, apropriada para distribuições angulares, justificando a utilização das técnicas para esse tipo de dados.

A interpretabilidade dos Componentes Principais tem sido objeto de estudo sob diferentes enfoques. Enki, Trendafilov e Jolliffe (2013) consideraram um novo método para se obter Componentes Principais Interpretáveis. Primeiramente realizaram uma análise de cluster (agrupamento) das variáveis, utilizando as técnicas multivariadas já existentes e, após a identificação dos grupos similares, foram obtidos os Componentes Interpretáveis a partir das matrizes de correlação das variáveis já agrupadas.

A metodologia para avaliar os Componentes Principais na presença de *outliers* e posterior recomendação do uso dos Componentes Interpretáveis, dar-se-á na utilização de recursos computacionais por meio de simulação Monte Carlo. Neste contexto amostras de variáveis com distribuição normal multivariada são geradas, sendo algumas unidades provenientes de outra população, caracterizando uma mistura de distribuições. Tal procedimento caracteriza a distribuição normal contaminada.

Como a contaminação é feita na amostra, torna-se necessária a obtenção de alguns critérios para análise desse efeito nos valores médios angulares entre os componentes. Dessa forma, justifica-se a utilização dos conceitos de média e distância circulares nos ângulos obtidos pelos componentes, contribuindo para a disseminação de técnicas estatísticas angulares.

Partindo dessa premissa, o presente trabalho foi realizado com os objetivos de apresentar as seguintes contribuições:

a) Avaliar o efeito de estruturas de correlação, *outliers* e tamanho amostral na construção de Componentes Principais e na sua reparametrização, dita por Componentes Interpretáveis.
b) Utilizar medidas angulares para discriminar este efeito em função de amostras com diferentes graus de simetria e curtose.

2 REFERENCIAL TEÓRICO

2.1 Estatística para dados circulares

Ao se tratar de medidas angulares, encontram-se certas peculiaridades que não ocorrem em situações na reta. Este fato faz com que se tornem necessárias algumas definições específicas paras as medidas de posição e dispersão, bem como de modelos probabilísticos específicos. A literatura sobre o assunto pode ser encontrada em Mardia (1972), Batschelet (1981) e Fisher (1993). A Figura 1 representa, graficamente, uma amostra referente a dados angulares.

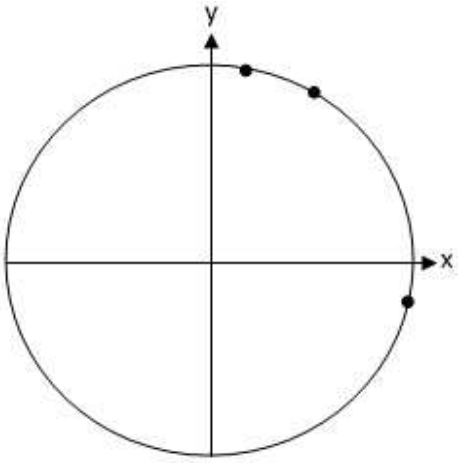


Figura 1 Representação gráfica de uma amostra com medidas angulares

2.1.1 Direção Média

O problema do cálculo da direção média para dados circulares foi ilustrado por Barriga (1997), considerando a representação na Figura 1 de três

direções dadas pelos ângulos θ_1=80º, θ_2=350º e θ_3=50º. Visualmente espera-se que o ângulo médio assuma um valor entre 0º e 50º. Ao calcular-se a média aritmética $(\theta_1+\theta_2+\theta_3)/3$ obtém-se como média o valor 160º, que não corresponde à situação. De acordo com Barriga (1997), o simples procedimento de obtenção de uma média aritmética pode não representar uma estatística adequada para representação da direção média.

Considerando, então, a representação gráfica da disposição dos ângulos de uma amostra em um círculo, pode se associar a cada ponto uma massa de igual valor *M* e encontrar seu centro de massa *G* ou centro de gravidade (BARRIGA, 1997). O vetor $\overrightarrow{OG}$ que aponta para o centro de gravidade é definido como vetor médio amostral e determina uma direção média amostral que é o ângulo médio $\overline{\theta}$. A Figura 2 representa o centro de gravidade:

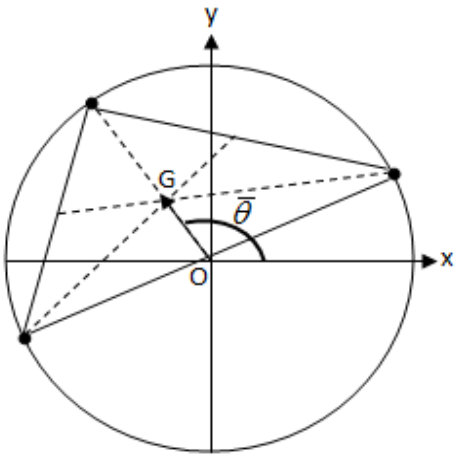


Figura 2 Representação gráfica do centro de gravidade de uma amostra com medidas angulares

Fonte: Adaptado de Barriga (1997)

Duas formas para determinação do vetor média (ou ângulo médio) podem ser utilizadas: determinação por meio de álgebra vetorial e determinação pelas funções trigonométricas.

a) Determinação do vetor média por meio de álgebra vetorial

Seja uma amostra angular $\theta_1,...,\theta_n$ as quais estão associados os vetores unitários correspondentes $\overrightarrow{OP}_1,\overrightarrow{OP}_2,\ldots,\overrightarrow{OP}_n$. Na Figura 3 tem-se a representação de um vetor

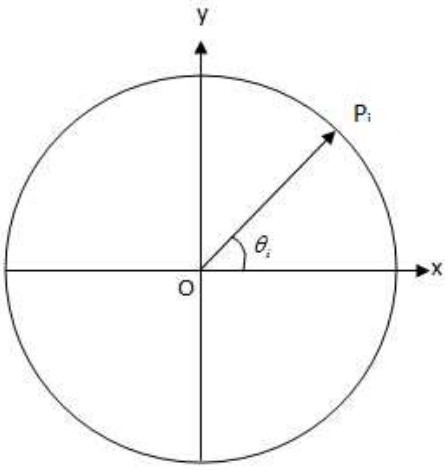


Figura 3 Representação do i -ésimo ponto amostral
Fonte Barriga (1997)

Conforme Figura 2, atribuindo massa *M* a cada ponto, o vetor

$$\overrightarrow{OG} = \frac{1}{n}\sum_{i=1}^{n}\overrightarrow{OP_i} \qquad ((1)$$

Aponta para o centro de massa de $P_1, P_2, ..., P_n$. Dessa forma $\overrightarrow{OG}$ é o vetor média da amostra. Considera-se, ainda, *R* como comprimento resultante e *r* o comprimento do vetor média, ou seja:

$$R = \left\| \sum_{i=1}^{n} \overrightarrow{OP_i} \right\| \quad \text{e} \quad r = \left\| \overrightarrow{OG} \right\| = \frac{R}{n} \tag{2}$$

b) Determinação do vetor média por meio de funções trigonométricas

Na ocorrência de vários vetores, algebricamente, calculam-se as médias com base nas coordenadas do centro de massa do sistema (MARDIA, 1972), que são:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} \cos(\theta_i) = \overline{C} \text{ e } \overline{y} = \frac{1}{n} \sum_{i=1}^{n} sen(\theta_i) = \overline{S} \tag{(3)}$$

O comprimento do vetor média é então:

$$r = \left(\overline{C}^2 + \overline{S}^2 \right)^{\frac{1}{2}} \tag{(4)}$$

e, assim, a partir das expressões apresentadas em (3), a direção média $\overline{\theta}$ é dada por:

$$\bar{\theta} = \begin{cases} \arctan\left(\dfrac{\bar{S}}{\bar{C}}\right) & \text{se } \bar{S} > 0, \bar{C} > 0 \\ 180^{o} + \arctan\left(\dfrac{\bar{S}}{\bar{C}}\right) & \text{se } \bar{C} < 0 \\ 360^{o} + \arctan\left(\dfrac{\bar{S}}{\bar{C}}\right) & \text{se } \bar{S} < 0, \bar{C} > 0 \end{cases} \quad ((5)$$

São casos específicos:

$$\bar{\theta} = \begin{cases} 90^{o} & \text{se } \bar{S} > 0, \bar{C} = 0 \\ 270^{o} & \text{se } \bar{S} < 0, \bar{C} = 0 \\ \text{não determinado} & \text{se } \bar{S} = 0, \bar{C} = 0 \end{cases} \quad (6)$$

2.1.2 Variância Circular

Considerando $P_1, P_2, ..., P_n$ como vetores unitários e sabendo que $r = \left(\bar{C}^2 + \bar{S}^2\right)^{\frac{1}{2}}$, então, $0 \le r \le 1$. Se os ângulos $\theta_1, \theta_2, ..., \theta_n$ estão mais agrupados *r* é mais próximo de 1. Por outro lado, se $\theta_1, \theta_2, ..., \theta_n$ estão mais dispersos *r* será mais próximo de 0 (MARDIA, 1972). Assim nota-se que *r* é uma medida de concentração do conjunto de dados. Ainda, se $r = 0$ mostra que todos os pontos estão em uma dispersão uniforme pelo círculo e se $r = 1$, todos os pontos são coincidentes.

A variância circular amostral é, então, definida como:

$$V_c = 1 - r \quad (7)$$

Salienta-se, ainda, que $0 \leq V_c \leq 1$, o que não acontece em dados na reta. Assim, quanto menor o valor de V_c, mais homogênea é a amostra.

2.1.3 Desvio padrão circular

O desvio padrão circular amostral é definido como (MARDIA, 1972):

$$s_c = \left[-2\ln\left(1-V_c\right)\right]^{\frac{1}{2}} \tag{8}$$

Em que V_c é a variância circular.

Para casos onde V_c tende a zero ou assume valores muito pequenos, utiliza-se uma aproximação de s_c, dada por:

$$s_c \approx \left(2V_c\right)^{\frac{1}{2}} \tag{9}$$

O desvio padrão circular não pode ser obtido simplesmente como a raiz quadrada da variância circular.

2.1.4 Amplitude circular

De acordo com Mardia (1972), a amplitude circular é o menor arco que contém todas as observações. Para sua determinação consideram-se os *n* ângulos, $\theta_1, \theta_2, ..., \theta_n$ no intervalo $0 \leq \theta_i \leq 2\pi$. Sendo $\theta_{(1)} \leq \ldots \leq \theta_{(n)}$ as estatísticas de ordem de $\theta_1, \theta_2, ..., \theta_n$, o comprimento do arco entre as observações adjacentes são:

$$T_i = \theta_{(i+1)} - \theta_{(i)},\ i = 1,...,n-1 \text{ e} \tag{10}$$

$$T_n = 360^{\circ} - \theta_{(n)} + \theta_{(1)} \tag{11}$$

Em que T_i são medidas dos comprimentos de arco entre pontos consecutivos.

Assim a amplitude circular *w* é dada por:

$$w = 360^{\circ} - \max(T_1,...,T_n) \tag{12}$$

2.1.5 Distribuição de Von Mises

A distribuição de Von Mises é a base das inferências estatísticas para dados circulares (BARRIGA, 1997). É considerada um caso análogo no círculo a distribuição normal em dados lineares. Esta distribuição foi introduzida por Von Mises, em 1918, para estudar desvios de medidas de pesos atômicos.

Se Θ é uma variável aleatória circular, então, a função densidade de probabilidade de uma distribuição Von Mises é dada por (ABUZAID et al., 2012):

$$f(x) = \frac{1}{2\pi I_0(q)} \exp\left[q\cos(\theta - \mu)\right],\ 0 \le \theta \le 2\pi,\ q \ge 0 \tag{13}$$

Em que μ é o parâmetro que representa a direção média, q um parâmetro associado à concentração e $I_0(q)$ é a função de Bessel modificada de ordem zero:

$$I_0(q) = \sum_{r=0}^{\infty} (r!)^{-2} \left(\frac{1}{2}q\right)^{2r} \quad ((14)$$

Mais detalhes são encontrados em Mardia (1972).

Denotamos a distribuição Von Mises por $VM(\mu, q)$. Quanto ao parâmetro *q* da distribuição, quanto mais próximo de zero, mais os dados estão uniformemente distribuídos ao redor do círculo, ao passo que quanto mais tender ao infinito, mais os dados estarão concentrados em sua direção média.

2.2 Detecção de *outliers* em dados circulares

Em se tratando da aplicabilidade de metodologias para detecção de *outliers* em dados circulares, Ibrahim et al. (2013) propuseram uma metodologia para um modelo de regressão circular utilizando a estatística *COVRATIO*.

Um modelo de regressão para duas variáveis aleatórias circulares *U* e *V* chamado modelo de regressão circular JS (JAMMALAMADAKA; SARMA, 1993) pode ser escrito em termos da esperança condicional e^{iv} dado *u* tal que

$$E\left(e^{iv} \mid u\right) = \rho(u) e^{i\mu(u)} = g_1(u) + i g_2(u) \quad ((15)$$

em que $e^{iv} = \cos(v) + i\text{sen}(v)$, $\mu(u)$ representa a direção média condicional de *v* dado *u* e $\rho(u)$ o parâmetro de concentração para as funções periódicas $g_1(u)$ e $g_2(u)$ (JAMMALAMADAKA; SARMA, 1993). Pode se escrever:

$$\begin{aligned} E\left(\cos(v)\,|\,u\right) &= g_1(u) \\ E\left(\operatorname{sen}(v)\,|\,u\right) &= g_2(u) \end{aligned} \qquad ((16)$$

Segundo os autores, *v* pode ser predito da seguinte forma:

$$\mu(u) = \hat{v} = \arctan\left(\frac{g_2(u)}{g_1(u)}\right) = \begin{cases} \arctan\left(\dfrac{g_2(u)}{g_1(u)}\right) & \text{se } g_1(u) > 0 \\ 180^{\circ} + \arctan\left(\dfrac{g_2(u)}{g_1(u)}\right) & \text{se } g_1(u) \le 0 \\ \text{indefinido} & \text{se } g_1(u) = g_2(u) = 0 \end{cases} \qquad (17)$$

As aproximações utilizadas são polinômios trigonométricos ajustados, de grau *m* e da forma:

$$\begin{aligned} g_1(u) &\approx \sum_{h=0}^{m}\left(A_h\cos(hu) + B_h\operatorname{sen}(hu)\right) \\ g_2(u) &\approx \sum_{h=0}^{m}\left(C_h\cos(hu) + D_h\operatorname{sen}(hu)\right) \end{aligned} \qquad ((18)$$

Com isso, têm-se os seguintes modelos:

$$\begin{aligned} \cos(v) &= \sum_{h=0}^{m}\left(A_h\cos(hu) + B_h\operatorname{sen}(hu)\right) + \varepsilon_1 \\ \operatorname{sen}(v) &= \sum_{h=0}^{m}\left(C_h\cos(hu) + D_h\operatorname{sen}(hu)\right) + \varepsilon_2 \end{aligned} \qquad ((19)$$

Em que $[0\ 0]^T$ $\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$ é um vetor dos erros aleatórios seguindo uma distribuição normal com vetor de médias $\mathbf{0}$ e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$ desconhecidas. Os parâmetros, A_h, B_h, C_h, D_h, os erros padrão e a matriz $\boldsymbol{\Sigma}$ podem ser estimados.

A estatística *COVRATIO* foi proposta por Belsley, Edwin e Roy (1980) e consiste na investigação do impacto de se eliminar uma linha por vez do conjunto de dados nos coeficientes estimados, valores ajustados, resíduos e matriz de covariâncias.

Ibrahim et al. (2013) desenvolveram, então, procedimento similar para dados circulares, atuando na matriz de variâncias e covariâncias do modelo de regressão circular JS. Para tal, elimina-se uma linha dos dados e verifica-se o efeito da eliminação na razão entre a matriz de variâncias e covariâncias, estimada com todas as observações disponíveis e a matriz de variâncias e covariâncias com a *j*-ésima observação eliminada.

$$COVRATIO_{(-j)} = \frac{|COV|}{\left|COV_{(-j)}\right|} \qquad ((20)$$

Em que $|COV|$ é o determinante da matriz de covariâncias para conjunto de dados completo e $\left|COV_{(-j)}\right|$ o determinante da matriz de covariâncias que exclui a *j*-ésima linha.

É utilizada, como teste de detecção, a quantidade $\left|COVRATIO_{(-j)} - 1\right|$, obtendo um ponto de corte, tabelado, obtido pelas simulações prévias que estabeleceram percentis superiores a 5% para vários tamanhos amostrais. Para

obtenção dos percentis foram geradas variáveis de uma distribuição Von-Mises e, também, erros aleatórios (ε_1 e ε_2) de tamanho *n* considerando uma distribuição Normal, como vetor de médias $[0\ 0]^T$ e matriz de variâncias e covariâncias $\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$. A partir deste ponto é ajustado um modelo de regressão circular e calculado $|COV|$. Exclui-se a *j*-ésima linha da amostra gerada e para cada *j* é reajustado o modelo obtendo $|COV_{(-j)}|$. A quantidade $|COVRATIO_{(-j)} - 1|$ também é obtida para cada *j* e observa-se o seu máximo valor, obtendo os pontos de corte tabelados.

2.3 Distâncias para dados circulares

Considerando $\theta_1, \theta_2, ..., \theta_n$ observações circulares alocadas na circunferência de círculo unitário, Jammalamadaka e Sengupta (2001) definiram a distância circular entre dois ângulos, θ_i e θ_j como:

$$d_{ij} = 1 - \cos(\theta_i - \theta_j) \tag{21}$$

Como medida de dissimilaridade, Abuzaid Mohamed e Hussin (2009) utilizaram a distância circular para propor um teste de discordância, dado pela estatística *B* definida por:

$$B = \max_{1 \le j \le n} \left\{ \frac{D_j}{2(n-1)} \right\} \tag{22}$$

Em que $D_j = \sum_{i=1}^{n} d_{ij}$. *B* fornece, então, um ponto para análise de possível ocorrência de observações discrepantes. Se θ_j é um *outlier*, o valor de D_j é aumentado, atuando diretamente na estatística, que é comparada a percentis pré-estabelecidos e tabelados, baseados no parâmetro de concentração *q* da distribuição Von-Mises.

Jammalamadaka e Sengupta (2001) consideram uma definição alternativa θ_{ij} da distância circular, em termos de ângulo, entre dois pontos θ_i e θ_j. Para tal, considera-se a representação dos pontos na Figura 4.

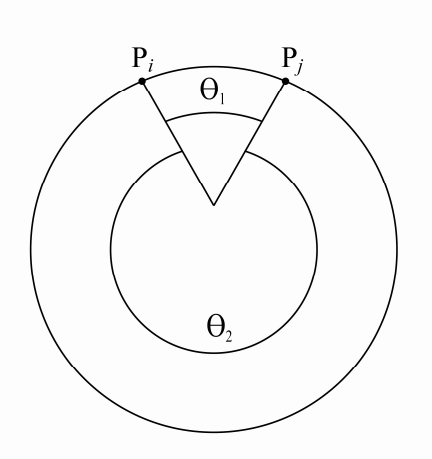


Figura 4 Representação das distâncias angulares entre dois pontos

Pode-se observar na Figura 4, a existência de duas possíveis distâncias θ_{ij} formadas entre P_i e P_j, pelos ângulos θ_1 e θ_2. Dessa forma ao calcular a distância pela expressão:

$$\theta_{ij} = 180^{\circ} - \left| 180^{\circ} - \left| \theta_i - \theta_j \right| \right|, \ \theta_{ij} \in \left[0,180^{\circ} \right] \tag{23}$$

É assegurado que θ_{ij} assuma valores com menores ângulos entre P_i e P_j.

2.4 Componentes Principais

A Análise de Componentes Principais (ACP) é uma técnica estatística multivariada introduzida por Karl Pearson em 1901, tendo sua consolidação em Hotelling (1933).

O principal objetivo da ACP é explicar a estrutura da variância de um vetor aleatório, composto de p-variáveis, utilizando combinações lineares das variáveis originais. Estas combinações lineares são chamadas de CP e são não correlacionadas entre si.

É possível reter tantos Componentes Principais quanto forem o número p de variáveis, ou seja, considerando p variáveis, consegue-se reter p Componentes Principais, onde cada componente retém uma porcentagem da variação original dos dados.

Haja vista que um dos principais objetivos é a redução de dimensão, não há sentido prático em reter p componentes. Procura-se, então, um número $k < p$ de componentes que explique, satisfatoriamente, a variação total contida no conjunto de dados. Assim, condensa-se a informação contida em um conjunto de p variáveis em um número menor de k variáveis, procurando uma perda mínima de informação (HAIR JÚNIOR et al., 2005).

Segundo Johnson e Wichern (2007), geometricamente, as combinações lineares representam a seleção de um novo sistema de coordenadas, obtidas pela rotação dos eixos do sistema original de coordenadas. As Figuras 5 e 6 abaixo exemplificam o sistema:

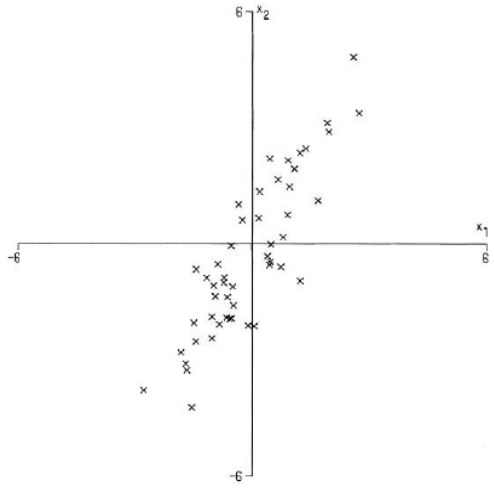


Figura 5 Eixo original do sistema

Fonte Jolliffe (2002)

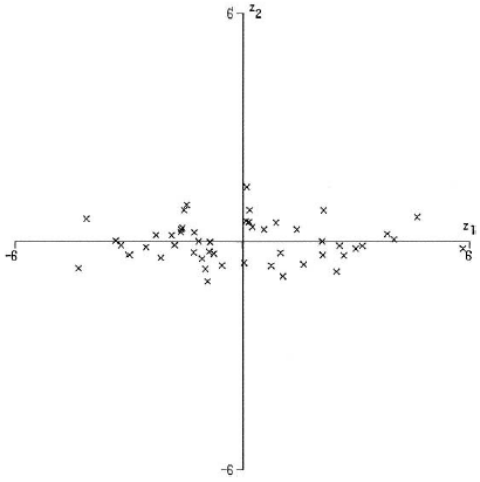


Figura 6 Novo eixo formado pelos Componentes Principais

Fonte Jolliffe (2002)

Observa-se, então, na Figura 6, que os Componentes Principais promoveram uma rotação do eixo de coordenadas no sentido de maior variação dos dados.

Supondo um vetor $\mathbf{X}$ de p variáveis aleatórias com matriz de variância e covariância $\mathbf{\Sigma}$, o primeiro passo é procurar por uma função linear $\mathbf{e}_1^T\mathbf{X}$ dos elementos de $\mathbf{X}$ que tenha máxima variância. Essa função tem a seguinte forma:

$$\mathbf{e}_1^T\mathbf{X} = \mathbf{e}_{11}x_1 + \mathbf{e}_{12}x_2 + \ldots + \mathbf{e}_{1p}x_p \tag{24}$$

O próximo passo é procurar por uma segunda função linear $\mathbf{e}_2^T\mathbf{X}$, não correlacionada com $\mathbf{e}_1^T\mathbf{X}$ também com uma variância máxima e, assim, sucessivamente. O p -ésimo componente principal será dado por $\mathbf{e}_p^T\mathbf{X}$.

Considerando $\mathbf{e}_1^T\mathbf{X}$, $\mathbf{e}_1$ deve-se maximizar $\text{var}\left[\mathbf{e}_1^T\mathbf{X}\right] = \mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1$. Para tal, deve se impor a restrição $\mathbf{e}_1^T\mathbf{e}_1 = 1$, a fim de que o máximo seja atingido para um $\mathbf{e}_1$ finito.

A maximização de $\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1$ sujeita à restrição $\mathbf{e}_1^T\mathbf{e}_1 = 1$ pode ser obtida, utilizando a técnica de multiplicadores de Lagrange, ou seja, maximizar

$$\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 - \lambda\left(\mathbf{e}_1^T\mathbf{e}_1 - 1\right) \tag{25}$$

onde λ é o multiplicador de Lagrange. A diferenciação em relação a $\mathbf{e_1}$ resulta em:

$$\mathbf{\Sigma}\mathbf{e}_1 - \lambda\mathbf{e}_1 = 0 \text{ ou } \left(\mathbf{\Sigma} - \lambda\mathbf{I}_p\right)\mathbf{e}_1 = 0 \tag{26}$$

em que $\mathbf{I}_p$ é uma matriz identidade de ordem p. Assim, λ é um autovalor de $\mathbf{\Sigma}$ e $\mathbf{e}_1$ o autovetor correspondente. Para decidir a respeito de qual dos p autovetores resultará na máxima variância para $\mathbf{e}_1^T\mathbf{X}$, basta notar que

$$\mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 = \mathbf{e}_1^T\lambda\mathbf{e}_1 = \lambda\mathbf{e}_1^T\mathbf{e}_1 = \lambda \tag{27}$$

Dessa forma λ deve ser tão grande quanto possível. $\mathbf{e}_1$ é então, o autovetor correspondente ao maior autovalor de $\mathbf{\Sigma}$ e $\text{var}\left[\mathbf{e}_1^T\mathbf{X}\right] = \mathbf{e}_1^T\mathbf{\Sigma}\mathbf{e}_1 = \lambda_1$, o maior autovalor (JOLLIFFE, 2002).

No geral, o p-ésimo componente principal de $\mathbf{X}$ é $\mathbf{e}_p^T\mathbf{X}$ e a $\text{var}\left[\mathbf{e}_p^T\mathbf{X}\right] = \lambda_p$, em que λ_p é o p-ésimo maior autovalor de $\mathbf{\Sigma}$.

2.4.1 Componentes Interpretáveis

Em alguns casos, os coeficientes dos Componentes Principais podem requerer algum tipo de interpretação ou, ainda, ter identificadas as direções, sendo estas dificultadas pelos valores assumidos pelos mesmos.

Para interpretar os componentes, existe uma alternativa que é filtrar os coeficientes das combinações lineares e identificar padrões. Chipman & Gu (2005) apresentam métodos para simplificar as combinações, tornando-as mais interpretáveis. Outros métodos de simplificação das combinações podem ser encontrados em Vines (2000).

Por coeficiente interpretável entende-se uma redução nas possíveis combinações que os coeficientes podem assumir. Por exemplo, muitos coeficientes podem assumir valores zero ou, ainda, os coeficientes dos

componentes podem assumir alguns poucos e distintos valores, que é o caso apresentado pela restrição de homogeneidade.

2.4.1.1 Restrição de Homogeneidade

Considerando p variáveis, existem métodos para identificar direções mais interpretáveis para os componentes sob restrição. Esses Componentes Interpretáveis são chamados de $\boldsymbol{\alpha}_i$, $i = 1,..., p$.

A i-ésima direção de $\boldsymbol{\alpha}_i$ pode ser mais interpretável se seus elementos assumirem poucos e distintos valores, como 0 ou $\pm c$, considerando um valor de c que permita que $\boldsymbol{\alpha}_i^T \boldsymbol{\alpha} = 1$. Esta restrição de homogeneidade pode corresponder a uma direção que é a média de algumas variáveis (CHIPMAN; GU, 2005).

A restrição de homogeneidade fixa como $\pm c$, a quantidade $\pm \frac{1}{\sqrt{k}}$, com $k = 1,..., p$ variáveis, sendo uma constante normalizadora.

Existem 3^p possíveis valores de $\boldsymbol{\alpha}_i$. Para encontrar o melhor, minimiza-se o $\arccos\left(\mathbf{e}_i^T \boldsymbol{\alpha}_i\right)$, que é o ângulo entre a i-ésima direção do componente principal e o componente interpretável. Equivalentemente pode-se, também, proceder à maximização do produto direto entre $\mathbf{e}_i^T \boldsymbol{\alpha}_i$ (CHIPMAN; GU, 2005).

Como um exemplo de ilustração, suponha um vetor de coeficientes $\mathbf{e}_1$ de um primeiro componente principal, com p =4. Seja então:

$$\mathbf{e}_1 = \begin{bmatrix} 0,41 & -0,03 & -0,42 & 0,81 \end{bmatrix}^T$$

O próximo passo é encontrar o $\boldsymbol{\alpha}_i$ que seja o mais próximo possível de $\mathbf{e}_1$. Como a regra é procurar o $\boldsymbol{\alpha}_i$ em $\pm\frac{1}{\sqrt{k}}$, tem-se as seguintes opções:

$$\pm\frac{1}{\sqrt{1}}\ ,\ \pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{3}} \text{ ou} \pm\frac{1}{\sqrt{4}}$$

Assim, alguns possíveis candidatos são:

$$\boldsymbol{\alpha}_1 = \frac{\begin{bmatrix}0 & 0 & 0 & 1\end{bmatrix}^T}{\sqrt{1}},\ \boldsymbol{\alpha}_1 = \frac{\begin{bmatrix}0 & 0 & -1 & -1\end{bmatrix}^T}{\sqrt{2}},$$
$$\boldsymbol{\alpha}_1 = \frac{\begin{bmatrix}1 & 0 & -1 & 1\end{bmatrix}^T}{\sqrt{3}},\ \boldsymbol{\alpha}_1 = \frac{\begin{bmatrix}1 & -1 & -1 & 1\end{bmatrix}^T}{\sqrt{4}}$$

Neste caso, o $\boldsymbol{\alpha}_1$ mais próximo de $\mathbf{e}_1 = \begin{bmatrix}0{,}41 & -0{,}03 & -0{,}42 & 0{,}81\end{bmatrix}^T$ é $\boldsymbol{\alpha}_1 = \frac{\begin{bmatrix}1 & 0 & -1 & 1\end{bmatrix}^T}{\sqrt{3}}$, com um ângulo de 18,8 graus. Observa-se, ainda, que existe uma correspondência de sinais de elementos não próximos a zero.

Note que agora os componentes são mais interpretáveis, visto que os valores que c assume são, 0 e ± 1. A constante normalizadora $\sqrt{3}$ pode ser omitida para efeito de comparação, já que é comum a todos os elementos.

Um exemplo prático de aplicação dos Componentes Interpretáveis foi apresentado em Chipman e Gu (2005), em um estudo sobre características de carros vendidos nos Estados Unidos em 1993, representados na Tabela 1. As características avaliadas foram: preço mínimo, preço, preço máximo, consumo na cidade, consumo em rodovia, tamanho do motor, HP, RPM, revoluções por milhas, capacidade do tanque, passageiros, comprimento, distância entre pneus,

largura, distância de viragem, assento traseiro e peso. São apresentados os coeficientes dos Componentes Principais e os coeficientes dos Componentes Interpretáveis correspondentes.

Para os Componentes Interpretáveis apresentados na tabela, ressaltou-se que o primeiro componente referiu-se ao tamanho dos carros, com coeficientes positivos relacionados, positivamente, com o tamanho e coeficientes negativos seguindo a lógica contrária. O segundo Componente Interpretável pode ser interpretado como um contraste entre carros baratos, fracos e grandes versus carros caros, potentes e pequenos (CHIPMAN; GU, 2005).

Tabela 1 Direções interpretáveis e ângulos correspondentes às direções dos componentes

	Componentes Principais				Componentes Interpretáveis			
Variável	1	2	3	4	1	2	3	4
Preço Min	0,230	-0,376	-0,118	-0,154	1	-1	0	0
Preço	0,220	-0,421	-0,131	-0,114	1	-1	-1	0
Preço Max	0,203	-0,439	-0,136	-0,077	1	-1	-1	0
Consumo C.	-0,265	0,002	-0,103	-0,450	-1	0	0	-1
Consumo. R.	-0,247	0,013	-0,005	-0.611	-1	0	0	-1
Motor	0,282	0,050	0,184	-0,202	1	0	1	-1
HP	0,243	-0,289	0,190	-0,005	1	-1	1	0
RPM	-0,141	-0,411	-0,149	0,140	-1	-1	-1	0
Rev/milha	-0,241	-0,135	-0,344	0,126	-1	0	-1	0
Tanque	0,273	0,004	-0,064	0,214	1	0	0	1
Passageiro	0,192	0,321	-0,461	0,231	1	1	-1	1
Comprimento	0,263	0,073	0,058	-0,295	1	0	0	-1
Dist.Pneus	0,275	0,108	-0,172	-0,130	1	0	-1	0

"Tabela 1, conclusão"

	Componentes Principais				Componentes Interpretáveis			
Variável	1	2	3	4	1	2	3	4
Largura	0,271	0,163	0,189	-0,105	1	1	1	0
Dist.Viragem	0,247	0,175	0,196	-0,117	1	1	1	0
Assento T	0,178	0,195	-0,637	-0,260	1	1	-1	-1
Peso	0,295	0,011	0,017	0,097	1	0	0	0
Ângulo (º)					10^{o}	22^{o}	33^{o}	31^{o}

Fonte: Chipman e Gu (2005).

2.5 Distribuições Assimétricas

Estudos envolvendo distribuições assimétricas são encontrados nas mais diversas áreas.

O objetivo de desenvolvimento de trabalhos dessa classe de distribuições foi a obtenção de distribuições paramétricas que representassem uma transição da normalidade para a não normalidade, considerando parâmetros específicos que controlassem posição, escala e forma da distribuição.

Azzalini (1985) apresenta o seguinte lema, para definir uma distribuição assimétrica:

Lema: Seja f uma função densidade de probabilidade simétrica em torno de 0, e G uma função de distribuição acumulada, absolutamente contínua, tal que G' é simétrica em torno de 0. Então,

$$2f(x)G(\lambda x) \tag{28}$$

é uma função densidade de probabilidade para qualquer número real λ. Observa-se, ainda, que $(-\infty < x < \infty)$.

Levando em consideração o lema e a função descrita em (28), muitas distribuições podem ser propostas, sendo mais comum a manipulação de distribuições que possuam função densidade de probabilidade e função de distribuição acumulada mais conhecida.

Ao se definir, por exemplo, f e G, como a função densidade de probabilidade e a função de distribuição acumulada de uma variável aleatória normal padrão, respectivamente, tem-se como resultado a distribuição normal assimétrica, que será definida a seguir.

2.5.1 Distribuição Normal Assimétrica

A distribuição normal assimétrica foi inicialmente introduzida por Azzalini (1985), a partir do lema da seção 2.5.

Definição: Se uma variável aleatória Z, com parâmetro de assimetria λ, tem a seguinte função densidade:

$$f(z;\lambda) = 2\varphi(z)\Phi(\lambda z) \qquad (-\infty < z < \infty) \tag{29}$$

Em que λ é definido em $\square$, φ e Φ são, respectivamente, a densidade e a função distribuição da normal padrão, então Z segue uma distribuição normal assimétrica com parâmetro λ. Resumidamente, denota-se: $Z \sim NA(\lambda)$.

Propriedades:

A densidade de $NA(0)$ é igual à densidade de $N(0,1)$. Em outras palavras, uma distribuição normal assimétrica, com parâmetro de assimetria igual a 0, é uma distribuição normal padrão.

a) Se Z é uma variável aleatória $NA(\lambda)$, então $-Z$ é uma variável aleatória $NA(-\lambda)$.

b) Se $Z \sim NA$, então $Z^2 \sim \chi^2_1$.

2.5.2 Distribuição Normal Assimétrica com parâmetros de posição e escala

A forma da distribuição normal assimétrica apresentada em (29), ainda, pode ser reescrita com adição de parâmetros de posição e escala.

Definição: Uma variável aleatória Y tem distribuição normal assimétrica com parâmetro de assimetria λ, parâmetro de posição μ e, também, parâmetro de escala σ, com a seguinte densidade:

$$f(y) = 2\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\left(\frac{y-\mu}{\sigma}\right)\right), y \in \mathbb{R} \quad (30)$$

Em que $\phi(.)$ e $\Phi(.)$ são, respectivamente, a função densidade de probabilidade e a função de distribuição acumulada da distribuição normal padrão. Salienta-se, ainda, que ao se considerar duas variáveis aleatórias, Z e Y, com $Z \sim NA(\lambda)$ e $Y = \mu + \sigma Z$, então $Y \sim NA(\mu, \sigma, \lambda)$. Verifica-se, então, que qualquer combinação linear de uma variável aleatória normal assimétrica padrão, também, tem distribuição normal assimétrica.

Propriedades (AZZALINI, 1986):

a) Se $Y \sim NA(\mu,\sigma,\lambda)$ então $X = a + bY \sim NA(a+b\mu, b\sigma, \lambda), a, b \in \mathbb{R}$.

b) A função geradora de momentos de Y

$$M_y(t) = 2\exp\left(\frac{(t-\mu)^2}{2\sigma^2}\right)\Phi\left(\delta\left(\frac{y-\mu}{\sigma}\right)\right) \quad (31)$$

em que $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

c) Y tem como média:

$$E(Y) = \mu + \sigma\delta\sqrt{\frac{2}{\pi}} \quad (32)$$

d) Y, tem como variância:

$$Var(Y) = \sigma^2\left(1 - \frac{2}{\pi}\delta^2\right) \quad (33)$$

A Figura 7 apresenta o comportamento da distribuição normal assimétrica, considerando diferentes valores do parâmetro de assimetria.

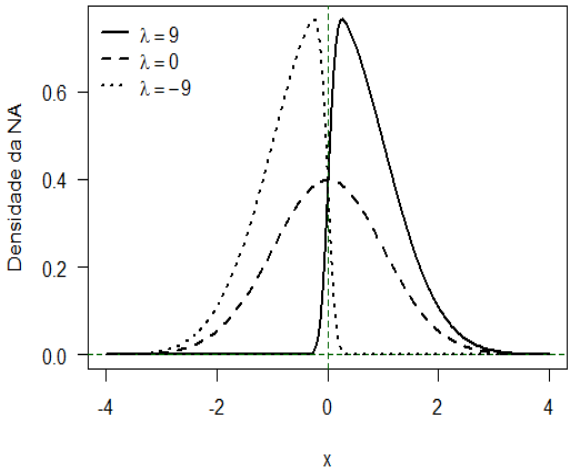


Figura 7 Funções densidade de NA(9),NA(-9) e NA(0)

2.5.3 Distribuição Normal Assimétrica Multivariada

A distribuição normal assimétrica multivariada foi, originalmente, introduzida de forma sucinta por Azzalini (1985) como uma extensão da normal assimétrica apresentada em (29) e, posteriormente, desenvolvida de forma mais completa por Azzalini e Valle (1996).

A necessidade da generalização do caso univariado para o multivariado deu-se, segundo os autores, pelo relevante potencial de aplicação de tal distribuição, afirmando que, no caso multivariado, há uma maior escassez de distribuições disponíveis para tratamento de dados multivariados e não normais.

As famílias das distribuições normais multivariadas assimétricas possuem, como distribuições marginais, as normais assimétricas univariadas e como um de seus membros, a distribuição normal multivariada.

Para apresentação do caso multivariado, considera-se como exemplo p características ou variáveis. Segundo Azzalini e Valle (1996), uma variável

aleatória Z, p-dimensional, tem uma distribuição normal assimétrica multivariada, se é contínua e com a seguinte função densidade:

$$f_p(\mathbf{z}) = 2\varphi_p(\mathbf{z}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\alpha}^T\mathbf{z}), \text{ com } \mathbf{z} \in \mathbb{R}^p \quad (34)$$

em que $\varphi_p(\mathbf{z}, \boldsymbol{\Sigma})$ representa a densidade da distribuição normal p-multivariada com vetor de média 0 e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$; $\Phi(.)$ é uma função distribuição normal padrão e $\boldsymbol{\alpha}$ é um vetor p-dimensional do parâmetro de forma.

Assim, quando $\boldsymbol{\alpha}$ é igual a $\mathbf{0}$, a função densidade (34) reduz-se à normal multivariada.

2.5.4 Distribuição Normal Assimétrica Multivariada com parâmetros de posição e escala

A densidade apresentada em (34) não incorpora parâmetros de posição e escala que são essenciais para trabalhos estatísticos práticos (AZZALINI; CAPITANIO, 1999).

Com esta premissa, Azzalini e Capitanio (1999) introduziram estes parâmetros, até então omitidos, na função densidade de $\mathbf{Z}$. Assim, considera-se então:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\omega}\mathbf{Z} \quad (35)$$

Em que $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$ e $\boldsymbol{\omega} = diag(\sigma_{11}, \ldots, \sigma_{pp})$ são, respectivamente, os parâmetros de posição e escala. Os componentes de $\boldsymbol{\omega}$ são positivos.

A função densidade de $\mathbf{Y}$ é então:

$$f_p(\mathbf{y}) = 2\varphi_p(\mathbf{y}-\boldsymbol{\mu},\boldsymbol{\Sigma})\Phi\left\{\boldsymbol{\alpha}^T\boldsymbol{\omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\} \quad (36)$$

Em que $\boldsymbol{\Sigma}$ é uma matriz de covariância e a notação (SILVA; PINTO JUNIOR, 2010):

$$\mathbf{Y} \sim NA_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha}) \quad (37)$$

é utilizada para indicar que $\mathbf{Y}$ tem uma função densidade multivariada assimétrica com parâmetro de posição e escala, conforme (36).

A Figura 8 é uma representação gráfica, para comparação, de uma normal bivariada e uma normal assimétrica bivariada com $\boldsymbol{\mu} = [-0{,}1 \quad 0{,}1]$, $\boldsymbol{\alpha} = [-5 \quad 5]^T$ e $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{bmatrix}$.

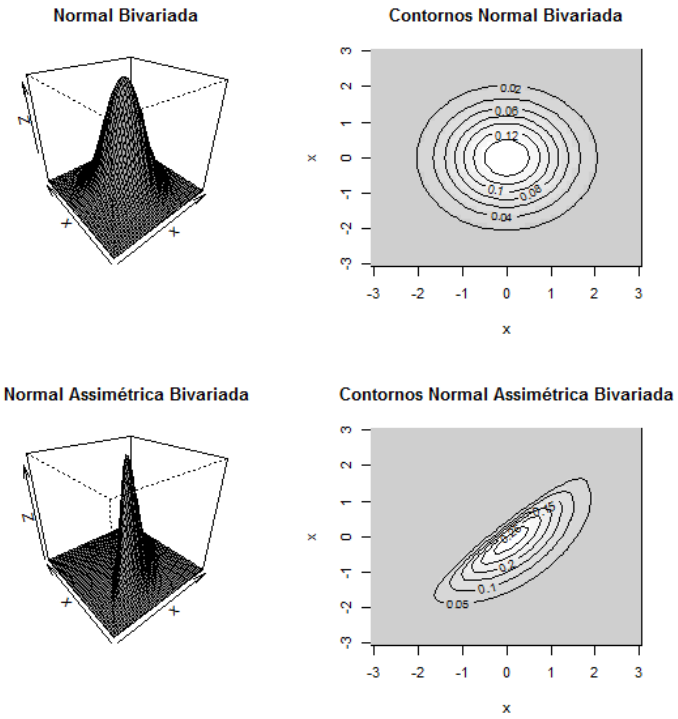


Figura 8 Representação gráfica da normal bivariada

2.6 Distribuição normal multivariada contaminada

A distribuição normal multivariada contaminada é muito importante para realização de certos tipos de estudo, principalmente, os de simulação em que se envolvem *outliers*.

Seja um vetor aleatório $\mathbf{X} = \left[X_1, \ldots, X_p \right]^T \in \mathbb{R}^p$ com distribuição normal multivariada contaminada. Sua função de densidade de probabilidade será

$$f(\mathbf{x}) = (1-\delta)(2\pi)^{-\frac{p}{2}}\left|\boldsymbol{\Sigma}_1\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right] - \\ +\delta(2\pi)^{-\frac{p}{2}}\left|\boldsymbol{\Sigma}_2\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \quad (38)$$

Em que $(1-\delta)$ é a probabilidade do processo ser realizado por uma $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ e δ a probabilidade que o processo tem de ser realizado por $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. $\boldsymbol{\Sigma}_i$ é uma matriz de variâncias e covariâncias e $\boldsymbol{\mu}_i$ o vetor de médias, $i = 1,2$ e $0 \leq \delta \leq 1$ (JOHNSON, 1987).

2.7 Distribuição t-Student multivariada

A distribuição t-Student multivariada pertence à família das distribuições elípticas e pode ser, também, utilizada para avaliar desvios de normalidade dos dados.

Para defini-la, considere um vetor $\mathbf{X} = \left[X_1, \ldots, X_p\right]^T \in \mathbb{R}^p$ com

$$f(\mathbf{x}) = \frac{\left|\boldsymbol{\Sigma}\right|^{-1/2}\boldsymbol{\Gamma}\left[(\nu+p)/2\right]}{\left[\boldsymbol{\Gamma}(1/2)\right]^p\boldsymbol{\Gamma}(\nu/2)\nu^{p/2}}\left[1+\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+p}{2}} \quad (39)$$

Assim, $\mathbf{X}$ tem distribuição t multivariada com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ com ν graus de liberdade com a notação $\mathbf{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ (LANGE; RODERICK; TAYLOR, 1989).

2.8 Distribuição log-normal multivariada

A distribuição log-normal multivariada pode ser facilmente declarada:

Se $\mathbf{X}$ tem distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$, segundo Kotz, Balakrishnan e Johnson (2004), $\mathbf{Y}$ tem uma distribuição log-normal multivariada, se seu logaritmo converge em distribuição para $\mathbf{X}$, ou seja, $\log \mathbf{Y} \xrightarrow{d} \mathbf{X}$. Sua função densidade é dada por:

$$f(\mathbf{y}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \mathbf{y}^{-1} \exp\left[-\frac{1}{2} (\ln \mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\ln \mathbf{y} - \boldsymbol{\mu}) \right] \qquad ((40)$$

Assim, $\mathbf{Y}$ segue uma distribuição log-normal multivariada com parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ e tem a notação $\mathbf{X} \sim LN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

3 METODOLOGIA

Em consonância com os objetivos propostos, a avaliação do comportamento dos Componentes Interpretáveis será feita pelos valores assumidos pelos ângulos dos mesmos em relação ao componente principal original. Para tal, foram geradas amostras com *outliers* em termos computacionais que, posteriormente, foram submetidas às técnicas de Componentes Principais e Componentes Interpretáveis.

3.1 Mistura de Distribuições

Para criação das amostras com *outliers*, utilizou-se, então, uma mistura de distribuições, que caracterizará a amostra por elementos predominantes de uma distribuição e elementos de uma segunda distribuição.

O modelo de mistura utilizado foi:

$$f(\mathbf{x}) = (1-\gamma)\, f_1(\mathbf{x}) + \gamma f_2(\mathbf{x}) \tag{41}$$

Em que $(1-\gamma)$ representa a probabilidade do processo ser realizado por $f_1(\mathbf{x})$ e γ a probabilidade do processo ser realizado por $f_2(\mathbf{x})$. Os valores assumidos por γsão 0,05, 0,15 e 0,30.

Dessa forma tem-se $f_1(\mathbf{x})$ como a distribuição de referência, sempre assumida como $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, que é uma normal multivariada, e $f_2(\mathbf{x})$ assumiu:

a) Normal multivariada assimétrica:

$$\mathbf{X} \sim NA_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}\right),\ \boldsymbol{\alpha} = \begin{bmatrix} -20 & -20 & -20 \end{bmatrix}^T \tag{42}$$

b) *t*-Student multivariada:

$$\mathbf{X} \sim t_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu = 5\right) \tag{43}$$

c) log-normal multivariada:

$$\mathbf{X} \sim LN_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{44}$$

Para as distribuições utilizadas na obtenção das amostras foi utilizado um vetor de médias $\boldsymbol{\mu} = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}^T$ e para as matrizes de covariâncias $\boldsymbol{\Sigma}$ de ordem *p* foram consideradas três diferentes estruturas de correlação, chamadas $\mathbf{R}_1, \mathbf{R}_2$ e $\mathbf{R}_3$.

De acordo com Diggle et al. (2002) e Diggle (1988), a matriz de correlação deve apresentar flexibilidade para englobar fontes de variação, em função dos efeitos aleatórios, variação explicada por correlação serial, em que se espera que as observações mais próximas sejam fortemente correlacionadas e, ainda, variação em virtude dos erros de medida. Dessa forma, foram assumidas três diferentes estruturas de correlação, a fim de que se obtivesse maior abrangência dos aspectos citados pelo autor.

As estruturas adotadas têm a seguinte representação (LITTEL; PENDERGAST; NATARAJAN, 2000; CAMARINHA FILHO, 2002):

1) Autoregressiva de ordem 1 AR(1) representada por $\mathbf{R}_1$, possui estrutura com variâncias homogêneas. A correlação entre dois elementos adjacentes é igual a ρ, entre dois elementos separados por um terceiro, igual a ρ^2 e, assim, sucessivamente. Observa-se, então, que a estrutura especifica

correlações diferentes entre variáveis, que decrescem para zero com o aumento do *lag*. Este modelo de estrutura de correlação é bastante utilizado em dados provenientes de medidas repetidas e longitudinais.

$$\mathbf{R}_1 = \begin{bmatrix} 1 & \rho^{2-1} & \cdots & \rho^{p-1} \\ \rho^{2-1} & 1 & \cdots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{bmatrix} \tag{45}$$

2) Simetria Composta de Variância Homogênea (CS), representada por $\mathbf{R}_2$ e que possui estrutura de correlação homogêneas. É assumido que as variáveis tenham a mesma correlação, que é ocorrência comum em alguns estudos experimentais.

$$\mathbf{R}_2 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \tag{46}$$

3) Estrutura Toeplitz, representada por $\mathbf{R}_3$, especifica que a correlação depende de um *lag*, mas não de forma exponencial, como o caso AR(1), podendo ser declarados coeficientes de correlação diferentes para cada variável, dependendo da dimensão da matriz.

$$\mathbf{R}_3 = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_p \\ \rho_1 & 1 & \cdots & \rho_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_p & \rho_{p-1} & \cdots & 1 \end{bmatrix} \tag{(47}$$

Foram assumidos os coeficientes de correlação $\rho = 0,5$ e $\rho = 0,8$ para $\mathbf{R}_1$ e $\mathbf{R}_2$. Para $\mathbf{R}_3$ é necessário assumir conjuntamente tantos coeficientes de correlação quanto for o número p de variáveis. Assim para $\mathbf{R}_3$ utilizou-se um vetor de correlações $\boldsymbol{\rho}_1 = [0,9\ 0,8\ 0,7]^T$, representando altas correlações e um vetor $\boldsymbol{\rho}_2 = [0,6\ 0,5\ 0,4]^T$ para médias correlações, afim de que possam ser comparados com as outras estruturas.

No processo de simulação recorreu-se, ainda, a diferentes tamanhos amostrais (n igual a 50, 100 e 200) e $p = 3$ variáveis e diferentes probabilidades de mistura.

Assim, considerando $\boldsymbol{\mu}$, alternando as estruturas de correlação de $\boldsymbol{\Sigma}$ entre $\mathbf{R}_1$, $\mathbf{R}_2$ e $\mathbf{R}_3$, diferentes misturas de distribuições foram geradas, obedecendo aos valores de γ previamente informados.

3.2 Componentes Principais e Componentes Interpretáveis

Para cada situação descrita na seção 3.1, ou seja, amostras geradas com uma população de referência normal multivariada e contaminadas com observações de outras distribuições foram obtidos $p = 3$ Componentes Principais.

A partir do vetor $\mathbf{X}$ de $p = 3$ variáveis aleatórias, foram encontradas 3 combinações lineares, não correlacionadas, dos elementos de $\mathbf{X}$ que tenham máxima variância. A combinação é dada por $\mathbf{e}_i^T \mathbf{X}$, $i = 1, 2, 3$.

Dessa forma, há três Componentes Principais:

$$\begin{aligned}\mathbf{e}_1^T\mathbf{X} &= e_{11}x_1 + e_{12}x_2 + e_{13}x_3 \\ \mathbf{e}_2^T\mathbf{X} &= e_{21}x_1 + e_{22}x_2 + e_{23}x_3 \\ \mathbf{e}_3^T\mathbf{X} &= e_{31}x_1 + e_{32}x_2 + e_{33}x_3\end{aligned} \tag{48}$$

Partindo das equações apresentadas em 48, procedeu-se, então, à obtenção dos Componentes Interpretáveis $\boldsymbol{\alpha}_i$, $i = 1,..., p$ considerando a restrição de homogeneidade em que $\boldsymbol{\alpha}_i$ assumiu os valores $\pm c$, sendo $\pm c$ proposto como $\pm\frac{1}{\sqrt{k}}$, em que $k = 1, 2,..., p$. Esse processo reduziu ainda mais a quantidade dos coeficientes a serem assumidos pelos Componentes Principais. Dessa forma obtiveram-se os Componentes Interpretáveis:

$$\begin{aligned}\boldsymbol{\alpha}_1^T\mathbf{X} &= \alpha_{11}x_1 + \alpha_{12}x_2 + \alpha_{13}x_3 \\ \boldsymbol{\alpha}_2^T\mathbf{X} &= \alpha_{21}x_1 + \alpha_{22}x_2 + \alpha_{23}x_3 \\ \boldsymbol{\alpha}_3^T\mathbf{X} &= \alpha_{31}x_1 + \alpha_{32}x_2 + \alpha_{33}x_3\end{aligned} \tag{(49)}$$

Sobre todos e_{ij} (coeficientes dos Componentes Principais) representados em (48) foi feita uma correspondência com cada α_{ij} em (49). Dessa forma, para cada e_{ij} houve um α_{ij} correspondente;

O algoritmo de procura executa os seguintes passos:

a) Fixa os elementos de $\boldsymbol{\alpha}_i$ em $\pm\frac{1}{\sqrt{k}}$.

b) Faz a correspondência dos sinais dos coeficientes α_{ij} com os coeficientes de e_{ij}.

c) Considera-se a restrição $\boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_i = 1$.

d) Obtém-se o ângulo entre $\boldsymbol{\alpha}_i$ e $\mathbf{e}_i$, por meio do $\arccos\left(\mathbf{e}_i^T \boldsymbol{\alpha}_i\right)$.

A avaliação dos ângulos foi feita por meio dos valores médios angulares obtidos pelos $a = \arccos\left(\mathbf{e}_i^T \boldsymbol{\alpha}_i\right)$ nas simulações. Como resultado, obtiveram-se valores em radianos que, posteriormente, foram transformados em graus. Quanto mais próximo de zero for o valor médio angular, mais próxima da direção do componente principal é a direção do componente interpretável.

Para obtenção dos Componentes Interpretáveis foram feitas 2000 simulações Monte Carlo, utilizando o software R (R DEVELOPMENT CORE TEAM, 2013), por meio do desenvolvimento de uma rotina computacional, que se encontra no Anexo C. Dessa forma foram obtidos os valores de $\mathbf{e}_i$ e seus $\boldsymbol{\alpha}_i$ correspondentes.

Para a obtenção dos valores médios angulares resultantes da simulação, foram utilizados os conceitos de estatística circular.

3.3 Procedimento para discriminar o efeito de *outliers* nos ângulos formados entre os eixos CP e CI com aprimoramento da distância de Jammalamadaka & Sengupta

A fim de se visualizar os possíveis valores esperados dos ângulos mais afastados dos demais, em razão do caso de presença de *outliers* na amostra gerada, foram simuladas, para efeito de comparação, amostras provenientes de uma Normal Multivariada. Para tal, fixa-se $\gamma = 0$ no processo de simulação, garantindo, assim, ocorrências de amostras originalmente distribuídas por $N_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, isto é, livre de *outliers*, considerando as estruturas de correlação $\mathbf{R}_1$, $\mathbf{R}_2$ e $\mathbf{R}_3$.

Dado o propósito de identificar o efeito de *outliers* na distribuição empírica dos ângulos formados pelos eixos entre os Componentes Principais e Interpretáveis, procedeu-se a uma modificação na distância apresentada por Jammalamadaka e Sengupta (2001), reescrevendo-a por:

$$\theta_{ij} = 180^{\circ} - \left| 180^{\circ} - \left| \hat{a}^{*}_{\gamma k} - \hat{a}_{\gamma k} \right| \right|, \ \theta_{ij} \in \left[0,180^{\circ} \right], \text{ em que} \qquad ((50)$$

$\hat{a}^{*}_{\gamma k}$ representa o valor esperado dos ângulos correspondentes a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com $(\gamma = 0)$, sendo *k* um indexador dos componentes (*k =1,2,3* componentes);

$\hat{a}_{\gamma k}$ representa o valor esperado dos ângulos das variáveis na presença de *outliers* (obtidos de acordo com a mistura de distribuições);

Obtiveram-se, dessa forma, as distâncias em cada situação avaliada, em relação a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

A modificação proposta está no fato de que as distâncias são sempre calculadas em relação a uma amostra de referência, sem presença de *outliers*. A identificação do efeito dos *outliers* nos ângulos foi, então, obtida por essa medida de distância proposta, que identificou dissimilaridades entre ângulos obtidos de uma amostra sem *outliers* e ângulos de outra amostra com *outliers.*

Após obtenção dos valores esperados dos ângulos e das distâncias θ_{ij} procedeu-se à representação gráfica angular e distâncias no círculo trigonométrico, com o objetivo de identificar possíveis padrões. A representação angular foi feita, considerando conjuntamente os dois coeficientes de correlação. Foram utilizados gráficos do tipo Dot-Plot (WILKINSON, 1999), que representaram cada observação obtida em uma escala horizontal e, também, permitiram a visualização de diferenças entre os dois coeficientes de correlação

(médios e altos). Os gráficos foram elaborados para os valores esperados dos ângulos e distâncias, a fim de se verificar, visualmente, pontos discrepantes.

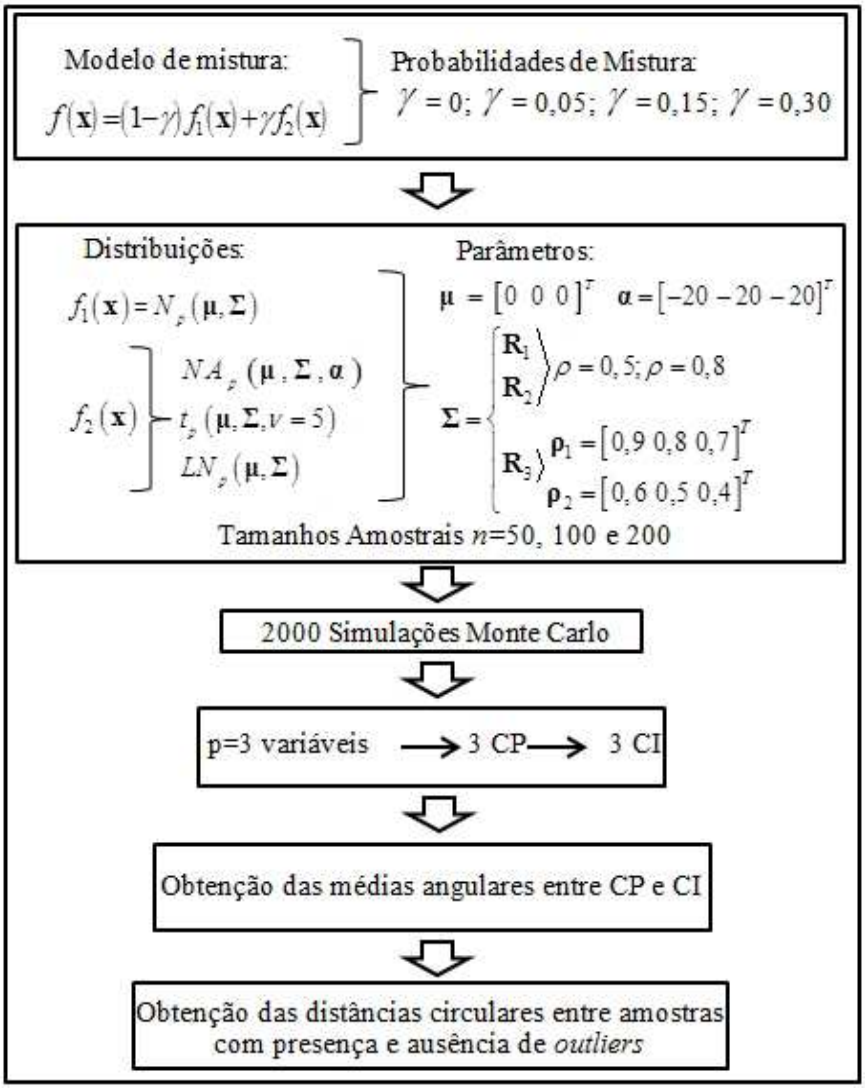


Figura 9 Fluxograma do processo de simulação Monte Carlo para computar as distâncias obtidas em (50) (seção 3.3)

4 RESULTADOS E DISCUSSÃO

4.1 Médias angulares dos componentes

Tendo por base os cenários avaliados na simulação Monte Carlo, representados pelas combinações, envolvendo diferentes estruturas e graus de correlação, proporção de outliers (γ), em relação ao tamanho amostral (n), os resultados descritos nas Tabelas 2-5 correspondem aos valores esperados angulares entre os eixos formados pelos Componentes Principais e os eixos gerados pelos Componentes Interpretáveis, obtidos por meio da distribuição empírica resultante das realizações Monte Carlo.

Neste contexto, os resultados e discussão são descritos, de modo que os valores encontrados na Tabela 2 correspondem às médias angulares dos componentes estimados em amostras sem a presença de *outliers*.

De forma análoga, os resultados nas Tabelas 3-5 são descritos, considerando as amostras contaminadas com diferentes proporções de *outliers*.

Tabela 2 Média dos ângulos em graus considerando a distribuição Normal Multivariada

Estrutura de Correlação	n	$\rho = 0{,}80$			$\rho = 0{,}50$		
		$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
AR(1)	50	1,95	30,59	19,74	5,91	31,00	18,25
	100	1,97	31,38	19,02	4,93	31,04	17,97
	200	1,96	33,07	17,91	4,36	32,85	16,32
CS	50	0,62	25,19	23,80	2,55	27,55	23,10
	100	0,43	25,35	24,32	1,87	26,47	24,51

	200	0,32	25,22	23,76	1,38	25,85	25,09

"Tabela 2, conclusão"

Estrutura de Correlação	n	$\rho = 0,80$			$\rho = 0,50$		
		$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
		$\boldsymbol{\rho}_1 = [0,9\ 0,8\ 0,7]$			$\boldsymbol{\rho}_2 = [0,6\ 0,5\ 0,4]$		
Toeplitz	50	1,23	30,58	19,87	2,01	30,56	19,48
	100	1,21	32,06	19,07	2,04	32,36	18,22
	200	1,15	33,02	18,78	1,95	33,12	17,94

Os resultados encontrados na Tabela 2 evidenciaram que os valores esperados dos ângulos entre os eixos dos Componentes Principais e Interpretáveis são menos influenciados pelo efeito do tamanho amostral. Entretanto, notou-se um maior impacto, ao considerar a estrutura de correlação e os coeficientes de correlação, uma vez que os resultados obtidos para a correlação de simetria composta (CS) foram mais contrastantes em relação às demais estruturas.

Em se tratando de análise de Componentes Principais, de acordo com Morrison (1990), a estrutura CS garante a explicação da maior parte da variação em um único componente principal em situações de alta correlação entre as variáveis, possuindo uma dimensão que tem uma orientação com ângulos iguais entre os eixos das variáveis originais, garantindo coeficientes muito próximos para os Componentes.

Uma vez obtidos os Componentes Principais no maior coeficiente de correlação ($\rho = 0,80$), o primeiro Componente Principal explicou a maior parte da variância, garantindo coeficientes não muito dispersos pelo uso da estrutura citada.

Contextualizando com os resultados observados quanto à estrutura CS, no estudo dos Componentes Interpretáveis, esta estrutura foi a que apresentou os

menores ângulos entre os primeiros componentes conforme observado na Tabela 2. Como os coeficientes dos primeiros Componentes Principais não são dispersos, os coeficientes dos primeiros Componentes Interpretáveis estão bem próximos, garantindo os menores ângulos.

Por outro lado, a estrutura AR(1) apresenta correlações diferentes entre as variáveis em virtude do *lag* do coeficiente de correlação. Assim explicação da variação pelo componente diminui, elevando o valor do ângulo formado entre os primeiros componentes, quando comparados os dois coeficientes de correlação.

Frente ao exposto, sugere-se uma nova investigação do efeito da estrutura com ênfase nas amostras contaminadas, cujos resultados são discutidos a seguir.

Tabela 3 Média dos ângulos em graus considerando a distribuição Normal Assimétrica com $\gamma = 0,05$ e $\gamma = 0,30$

Estrutura de Correlação	γ	n	$\rho = 0,80$			$\rho = 0,50$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
AR(1)	0,05	50	1,91	30,38	19,79	5,98	31,03	19,05
		100	1,95	31,84	18,55	5,94	32,06	16,91
		200	2,02	32,71	18,01	5,15	32,84	15,82
AR(1)	0,30	50	2,07	30,19	19,77	6,37	30,93	19,12
		100	2,29	31,96	18,34	5,79	31,77	17,31
		200	2,07	32,41	18,17	5,42	32,03	16,81
CS	0,05	50	0,72	25,28	25,46	3,08	27,27	24,16
		100	0,46	25,69	25,00	1,56	26,67	24,02
		200	0,29	25,12	25,53	1,22	25,76	25,20
CS	0,30	50	0,65	25,71	25,05	2,85	27,41	23,66
		100	0,52	25,90	24,84	1,85	25,95	24,94
		200	0,34	25,03	25,67	1,22	26,67	23,87

"Tabela 3, conclusão"

Estrutura de Correlação	γ	n	$\rho = 0{,}80$			$\rho = 0{,}50$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
			$\boldsymbol{\rho}_1 = [0{,}9\ 0{,}8\ 0{,}7]$			$\boldsymbol{\rho}_2 = [0{,}6\ 0{,}5\ 0{,}4]$		
		50	1,21	30,78	19,81	1,92	31,42	18,86
Toeplitz	0,05	100	1,20	32,19	18,99	1,95	32,58	18,16
		200	1,18	33,19	18,68	1,68	33,06	18,16
		50	1,42	31,06	19,50	2,20	30,52	19,22
Toeplitz	0,30	100	1,22	32,96	18,65	1,88	32,93	18,06
		200	1,17	32,50	18,83	1,81	33,34	17,95

Em concordância com os resultados obtidos em amostras não contaminadas (Tabela 2), os resultados descritos na Tabela 3 evidenciaram que, indiferente do grau de contaminação, a estrutura e o grau de correlação entre as variáveis, de fato, apresentam um efeito mais perturbador nos valores esperados dos ângulos formados entre os eixos dos componentes.

De forma mais específica, notou-se que ao assumir a estrutura de correlação AR(1), o ângulo $\hat{a}_1$, formado entre os eixos representados pelo Componente Interpretável e o primeiro Componente Principal, assumiu menor valor quando as variáveis foram altamente correlacionadas ($\rho = 0{,}80$). Ao comparar os ângulos entre os eixos formados pelos segundos Componentes Principais e Interpretáveis, com suas respectivas parametrizações e nos dois coeficientes de correlação, notou-se uma variação mínima entre as médias angulares, explicada meramente pelo erro Monte Carlo.

Em se tratando da estrutura de Simetria Composta (CS), as menores médias angulares foram identificadas nos primeiros componentes, com valores

$\hat{a}_1$ menores em relação aos ângulos obtidos ao se considerar as estruturas AR(1) e Toeplitz.

A ocorrência de menores ângulos nos eixos formados entre os primeiros Componentes Principais e Interpretáveis estão de acordo com resultados apresentados em Chipman e Gu (2005) e Vines (2000), que obtiveram a mesma relação para os primeiros componentes, porém não consideraram *outliers* ou diferentes estruturas de correlação em seus estudos.

Ainda sobre a estrutura CS, a média angular manteve-se inferior para os segundos componentes ($\hat{a}_2$). Para o terceiro componente, apresentou elevação nas médias em relação à estrutura AR(1) e Toeplitz em situações de $\rho = 0{,}80$. Na estrutura CS, praticamente em todas as situações, o primeiro ângulo $\hat{a}_1$ apresentou médias inferiores a 1º para $\rho = 0{,}80$ e o segundo e terceiros ângulos, médias em torno de 25º.

Ao assumir a estrutura de correlação Toeplitz, observou-se que os valores angulares esperados para o primeiro Componente Interpretável ($\hat{a}_1$), foram inferiores aos valores esperados nas situações em que a estrutura AR(1) foi considerada. Porém, ressalta-se que dado diferentes graus de correlação um aumento nos valores esperados foi detectado, no entanto com menor variação, quando comparado às demais estruturas.

O terceiro componente apresentou valores esperados menores, próximos a 19º, nas estruturas AR(1) e Toeplitz em ambos os graus de correlação.

Em relação ao efeito do tamanho amostral, os resultados foram concordantes com os apresentados na Tabela 2, ou seja, os valores esperados dos ângulos são pouco influenciados pela variação do tamanho da amostra.

Quanto à variação de γ, verifica-se pelas Tabelas 3, 4 e 5 a ocorrência de mínimas variações nos valores médios dos ângulos, principalmente, nos ângulos $\hat{a}_1$ da estrutura AR(1). As variações são pequenas, não excedendo 1º.

As maiores variações continuam acontecendo na mudança de $\rho = 0,80$ para $\rho = 0,50$, no caso AR(1) e CS.

Os casos em que $\gamma = 0,15$ apresentam resultados similares e encontram-se no Anexo A (Tabelas 1-3).

Mantendo as mesmas situações paramétricas avaliadas na simulação Monte Carlo, os resultados encontrados na Tabela 4 referem-se à distribuição log-normal multivariada, da qual as amostras foram geradas.

Notou-se que o efeito do excesso de curtose, bem como alto grau de assimetria, caracterizado por esta distribuição, não apresentou diferenças expressivas, ou seja, as diferenças nas médias angulares foram perceptíveis na mudança da estrutura de correlação e nos casos dos valores assumidos de ρ.

A estrutura CS manteve-se com as menores médias angulares para os dois primeiros componentes, ao passo que a estrutura AR(1) e Toeplitz apresentou menores médias angulares para os terceiros componentes.

Em relação aos Componentes Interpretáveis, de um modo geral estes apresentaram médias angulares mínimas em relação aos primeiros Componentes Principais, em cada estrutura de correlação e os menores ângulos ocorrem no caso de maior coeficiente de correlação.

Ao se comparar as Tabelas 3-5 com a Tabela 2, verificou-se, ainda, que a distribuição dos ângulos não é afetada quando a amostra é perturbada pela contaminação.

Com as mesmas evidências estatísticas, ao simular amostras com a t-Student multivariada (Tabela 5), caracterizada por uma simetria e baixo nível de curtose, os resultados foram semelhantes aos casos anteriores.

Tabela 4 Média dos ângulos em graus considerando a distribuição log-Normal com $\gamma = 0,05$ e $\gamma = 0,30$

Estrutura de Correlação	γ	n	$\rho = 0,80$			$\rho = 0,50$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
		50	1,98	29,24	21,08	5,21	29,54	21,09
AR(1)	0,05	100	1,84	30,5	20,26	4,80	31,19	18,46
		200	1,72	32,11	18,58	4,53	32,13	16,64
		50	1,96	27,72	22,47	3,47	27,64	23,37
AR(1)	0,30	100	1,96	30,21	19,95	4,18	29,80	19,87
		200	1,57	29,66	20,66	4,12	30,24	20,46
		50	0,64	25,76	24,98	2,56	26,81	24,30
CS	0,05	100	0,52	24,81	26,19	1,72	26,21	24,80
		200	0,47	25,68	24,75	1,32	27,05	23,32
		50	0,85	25,60	25,22	2,68	26,37	25,69
CS	0,30	100	0,76	26,28	24,18	2,01	25,83	25,58
		200	0,87	26,17	24,55	1,33	26,11	24,66
			$\boldsymbol{\rho}_1 = [0,9\ 0,8\ 0,7]$			$\boldsymbol{\rho}_2 = [0,6\ 0,5\ 0,4]$		
		50	1,23	30,63	19,97	1,82	31,23	19,03
Toeplitz	0,05	100	1,11	31,69	19,31	1,68	31,94	18,77
		200	1,12	31,80	19,81	1,55	32,84	18,34
		50	1,26	29,35	20,96	1,34	30,13	20,45
Toeplitz	0,30	100	0,98	29,24	21,48	1,16	29,47	21,23
		200	1,34	28,49	21,50	1,19	32,09	19,14

Tabela 5 Média dos ângulos em graus considerando a distribuição t-Student com $\gamma = 0,05$ e $\gamma = 0,30$

Estrutura de Correlação	γ	n	$\rho = 0,80$			$\rho = 0,50$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
		50	1,86	30,81	19,55	5,95	30,60	19,83
AR(1)	0,05	100	2,01	31,63	18,67	5,05	31,97	16,79
		200	1,83	32,97	18,06	4,48	31,77	17,57
		50	2,50	29,73	20,28	6,35	30,36	20,64
AR(1)	0,30	100	2,03	31,17	19,00	5,49	31,31	18,04
		200	1,89	31,84	18,59	4,87	32,01	17,26
		50	0,68	25,17	25,57	3,31	28,04	23,20
CS	0,05	100	0,36	25,13	25,77	1,79	27,08	23,73
		200	0,30	25,69	25,08	1,16	25,32	25,90
		50	0,81	25,54	25,23	2,47	27,16	23,69
CS	0,30	100	0,57	26,21	24,45	2,12	26,72	24,14
		200	0,36	26,08	24,82	1,69	24,91	26,80
			$\boldsymbol{\rho}_1 = [0,9\ 0,8\ 0,7]$			$\boldsymbol{\rho}_2 = [0,6\ 0,5\ 0,4]$		
		50	1,16	30,85	20,07	2,23	31,42	18,63
Toeplitz	0,05	100	1,18	31,47	19,31	2,05	32,57	18,03
		200	1,10	33,22	18,66	1,95	31,83	18,58
		50	1,28	30,13	20,27	2,07	30,77	19,24
Toeplitz	0,30	100	1,21	31,89	19,20	2,21	32,57	17,92
		200	1,14	32,66	18,85	2,16	32,94	17,88

No que diz respeito a distribuições assimétricas aplicadas às medidas angulares, Fisher e Hall (1989) ressaltaram a importância do estudo de regiões de confiança nessas situações. Segundo os autores, quando um modelo paramétrico como Von-Mises não é apropriado, recorre-se à teoria assintótica, para elaboração de tais regiões. Ainda, para amostras menores de medidas angulares, que não possuem simetria rotacional, foram propostas pelos autores regiões de confiança *bootstrap* baseadas em uma quantidade pivotal que é função do ângulo entre a direção média verdadeira e a direção média amostral. No presente estudo, porém, não são levadas em consideração inferências sobre os ângulos, pois não foi utilizada a distribuição Von-Mises.

Dado o propósito de identificar o efeito de *outliers*, nas medidas de distância entre os ângulos, os resultados doravante apresentados na seção 4.2 são baseados na obtenção da distância entre as médias angulares.

4.2 Obtenção e representação das distâncias entre os ângulos

A maior parte das estatísticas de teste para identificação de *outliers* em medidas angulares utiliza o pressuposto de amostras e distâncias, identicamente, distribuídas por uma distribuição Von-Mises, para que sejam feitas as inferências. Partindo desta premissa, as Figuras 10 – 15 apresentam os gráficos P-P Plot referente às distâncias obtidas na estrutura CS, para o primeiro Componente Principal. Os demais casos encontram-se representados no Anexo B.

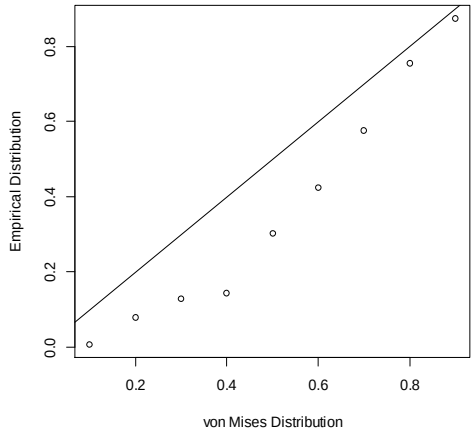


Figura 10 P-P Plot da distribuição Von-mises estrutura CS, n=50, ρ =0,5 e CP 1

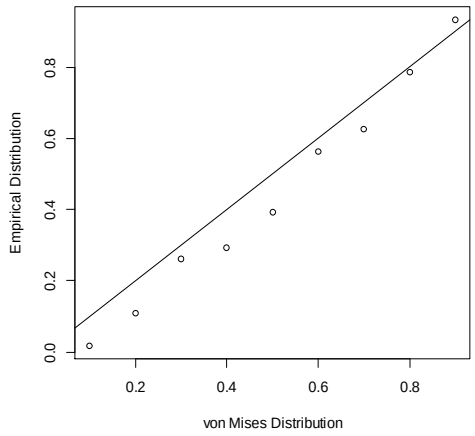


Figura 11 P-P Plot da distribuição Von-mises estrutura CS, n=100, ρ =0,5 e CP 1

As Figuras 10 e 11 representaram o ajuste da distância pela Distribuição Von-Mises, por meio de um gráfico Probabilidade-Probabilidade (P-P Plot). Uma vez que os pontos não se distribuem, uniformemente, sobre a reta obtida por uma distribuição empírica e a distribuição Von-Mises, as distâncias não são bem representadas por tal distribuição.

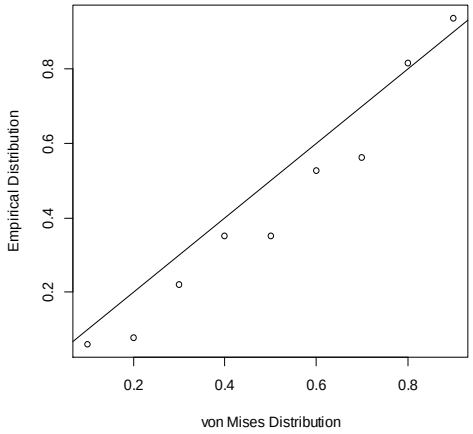


Figura 12 P-P Plot da distribuição Von-mises estrutura CS, n=200, ρ =0,5 e CP 1

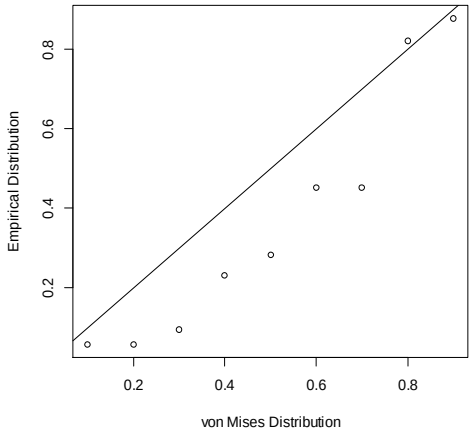


Figura 13 P-P Plot da distribuição Von-mises estrutura CS, n=50, ρ =0,8 e CP 1

Para um tamanho amostral n igual a 200 e ρ =0,5 as distâncias obtidas entre os primeiros Componentes não se ajustaram à distribuição Von-Mises (Figura 12). O mesmo ocorre para as condições de tamanho amostral n igual a 50 e ρ =0,8 (Figura 13). O mesmo pode ser observado nas Figuras 14 e 15.

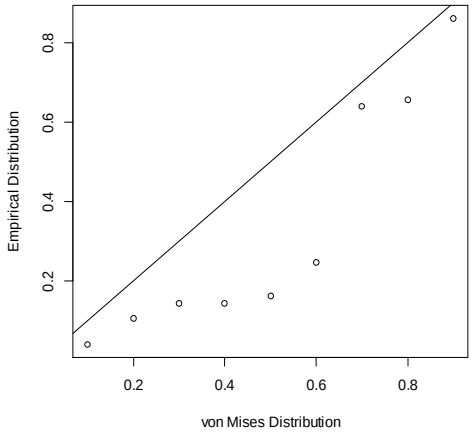


Figura 14 P-P Plot da distribuição Von-mises estrutura CS, n=100, ρ =0,8 e CP 1

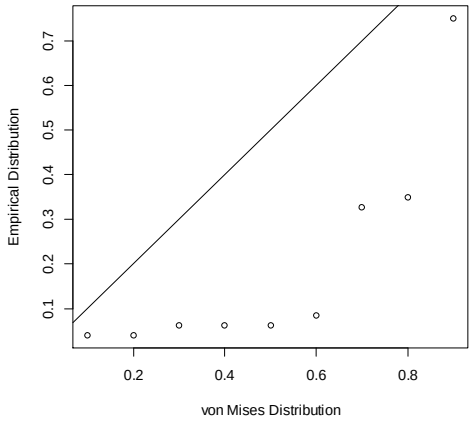


Figura 15 P-P Plot da distribuição Von-mises estrutura CS, n=200, ρ =0,8 e CP 1

Lewis, Fisher e Willcox (1981) afirmam que um exame preliminar de um dado conjunto de medidas angulares, por meio dos P-P Plots, é o passo inicial na identificação de *outliers.* Os resultados representados pelos P-P Plots,

estão de acordo com resultados apresentados por Jammalamadaka e Sengupta (2001) e Lewis e Fisher (1981) que utilizam a mesma ferramenta para um diagnóstico inicial de dados circulares. Os autores mostraram que a ocorrência de pontos fora da linha de referência indica possível presença de pontos discrepantes.

A partir dos resultados apresentados pelas Figuras 10-15 e conforme resultados descritos na seção anterior tornou-se necessário identificar técnicas específicas para estudo das distâncias entre as médias angulares. Para tal foi utilizado como medida de dissimilaridade:

$$\theta_{ij} = 180^{\circ} - \left| 180^{\circ} - \left| \hat{a}^{*}_{\gamma k} - \hat{a}_{\gamma k} \right| \right|, \ \theta_{ij} \in \left[0, 180^{\circ} \right] \qquad ((51)$$

Os resultados das distâncias encontram-se no Anexo A (Tabelas 4-6).

As Figuras 16-24 representam os ângulos referentes aos componentes e as distâncias entre amostras com *outliers* e amostras geradas de uma Normal Multivariada. Estão representados, também, gráficos do tipo Dot-Plot, a fim de que se visualize a dispersão das distâncias obtidas pela expressão 51.

Em relação às médias angulares dos componentes, considerando a estrutura AR(1) e, conjuntamente os dois coeficientes de correlação, $\rho = 0{,}80$ e $\rho = 0{,}50$, a Figura 16 representa a disposição dos valores médios angulares no círculo trigonométrico. Observa-se a presença de três agrupamentos distintos.

Os grupos são formados pelas médias angulares observadas para os três Componentes Principais em relação aos respectivos Componentes Interpretáveis. A Tabela 3 confirma os grupos representados, visto que apresenta para $\hat{a}_1$, uma variação entre 1,5º a, aproximadamente 6,0º, para $\hat{a}_2$ variação das médias angulares em torno de 30,0º e para $\hat{a}_3$ em torno de 19,0º, comprovando a existência de pontos dispersos em três grupos, como mostrado na Figura 16.

Pelos resultados dos grupos, observou-se uma concordância com resultados de uma nova metodologia proposta por Enki et al. (2013), que propuseram o estudo dos Componentes Interpretáveis sobre o agrupamento de variáveis. Primeiramente são identificadas as variáveis mais similares para construção dos agrupamentos. Posteriormente cada agrupamento é tratado como uma variável sendo submetido à análise de Componentes Principais e Interpretáveis, porém sob outras restrições, também, garantindo a interpretabilidade.

A distribuição das distâncias $\hat{\theta}_{ij}$ para a estrutura de correlação AR(1) encontram-se na Figura 17. Embora seja perceptível a existência de valores distintos, o diagrama circular apresenta uma difícil identificação de tal situação, por apresentar valores com baixa amplitude. Para tal identificação, utilizou-se, então, o Gráfico do tipo Dot-Plot, apresentado na Figura 18.

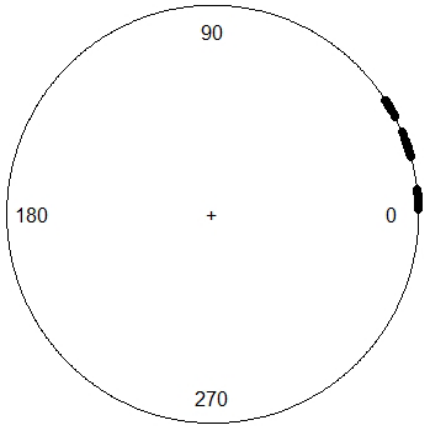


Figura 16 Representação angular dos componentes na estrutura AR(1) para $\rho = 0{,}80$ e $\rho = 0{,}50$

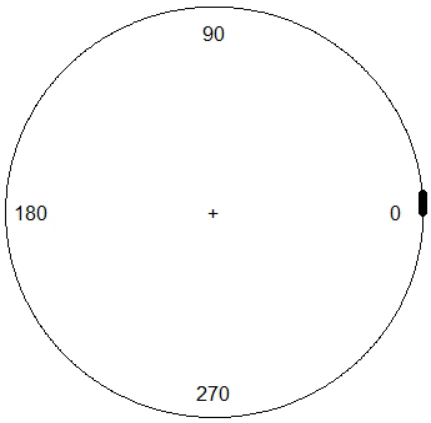


Figura 17 Representação angular das distâncias na estrutura AR(1)

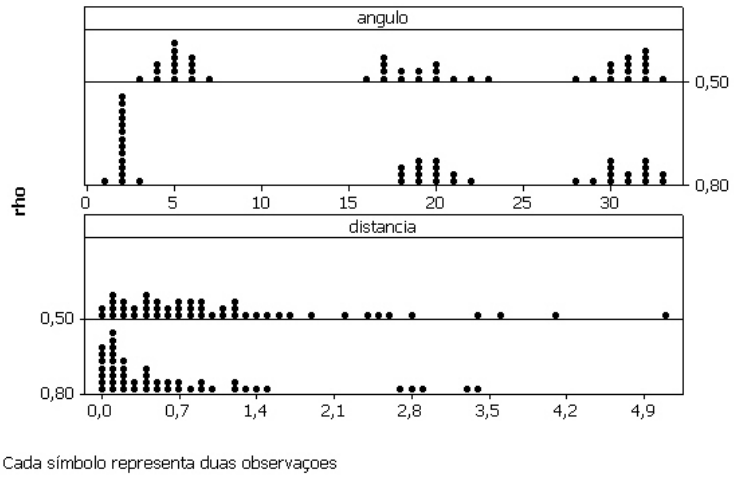


Figura 18 Dot-Plot para o ângulo e distância considerando a estrutura AR(1)

O Dot-Plot, para as médias angulares e distâncias para a estrutura AR(1) (Figura 18), apresenta uma melhor visualização dos pontos, separadamente, para cada ρ. A parte superior, que representa os ângulos, também, mostra a mesma diferenciação dos grupos das médias angulares, considerando os três

Componentes Principais, onde claramente percebem-se três concentrações distintas de pontos. Já para as distâncias, estas estão representadas na parte inferior do gráfico que mostra alguns valores dissimilares dos demais.

Abuzaid, Hussin e Mohamed (2008) utilizaram a mesma representação angular das Figuras 16 e 17, como forma de identificação de possíveis *outliers*, em um modelo de regressão circular, baseado na distância circular, em que propuseram uma nova definição de resíduo circular baseado nesse tipo de distância. Também foram utilizados gráficos do tipo P-P Plot para análise inicial.

Uma forma gráfica mais apropriada, para a detecção de *outliers*, foi proposta por Abuzaid et al. (2012) que é o chamado boxplot circular. Os autores utilizaram a técnica em um exemplo prático apresentado por Collet (1980), de medidas angulares referentes à direção tomada por sapos, após serem libertados de um confinamento, identificando, assim, de forma, também, gráfica os *outliers*.

Estes trabalhos recentes estão em consonância com os resultados aqui encontrados, confirmando que, no caso de medidas angulares, a representação gráfica é uma eficiente ferramenta de detecção.

Considerando o Anexo A (Tabela 4), é possível verificar que as distâncias mais dissimilares ocorrem nas situações de contaminação pela distribuição log-normal e seu valor mais extremo é 5,12º em ρ =0,5, γ= 0,30 correspondente ao terceiro componente principal. Logo, este é um possível ponto discrepante.

Collet (1980) utilizou como teste de possível dissimilaridade, o desvio de uma observação angular em relação à direção média amostral. A partir dos possíveis pontos discrepantes, estudou a eficiência de alguns testes na detecção de *outliers*. Foram avaliados 3 testes (C, D e M) dos quais 2 (D e M) identificaram o possível ponto candidato como discrepante.

Jammalamadaka e Sengupta (2001) utilizaram a definição alternativa de distância circular apresentada em 23, como identificação inicial de *outliers* em amostras, identicamente, distribuídas por uma distribuição Von-Mises, onde detectaram os possíveis pontos discrepantes. A distância diferencia-se da proposta atual pela identificação do efeito dos *outliers* nos ângulos provenientes de componentes de amostras contaminadas e não identicamente distribuídas.

A distância circular foi utilizada, também, com eficácia por Abuzaid, Hussin e Mohamed (1999) e Abuzaid et al. (2012), para a criação de uma estatística para identificação de pontos discrepantes, apresentada na equação 22. À medida que a distância entre duas observações se torna maior que as demais, influenciam diretamente no valor da estatística.

Os autores identificaram *outliers*, utilizando a estatística baseada na distância circular, no exemplo prático apresentado por Collet (1980). O teste, envolvendo a distância circular, apresenta como *outlier*, o mesmo ponto estudado por Collet (1980). Assim, pode-se se dizer que a distância circular é, realmente, uma boa ferramenta de detecção inicial, corroborando com os resultados encontrados neste trabalho.

No que se refere à estrutura de correlação de Simetria Composta (CS), encontram-se representados na Figura 19, os valores médios angulares para tal estrutura. Verifica-se a presença de dois grupos distintos de pontos que, também, correspondem à dispersão no círculo, dos ângulos $\hat{a}_1$, $\hat{a}_2$ e $\hat{a}_3$. A estrutura CS apresentou valores médios angulares menores que 1º para o primeiro componente e em torno de 25º para o segundo e terceiro componentes, o que caracteriza a visualização de somente dois grupos distintos no diagrama circular.

Quanto à representação das distâncias $\hat{\theta}_{ij}$ por apresentarem a maioria dos valores muito próximos, esta, também, não é bem visualizada no diagrama (Figura 20). Para tal visualização, utilizou-se, então, o Dot-Plot apresentado na Figura 21, em que a parte superior representa bem a distribuição dos ângulos em

dois grupos distintos nos dois casos de coeficiente de correlação, $\rho = 0{,}80$ e $\rho = 0{,}50$. Para as distâncias, representadas na parte inferior, identificam-se alguns pontos afastados dos demais. O mais extremo é o valor de $\hat{\theta}_{ij}$ igual a 2,59, em ρ =0,5, γ= 0,30, distribuição log-normal para o terceiro componente (ANEXO A, TABELA 5).

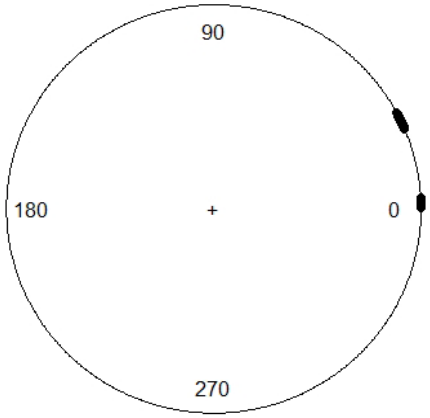


Figura 19 Representação angular dos componentes na estrutura CS para $\rho = 0{,}80$ e $\rho = 0{,}50$

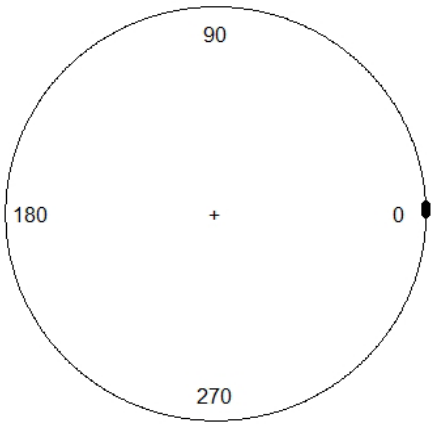


Figura 20 Representação angular das distâncias na estrutura CS

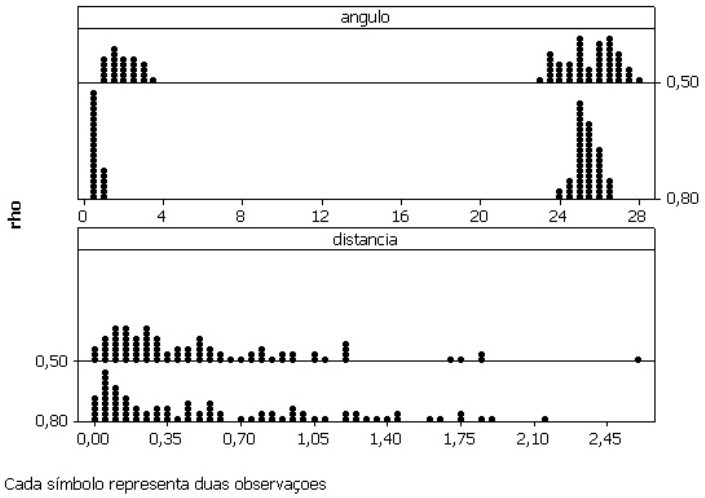


Figura 21 Dot-Plot para o ângulo e distância considerando a estrutura CS

Em relação à estrutura de correlação Toeplitz, os ângulos e distâncias encontram-se representados nas Figuras 22 – 24. Na Figura 22, visualizam-se 3 agrupamentos de valores médios angulares, como na estrutura AR(1). Os valores de $\hat{a}_1$, $\hat{a}_2$ e $\hat{a}_3$ estão em torno de 1º, 20º e 30º, respectivamente, caracterizando os 3 agrupamentos. Em relação à distribuição das distâncias $\hat{\theta}_{ij}$, estas encontram-se representadas na Figura 23 e no Dot-plot da Figura 24, para uma melhor visualização.

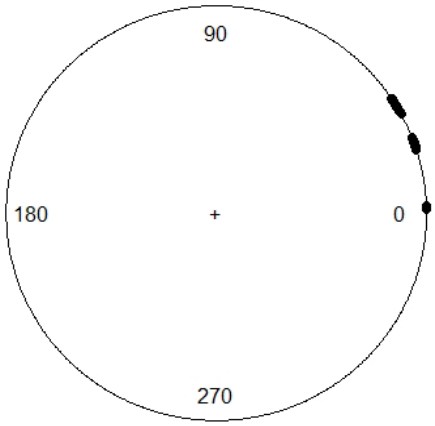


Figura 22 Representação angular dos componentes na estrutura Toeplitz para $\rho = 0{,}80$ e $\rho = 0{,}50$

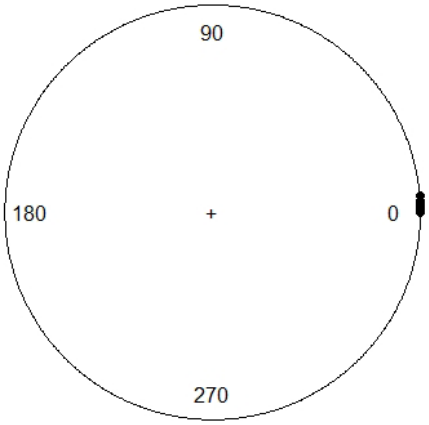


Figura 23 Representação angular das distâncias na estrutura Toeplitz

O valor mais extremo de $\hat{\theta}_{ij}$ é 4,53º na situação de ρ =0,8, γ= 0,30, distribuição log-normal para o segundo componente (ANEXO A, TABELA6).

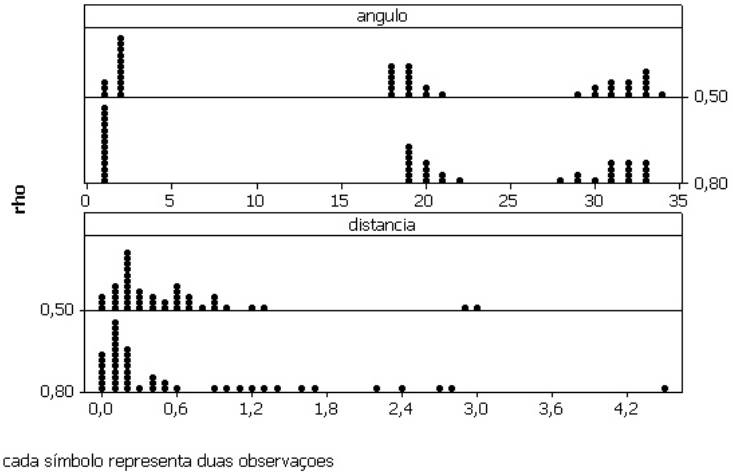


Figura 24 Dot-Plot para o ângulo e distância considerando a estrutura Toeplitz

A identificação de possíveis pontos discrepantes, também, é de grande importância em estudos de regressão envolvendo medidas angulares. Segundo Abuzaid, Hussin e Mohamed (2008), a ocorrência de *outliers* altera as estimativas dos parâmetros em um modelo de regressão linear para variáveis circulares. A identificação, em casos de modelos de regressão e estudo dos efeitos nas estimativas dos parâmetros, foi apresentada por Hussin, Fielle e Stillman (2004) e Hussin et al. (2010) no ajuste de um conjunto de dados envolvendo direção do vento. Como feito no presente trabalho, os autores verificaram, graficamente, a possível existência de pontos discrepantes, diferindo no fato de serem analisados pela estatística COVRATIO e verificados sua influência nas estimativas dos parâmetros do modelo de regressão, considerando a distribuição Von-Mises.

Em se tratando de modelos de regressão circular, a estatística COVRATIO, também, foi utilizada por Ibrahim et al. (2013), especificamente, no chamado modelo de regressão circular JS (JAMMALAMADAKA; SARMA,

1993). Foi utilizado o modelo no mesmo conjunto de direção dos ventos e foi identificada somente uma observação como *outlier*, ao passo que os trabalhos anteriores encontraram duas observações, utilizando modelos de regressão linear. Os trabalhos citados estão de acordo com o objetivo do presente trabalho, no que tange à identificação de possíveis *outliers*.

Um estudo da estatística robusta, aplicada a medidas angulares, foi apresentada por Agostinelli (2007), que apresentou uma metodologia de estimação dos parâmetros da distribuição Von-Mises. Utilizando como exemplo os dados apresentados em Collet (1980), procedeu-se a uma estimação robusta dos parâmetros, considerando o ponto discrepante, previamente, identificado.

4.3 Roteiro para aplicação das medidas angulares na identificação de ângulos discrepantes na seleção de componentes.

Para a utilização da metodologia elaborada no presente trabalho em dados reais, é apresentado o roteiro abaixo para obtenção de possíveis pontos *outliers* utilizando a distância circular.

a) Considerar uma amostra multivariada representada por um vetor $\mathbf{X} = \left[\mathbf{X}_{i1}, \mathbf{X}_{i2} ..., \mathbf{X}_{ip} \right]^T$, em que $(i = 1,...,n)$ e p representa o número total de variáveis. Matricialmente:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

b) Estimar a matriz de correlação amostral $\mathbf{R}$ referente às variáveis a serem analisadas.

c) Obter todos os *p* Componentes Principais ($\mathbf{e}_i$) e Componentes Interpretáveis ($\boldsymbol{\alpha}_i$) relacionados e computar os ângulos entre os mesmos, por meio do $\arccos\left(\mathbf{e}_i^T \boldsymbol{\alpha}_i\right)$.

d) Gerar, via simulação Monte Carlo, considerando a mesma dimensão do conjunto de dados, *n* amostras $N_p\left(\mathbf{0}, \mathbf{R}\right)$, ou seja, de uma distribuição Normal Multivariada com vetor de médias $\mathbf{0}$ e matriz de correlação $\mathbf{R}$ estimada no passo *b*.

e) Computar valores esperados dos ângulos obtidos entre os componentes das amostras simuladas no passo *d*.

f) Computar as distâncias entre os ângulos do passo *c* e os ângulos do passo *e*, por meio da expressão apresentada na seção 3.3:

$$\theta_{ij} = 180^\circ - \left|180^\circ - \left\|\hat{a}^*_{\gamma k} - \hat{a}_{\gamma k}\right\|\right|,\ \theta_{ij} \in \left[0, 180^\circ\right]$$

Considerando $\hat{a}^*_{\gamma k}$ como os ângulos obtidos no passo 5 e $\hat{a}_{\gamma k}$ os ângulos obtidos no passo 3.

g) Construir os gráficos descritivos para representar as possíveis distâncias dissimilares relacionadas aos ângulos obtidos.

5 CONCLUSÕES

A utilização da estatística circular é fundamental para a correta interpretação de resultados e inferências sobre os dados, pois como exposto, ocorrem diferenças nas formas de obtenção de algumas medidas estatísticas em dados circulares quando comparados a situações mais gerais.

O efeito de *outliers* não apresentou diferenças expressivas na distribuição dos ângulos entre o eixo dos Componentes Principais e o eixo dos Componentes Interpretáveis. Entretanto, ao utilizar a distância proposta neste trabalho tornou-se possível identificar quais ângulos foram discrepantes dentro dos cenários de simulação avaliados.

Em se tratando da distribuição dos *outliers*, observou-se que o excesso de curtose, causado pela contaminação pela distribuição log-normal, resultou nas maiores distâncias entre os ângulos. Em relação à assimetria e curtose, a curtose é fonte causadora de distâncias mais dissimilares.

REFERÊNCIAS

ABUZAID, A. H. et al. Statistics for a new test of discordance in circular data. **Communications in Statistics - Simulation and Computation**, New York, v. 41, n. 10, p. 1882-1890, 2012.

ABUZAID, A. H.; HUSSIN, A. G.; MOHAMED, I. B. Identifying single outlier in linear circular regression model based on circular distance. **Journal of Applied Probability & Statistics**, New York, v.3, n.1, p.107-117, 2008.

ABUZAID, A. H.; MOHAMED, I. B.; HUSSIN, A. G. Boxplot for circular variables. **Computational Statistics**, Heidelberg, v. 27, n. 3, p. 381-392, 2012.

ABUZAID, A. H.; MOHAMED, I. B.; HUSSIN, A. G. A new test of discordancy in circular data. **Communications in Statistics - Simulation and Computation**, New York, v. 38, n. 4, p. 682-691, 2009.

AGOSTINELLI, C. Robust estimation for circular data. **Computational Statistics & Data Analysis**, Amsterdam, v. 51, n. 12, p. 5867-5875, 2007.

AZZALINI, A. A class of distributions which includes the normal ones. **Scandinavian Journal of Statistics**, Stockholm, v. 12, n. 2, p. 171-178, 1985.

AZZALINI, A. Further results on a class of distributions which includes the normal ones. **Statistica**, Bologna, v. 46, n. 3, p. 199-208, 1986.

AZZALINI, A.; CAPITANIO, A. Statistical applications of the multivariate skew-normal distribution. **Journal of the Royal Statistical Society. Series B – Statistical Methodology**, London, v. 61, n. 3, p. 579-602, 1999.

AZZALINI, A.; VALLE, A.D. The multivariate skew-normal distributions. **Biometrika**, London, v. 83, n. 2, p. 715-726, 1996.

BARRIGA, G. D. C. **Inferência sobre medidas de posição e dispersão em dados circulares**. 1997. 124 p. Dissertação (Mestrado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 1997.

BATSCHELET, E. **Circular statistics in biology**. New York: Academic, 1981.

BELSLEY, D.A.; EDWIN, K.; ROY, E. W. **Regression diagnostics**: identifying influential data and sources of Collinearity. New York: J. Wiley, 1980. 292 p.

CAMARINHA FILHO, J. A. **Modelos lineares mistos**: estruturas de matrizes de variâncias e covariâncias e seleção de modelos. 2002. 85 p. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2002.

CHIPMAN, H. A.; GU, H. Interpretable dimension reduction. **Journal of Applied Statistics**, Abingdon, v. 32, n. 9, p. 969-987, 2005.

COLLET, D. Outliers in circular data. **Journal of Applied Statistics**, Abingdon, v. 29, n. 1, p. 50-57, 1980.

DIGGLE, P. J. An approach to the analysis of repeated measurements. **Biometrics**, Washington, v. 44, n. 4, p. 959-971, 1988.

DIGGLE, P. J. et al. **Analysis of longitudinal data**. 2. ed. Oxford: Oxford University, 2002.

ENKI, D.G.; TRENDAFILOV, N. T.; JOLLIFFE, T. A clustering approach to interpretable principal components. **Journal of Applied Statistics**, Abingdon, v. 40, n. 3, p. 583-599, 2013.

FISHER, N. I. **Statistical analysis of circular data**. Cambridge: University, 1993. 296 p.

FISHER, N. I.; HALL, P. Bootstrap confidence regions for directional data. **Journal of the American Statistical Association**, New York, v. 84, n. 408, p. 996-1002, 1989

FISHER, N. I.; LEWIS, T; WILLCOX, M. E. Tests of discordancy for samples from Fisher's distribution on the sphere. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, London, v. 30, n. 3, p. 230-237, 1981.

HAIR JÚNIOR, J. F. et al. **Análise multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005. 593 p.

HOTELLING, H. Review of the triumph of mediocrity in business. **Journal of the American Statistical Association**, New York, v. 28, n. 184, p. 463-465, Dec. 1933.

HUSSIN, A. G. et al. Asymptotic covariance and detection of influential observations in a linear functional relationship model for circular data with application to the measurements of wind directions. **ScienceAsia**, Bangkok, v. 36, n. 3, p. 249-253, 2010.

HUSSIN, A. G.; FIELLER, N. R. J.; STILLMAN, E. C. Linear regression for circular variables with application to directional data. **Journal of Applied Science and Technology**, Accra, v. 9, n. 1, p. 1-6, 2004.

IBRAHIM, S. et al. Outlier detection in a circular regression model using COVRATIO Statistic. **Communications in Statistics - Simulation and Computation**, New York, v. 42, n. 10, p. 2272-2280, 2013.

JAMMALAMADAKA, S. R.; SARMA, Y. R. Circular regression. In: STATISTICAL sciences and data analysis: proceedings of the Third Pacific Area Statistical Conference. [S.l.]: VSP Intl Science, 1993. v. 3, p 109-128.

JAMMALAMADAKA, S. R.; SENGUPTA, A. **Topics in circular statistics**. London: World Scientific Publication, 2001.

JOHNSON, M. E. **Multivariate statistical simulation**. New York: J. Wiley, 1987.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6. ed. Upper Saddle River: Pearson Prentice Hall, 2007. 773 p.

JOLLIFFE, I. T. **Principal component analysis**. 2. ed. New York: Springer Verlag, 2002. 487 p.

KOTZ, S.; BALAKRISHNAN, N.; JOHNSON, N. L. **Continuous multivariate distributions**. Hoboken: J. Wiley, 2004. v. 1, 218 p.

LANGE, K. L.; RODERICK, J. A. L.; TAYLOR, J. M. G. Robust statistical modeling using the t distribution. **Journal of the American Statistical Association**, New York, v. 84, n. 408, p. 881-896, Dec. 1989.

LEWIS, T.; FISHER, N. I. Graphical methods for investigating the fit of a Fisher distribution to spherical data. **Geophysical Journal of the Royal Astronomical Society**, Oxford, v. 69, n. 1, p. 1-13, 1982.

LITTELL, R. C.; PENDERGAST, J.; NATARAJAN, R. Modelling covariance structure in the analysis of repeated measures data. **Statistics in Medicine**, New York, v. 19, p. 1793-1819, 2000.

MARDIA, K. V. **Statistics of directional data**. London: Academic, 1972.

MORRISON, D. F. **Multivariate statistical methods**. 3. ed. New York: MxGraw-Hill, 1990. 495 p.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2013.

SILVA, G. F. da; PINTO JÚNIOR, D. L. Análise da performance de processos multivariados assimétricos. **Revista Eletrônica de Matemática**, Jataí, v. 1, n. 2, p. 1-7, 2010.

VINES, S. K. Simple principal components. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, London, v. 49, n. 4, p. 441-451, 2000.

WILKINSON, L. Dot plots. **The American Statistician**, Ames, v. 53, n. 3, p. 276-281, 1999.

ANEXOS

ANEXO A – Tabelas

Tabela 1 Média dos ângulos em graus considerando a distribuição Normal Assimétrica com. $\gamma = 0{,}15$

Estrutura de Correlação	γ	n	$\rho = 0{,}80$			$\rho = 0{,}50$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
AR(1)	0,15	50	1,92	30,09	19,89	6,80	31,37	18,78
		100	1,95	31,82	18,58	5,52	32,20	16,59
		200	1,85	32,78	18,13	5,02	32,42	16,17
CS	0,15	50	0,64	25,62	25,17	2,78	26,71	24,94
		100	0,33	26,08	24,62	2,05	26,97	23,59
		200	0,29	25,29	25,36	0,96	26,34	24,40
		n	$\boldsymbol{\rho}_1 = [0{,}9\ 0{,}8\ 0{,}7]$			$\boldsymbol{\rho}_2 = [0{,}6\ 0{,}5\ 0{,}4]$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
Toeplitz	0,15	50	1,16	30,73	19,80	1,68	31,21	19,15
		100	1,18	32,27	18,95	2,22	32,94	17,77
		200	1,22	32,60	18,76	2,03	32,90	17,85

Tabela 2 Média dos ângulos em graus considerando a distribuição log-Normal com. $\gamma = 0{,}15$

Estrutura de Correlação	γ	n	$\rho = 0{,}80$			$\rho = 0{,}50$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
AR(1)	0,15	50	1,87	29,63	21,17	4,17	28,81	22,39
		100	1,42	28,08	22,37	3,78	29,65	20,42
		200	1,91	31,54	18,68	4,01	29,45	19,96
CS	0,15	50	0,69	25,78	25,06	2,67	26,44	24,95
		100	0,75	26,32	24,06	1,64	26,49	24,32
		200	0,46	26,41	24,85	1,26	24,94	26,06
		n	$\boldsymbol{\rho}_1 = [0{,}9\ 0{,}8\ 0{,}7]$			$\boldsymbol{\rho}_2 = [0{,}6\ 0{,}5\ 0{,}4]$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
Toeplitz	0,15	50	1,07	29,26	21,44	1,35	29,93	20,40
		100	1,08	29,91	20,79	1,42	32,08	18,83
		200	1,13	31,61	19,38	1,41	33,68	18,27

Tabela 3 Média dos ângulos em graus considerando a distribuição t-Student com. $\gamma = 0{,}15$

Estrutura de Correlação	γ	n	$\rho = 0{,}80$			$\rho = 0{,}50$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
AR(1)	0,15	50	1,99	30,23	19,75	5,72	30,73	19,46
		100	2,09	31,32	18,91	5,06	31,45	18,06
		200	1,98	32,66	18,09	5,22	32,43	16,56
CS	0,15	50	0,72	25,97	24,82	2,94	27,62	23,35
		100	1,02	24,92	26,45	2,30	26,34	24,84
		200	0,35	25,54	24,67	1,17	25,82	25,03
		n	$\boldsymbol{\rho}_1 = [0{,}9\ 0{,}8\ 0{,}7]$			$\boldsymbol{\rho}_2 = [0{,}6\ 0{,}5\ 0{,}4]$		
			$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
Toeplitz	0,15	50	1,18	30,82	19,74	2,05	30,37	19,55
		100	1,26	31,77	19,19	1,82	31,75	18,64
		200	1,03	32,82	18,87	1,92	32,45	18,17

Tabela 4 Distância entre os ângulos considerando a Distribuição Normal Multivariada na estrutura AR(1)

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
50	0,50	na	PC1	0,05	5,98	0,07
50	0,50	na	PC1	0,15	6,80	0,89
50	0,50	na	PC1	0,30	6,37	0,46
50	0,50	t	PC1	0,05	5,95	0,04
50	0,50	t	PC1	0,15	5,72	0,19
50	0,50	t	PC1	0,30	6,35	0,44
50	0,50	ln	PC1	0,05	5,21	0,70
50	0,50	ln	PC1	0,15	4,17	1,74
50	0,50	ln	PC1	0,30	3,47	2,44
50	0,50	na	PC2	0,05	31,03	0,03
50	0,50	na	PC2	0,15	31,37	0,37
50	0,50	na	PC2	0,30	30,93	0,07
50	0,50	t	PC2	0,05	30,60	0,40
50	0,50	t	PC2	0,15	30,73	0,27
50	0,50	t	PC2	0,30	30,36	0,64
50	0,50	ln	PC2	0,05	29,54	1,46
50	0,50	ln	PC2	0,15	28,81	2,19
50	0,50	ln	PC2	0,30	27,64	3,36
50	0,50	na	PC3	0,05	19,05	0,80
50	0,50	na	PC3	0,15	18,78	0,53
50	0,50	na	PC3	0,30	19,12	0,87
50	0,50	t	PC3	0,05	19,83	1,58
50	0,50	t	PC3	0,15	19,46	1,21
50	0,50	t	PC3	0,30	20,64	2,39
50	0,50	ln	PC3	0,05	21,09	2,84
50	0,50	ln	PC3	0,15	22,39	4,14
50	0,50	ln	PC3	0,30	23,37	5,12
50	0,80	na	PC1	0,05	1,91	0,04
50	0,80	na	PC1	0,15	1,92	0,03
50	0,80	na	PC1	0,30	2,07	0,12

50	0,80	t	PC1	0,05	1,86	0,09

Tabela 4, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
50	0,80	t	PC1	0,15	1,99	0,04
50	0,80	t	PC1	0,30	2,50	0,55
50	0,80	ln	PC1	0,05	1,98	0,03
50	0,80	ln	PC1	0,15	1,87	0,08
50	0,80	ln	PC1	0,30	1,96	0,01
50	0,80	na	PC2	0,05	30,38	0,21
50	0,80	na	PC2	0,15	30,09	0,50
50	0,80	na	PC2	0,30	30,19	0,40
50	0,80	t	PC2	0,05	30,81	0,22
50	0,80	t	PC2	0,15	30,23	0,36
50	0,80	t	PC2	0,30	29,73	0,86
50	0,80	ln	PC2	0,05	29,24	1,35
50	0,80	ln	PC2	0,15	29,63	0,96
50	0,80	ln	PC2	0,30	27,72	2,87
50	0,80	na	PC3	0,05	19,79	0,05
50	0,80	na	PC3	0,15	19,89	0,15
50	0,80	na	PC3	0,30	19,77	0,03
50	0,80	t	PC3	0,05	19,55	0,19
50	0,80	t	PC3	0,15	19,75	0,01
50	0,80	t	PC3	0,30	20,28	0,54
50	0,80	ln	PC3	0,05	21,08	1,34
50	0,80	ln	PC3	0,15	21,17	1,43
50	0,80	ln	PC3	0,30	22,47	2,73
100	0,50	na	PC1	0,05	5,94	1,01
100	0,50	na	PC1	0,15	5,52	0,59
100	0,50	na	PC1	0,30	5,79	0,86
100	0,50	t	PC1	0,05	5,05	0,12
100	0,50	t	PC1	0,15	5,06	0,13
100	0,50	t	PC1	0,30	5,49	0,56
100	0,50	ln	PC1	0,05	4,80	0,13
100	0,50	ln	PC1	0,15	3,78	1,15

100	0,50	ln	PC1	0,30	4,18	0,75
100	0,50	na	PC2	0,05	32,06	1,02

Tabela 4, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
100	0,50	na	PC2	0,15	32,20	1,16
100	0,50	na	PC2	0,30	31,77	0,73
100	0,50	t	PC2	0,05	31,97	0,93
100	0,50	t	PC2	0,15	31,45	0,41
100	0,50	t	PC2	0,30	31,31	0,27
100	0,50	ln	PC2	0,05	31,19	0,15
100	0,50	ln	PC2	0,15	29,65	1,39
100	0,50	ln	PC2	0,30	29,80	1,24
100	0,50	na	PC3	0,05	16,91	1,06
100	0,50	na	PC3	0,15	16,59	1,38
100	0,50	na	PC3	0,30	17,31	0,66
100	0,50	t	PC3	0,05	16,79	1,18
100	0,50	t	PC3	0,15	18,06	0,09
100	0,50	t	PC3	0,30	18,04	0,07
100	0,50	ln	PC3	0,05	18,46	0,49
100	0,50	ln	PC3	0,15	20,42	2,45
100	0,50	ln	PC3	0,30	19,87	1,90
100	0,80	na	PC1	0,05	1,95	0,02
100	0,80	na	PC1	0,15	1,95	0,02
100	0,80	na	PC1	0,30	2,29	0,32
100	0,80	t	PC1	0,05	2,01	0,04
100	0,80	t	PC1	0,15	2,09	0,12
100	0,80	t	PC1	0,30	2,03	0,06
100	0,80	ln	PC1	0,05	1,84	0,13
100	0,80	ln	PC1	0,15	1,42	0,55
100	0,80	ln	PC1	0,30	1,96	0,01
100	0,80	na	PC2	0,05	31,84	0,46
100	0,80	na	PC2	0,15	31,82	0,44
100	0,80	na	PC2	0,30	31,96	0,58
100	0,80	t	PC2	0,05	31,63	0,25
100	0,80	t	PC2	0,15	31,32	0,06

100	0,80	t	PC2	0,30	31,17	0,21
100	0,80	ln	PC2	0,05	30,50	0,88

Tabela 4, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
100	0,80	ln	PC2	0,15	28,08	3,30
100	0,80	ln	PC2	0,30	30,21	1,17
100	0,80	na	PC3	0,05	18,55	0,47
100	0,80	na	PC3	0,15	18,58	0,44
100	0,80	na	PC3	0,30	18,34	0,68
100	0,80	t	PC3	0,05	18,67	0,35
100	0,80	t	PC3	0,15	18,91	0,11
100	0,80	t	PC3	0,30	19,00	0,02
100	0,80	ln	PC3	0,05	20,26	1,24
100	0,80	ln	PC3	0,15	22,37	3,35
100	0,80	ln	PC3	0,30	19,95	0,93
200	0,50	na	PC1	0,05	5,15	0,79
200	0,50	na	PC1	0,15	5,02	0,66
200	0,50	na	PC1	0,30	5,42	1,06
200	0,50	t	PC1	0,05	4,48	0,12
200	0,50	t	PC1	0,15	5,22	0,86
200	0,50	t	PC1	0,30	4,87	0,51
200	0,50	ln	PC1	0,05	4,53	0,17
200	0,50	ln	PC1	0,15	4,01	0,35
200	0,50	ln	PC1	0,30	4,12	0,24
200	0,50	na	PC2	0,05	32,84	0,01
200	0,50	na	PC2	0,15	32,42	0,43
200	0,50	na	PC2	0,30	32,03	0,82
200	0,50	t	PC2	0,05	31,77	1,08
200	0,50	t	PC2	0,15	32,43	0,42
200	0,50	t	PC2	0,30	32,01	0,84
200	0,50	ln	PC2	0,05	32,13	0,72
200	0,50	ln	PC2	0,15	29,45	3,40
200	0,50	ln	PC2	0,30	30,24	2,61
200	0,50	na	PC3	0,05	15,82	0,50
200	0,50	na	PC3	0,15	16,17	0,15

200	0,50	na	PC3	0,30	16,81	0,49
200	0,50	t	PC3	0,05	17,57	1,25

Tabela 4, conclusão

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
200	0,50	t	PC3	0,15	16,56	0,24
200	0,50	t	PC3	0,30	17,26	0,94
200	0,50	ln	PC3	0,05	16,64	0,32
200	0,50	ln	PC3	0,15	19,96	3,64
200	0,50	ln	PC3	0,30	20,46	4,14
200	0,80	na	PC1	0,05	2,02	0,06
200	0,80	na	PC1	0,15	1,85	0,11
200	0,80	na	PC1	0,30	2,07	0,11
200	0,80	t	PC1	0,05	1,83	0,13
200	0,80	t	PC1	0,15	1,98	0,02
200	0,80	t	PC1	0,30	1,89	0,07
200	0,80	ln	PC1	0,05	1,72	0,24
200	0,80	ln	PC1	0,15	1,91	0,05
200	0,80	ln	PC1	0,30	1,57	0,39
200	0,80	na	PC2	0,05	32,71	0,36
200	0,80	na	PC2	0,15	32,78	0,29
200	0,80	na	PC2	0,30	32,41	0,66
200	0,80	t	PC2	0,05	32,97	0,10
200	0,80	t	PC2	0,15	32,66	0,41
200	0,80	t	PC2	0,30	31,84	1,23
200	0,80	ln	PC2	0,05	32,11	0,96
200	0,80	ln	PC2	0,15	31,54	1,53
200	0,80	ln	PC2	0,30	29,66	3,41
200	0,80	na	PC3	0,05	18,01	0,10
200	0,80	na	PC3	0,15	18,13	0,22
200	0,80	na	PC3	0,30	18,17	0,26
200	0,80	t	PC3	0,05	18,06	0,15
200	0,80	t	PC3	0,15	18,09	0,18
200	0,80	t	PC3	0,30	18,59	0,68
200	0,80	ln	PC3	0,05	18,58	0,67
200	0,80	ln	PC3	0,15	18,68	0,77

200	0,80	ln	PC3	0,30	20,66	2,75

Tabela 5 Distância entre os ângulos considerando a Distribuição Normal Multivariada na estrutura CS

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
50	0,5	na	PC1	0,05	3,08	0,53
50	0,5	na	PC1	0,15	2,78	0,23
50	0,5	na	PC1	0,30	2,85	0,3
50	0,5	t	PC1	0,05	3,31	0,76
50	0,5	t	PC1	0,15	2,94	0,39
50	0,5	t	PC1	0,30	2,47	0,08
50	0,5	ln	PC1	0,05	2,56	0,01
50	0,5	ln	PC1	0,15	2,67	0,12
50	0,5	ln	PC1	0,30	2,68	0,13
50	0,5	na	PC2	0,05	27,27	0,28
50	0,5	na	PC2	0,15	26,71	0,84
50	0,5	na	PC2	0,30	27,41	0,14
50	0,5	t	PC2	0,05	28,04	0,49
50	0,5	t	PC2	0,15	27,62	0,07
50	0,5	t	PC2	0,30	27,16	0,39
50	0,5	ln	PC2	0,05	26,81	0,74
50	0,5	ln	PC2	0,15	26,44	1,11
50	0,5	ln	PC2	0,30	26,37	1,18
50	0,5	na	PC3	0,05	24,16	1,06
50	0,5	na	PC3	0,15	24,94	1,84
50	0,5	na	PC3	0,30	23,66	0,56
50	0,5	t	PC3	0,05	23,2	0,1
50	0,5	t	PC3	0,15	23,35	0,25
50	0,5	t	PC3	0,30	23,69	0,59
50	0,5	ln	PC3	0,05	24,3	1,2
50	0,5	ln	PC3	0,15	24,95	1,85
50	0,5	ln	PC3	0,30	25,69	2,59
50	0,8	na	PC1	0,05	0,72	0,1
50	0,8	na	PC1	0,15	0,64	0,02

50	0,8	na	PC1	0,30	0,65	0,03
50	0,8	t	PC1	0,05	0,68	0,06
50	0,8	t	PC1	0,15	0,72	0,1

Tabela 5, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
50	0,8	t	PC1	0,30	0,81	0,19
50	0,8	ln	PC1	0,05	0,64	0,02
50	0,8	ln	PC1	0,15	0,69	0,07
50	0,8	ln	PC1	0,30	0,85	0,23
50	0,8	na	PC2	0,05	25,28	0,09
50	0,8	na	PC2	0,15	25,62	0,43
50	0,8	na	PC2	0,30	25,71	0,52
50	0,8	t	PC2	0,05	25,17	0,02
50	0,8	t	PC2	0,15	25,97	0,78
50	0,8	t	PC2	0,30	25,54	0,35
50	0,8	ln	PC2	0,05	25,76	0,57
50	0,8	ln	PC2	0,15	25,78	0,59
50	0,8	ln	PC2	0,30	25,6	0,41
50	0,8	na	PC3	0,05	25,46	1,66
50	0,8	na	PC3	0,15	25,17	1,37
50	0,8	na	PC3	0,30	25,05	1,25
50	0,8	t	PC3	0,05	25,57	1,77
50	0,8	t	PC3	0,15	24,82	1,02
50	0,8	t	PC3	0,30	25,23	1,43
50	0,8	ln	PC3	0,05	24,98	1,18
50	0,8	ln	PC3	0,15	25,06	1,26
50	0,8	ln	PC3	0,30	25,22	1,42
100	0,5	na	PC1	0,05	1,56	0,31
100	0,5	na	PC1	0,15	2,05	0,18
100	0,5	na	PC1	0,30	1,85	0,02
100	0,5	t	PC1	0,05	1,79	0,08
100	0,5	t	PC1	0,15	2,3	0,43
100	0,5	t	PC1	0,30	2,12	0,25
100	0,5	ln	PC1	0,05	1,72	0,15
100	0,5	ln	PC1	0,15	1,64	0,23

100	0,5	ln	PC1	0,30	2,01	0,14
100	0,5	na	PC2	0,05	26,67	0,2
100	0,5	na	PC2	0,15	26,97	0,5

Tabela 5, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
100	0,5	na	PC2	0,30	25,95	0,52
100	0,5	t	PC2	0,05	27,08	0,61
100	0,5	t	PC2	0,15	26,34	0,13
100	0,5	t	PC2	0,30	26,72	0,25
100	0,5	ln	PC2	0,05	26,21	0,26
100	0,5	ln	PC2	0,15	26,49	0,02
100	0,5	ln	PC2	0,30	25,83	0,64
100	0,5	na	PC3	0,05	24,02	0,49
100	0,5	na	PC3	0,15	23,59	0,92
100	0,5	na	PC3	0,30	24,94	0,43
100	0,5	t	PC3	0,05	23,73	0,78
100	0,5	t	PC3	0,15	24,84	0,33
100	0,5	t	PC3	0,30	24,14	0,37
100	0,5	ln	PC3	0,05	24,8	0,29
100	0,5	ln	PC3	0,15	24,32	0,19
100	0,5	ln	PC3	0,30	25,58	1,07
100	0,8	na	PC1	0,05	0,46	0,03
100	0,8	na	PC1	0,15	0,33	0,1
100	0,8	na	PC1	0,30	0,52	0,09
100	0,8	t	PC1	0,05	0,36	0,07
100	0,8	t	PC1	0,15	1,02	0,59
100	0,8	t	PC1	0,30	0,57	0,14
100	0,8	ln	PC1	0,05	0,52	0,09
100	0,8	ln	PC1	0,15	0,75	0,32
100	0,8	ln	PC1	0,30	0,76	0,33
100	0,8	na	PC2	0,05	25,69	0,34
100	0,8	na	PC2	0,15	26,08	0,73
100	0,8	na	PC2	0,30	25,9	0,55
100	0,8	t	PC2	0,05	25,13	0,22
100	0,8	t	PC2	0,15	24,92	0,43

100	0,8	t	PC2	0,30	26,21	0,86
100	0,8	ln	PC2	0,05	24,81	0,54
100	0,8	ln	PC2	0,15	26,32	0,97

Tabela 5, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
100	0,8	ln	PC2	0,30	26,28	0,93
100	0,8	na	PC3	0,05	25	0,68
100	0,8	na	PC3	0,15	24,62	0,3
100	0,8	na	PC3	0,30	24,84	0,52
100	0,8	t	PC3	0,05	25,77	1,45
100	0,8	t	PC3	0,15	26,45	2,13
100	0,8	t	PC3	0,30	24,45	0,13
100	0,8	ln	PC3	0,05	26,19	1,87
100	0,8	ln	PC3	0,15	24,06	0,26
100	0,8	ln	PC3	0,30	24,18	0,14
200	0,5	na	PC1	0,05	1,22	0,16
200	0,5	na	PC1	0,15	0,96	0,42
200	0,5	na	PC1	0,30	1,22	0,16
200	0,5	t	PC1	0,05	1,16	0,22
200	0,5	t	PC1	0,15	1,17	0,21
200	0,5	t	PC1	0,30	1,69	0,31
200	0,5	ln	PC1	0,05	1,32	0,06
200	0,5	ln	PC1	0,15	1,26	0,12
200	0,5	ln	PC1	0,30	1,33	0,05
200	0,5	na	PC2	0,05	25,76	0,09
200	0,5	na	PC2	0,15	26,34	0,49
200	0,5	na	PC2	0,30	26,67	0,82
200	0,5	t	PC2	0,05	25,32	0,53
200	0,5	t	PC2	0,15	25,82	0,03
200	0,5	t	PC2	0,30	24,91	0,94
200	0,5	ln	PC2	0,05	27,05	1,2
200	0,5	ln	PC2	0,15	24,94	0,91
200	0,5	ln	PC2	0,30	26,11	0,26
200	0,5	na	PC3	0,05	25,2	0,11
200	0,5	na	PC3	0,15	24,4	0,69

200	0,5	na	PC3	0,30	23,87	1,22
200	0,5	t	PC3	0,05	25,9	0,81
200	0,5	t	PC3	0,15	25,03	0,06

Tabela 5, conclusão

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
200	0,5	t	PC3	0,30	26,8	1,71
200	0,5	ln	PC3	0,05	23,32	1,77
200	0,5	ln	PC3	0,15	26,06	0,97
200	0,5	ln	PC3	0,30	24,66	0,43
200	0,8	na	PC1	0,05	0,29	0,03
200	0,8	na	PC1	0,15	0,29	0,03
200	0,8	na	PC1	0,30	0,34	0,02
200	0,8	t	PC1	0,05	0,3	0,02
200	0,8	t	PC1	0,15	0,35	0,03
200	0,8	t	PC1	0,30	0,36	0,04
200	0,8	ln	PC1	0,05	0,47	0,15
200	0,8	ln	PC1	0,15	0,46	0,14
200	0,8	ln	PC1	0,30	0,87	0,55
200	0,8	na	PC2	0,05	25,12	0,1
200	0,8	na	PC2	0,15	25,29	0,07
200	0,8	na	PC2	0,30	25,03	0,19
200	0,8	t	PC2	0,05	25,69	0,47
200	0,8	t	PC2	0,15	25,54	0,32
200	0,8	t	PC2	0,30	26,08	0,86
200	0,8	ln	PC2	0,05	25,68	0,46
200	0,8	ln	PC2	0,15	26,41	1,19
200	0,8	ln	PC2	0,30	26,17	0,95
200	0,8	na	PC3	0,05	25,53	1,77
200	0,8	na	PC3	0,15	25,36	1,6
200	0,8	na	PC3	0,30	25,67	1,91
200	0,8	t	PC3	0,05	25,08	1,32
200	0,8	t	PC3	0,15	24,67	0,91
200	0,8	t	PC3	0,30	24,82	1,06
200	0,8	ln	PC3	0,05	24,75	0,99
200	0,8	ln	PC3	0,15	24,85	1,09

200	0,8	ln	PC3	0,30	24,55	0,79

Tabela 6 Distância entre os ângulos considerando a Distribuição Normal Multivariada na estrutura Toeplitz

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
50	0,50	na	PC1	0,05	1,92	0,09
50	0,50	na	PC1	0,15	1,68	0,33
50	0,50	na	PC1	0,30	2,20	0,19
50	0,50	t	PC1	0,05	2,23	0,22
50	0,50	t	PC1	0,15	2,05	0,04
50	0,50	t	PC1	0,30	2,07	0,06
50	0,50	ln	PC1	0,05	1,82	0,19
50	0,50	ln	PC1	0,15	1,35	0,66
50	0,50	ln	PC1	0,30	1,34	0,67
50	0,50	na	PC2	0,05	31,42	0,86
50	0,50	na	PC2	0,15	31,21	0,65
50	0,50	na	PC2	0,30	30,52	0,04
50	0,50	t	PC2	0,05	31,42	0,86
50	0,50	t	PC2	0,15	30,37	0,19
50	0,50	t	PC2	0,30	30,77	0,21
50	0,50	ln	PC2	0,05	31,23	0,67
50	0,50	ln	PC2	0,15	29,93	0,63
50	0,50	ln	PC2	0,30	30,13	0,43
50	0,50	na	PC3	0,05	18,86	0,62
50	0,50	na	PC3	0,15	19,15	0,33
50	0,50	na	PC3	0,30	19,22	0,26
50	0,50	t	PC3	0,05	18,63	0,85
50	0,50	t	PC3	0,15	19,55	0,07
50	0,50	t	PC3	0,30	19,24	0,24
50	0,50	ln	PC3	0,05	19,03	0,45
50	0,50	ln	PC3	0,15	20,40	0,92
50	0,50	ln	PC3	0,30	20,45	0,97
50	0,80	na	PC1	0,05	1,21	0,02

50	0,80	na	PC1	0,15	1,16	0,07

Tabela 6, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
50	0,80	na	PC1	0,30	1,42	0,19
50	0,80	t	PC1	0,05	1,16	0,07
50	0,80	t	PC1	0,15	1,18	0,05
50	0,80	t	PC1	0,30	1,28	0,05
50	0,80	ln	PC1	0,05	1,23	0,00
50	0,80	ln	PC1	0,15	1,07	0,16
50	0,80	ln	PC1	0,30	1,26	0,03
50	0,80	na	PC2	0,05	30,78	0,20
50	0,80	na	PC2	0,15	30,73	0,15
50	0,80	na	PC2	0,30	31,06	0,48
50	0,80	t	PC2	0,05	30,85	0,27
50	0,80	t	PC2	0,15	30,82	0,24
50	0,80	t	PC2	0,30	30,13	0,45
50	0,80	ln	PC2	0,05	30,63	0,05
50	0,80	ln	PC2	0,15	29,26	1,32
50	0,80	ln	PC2	0,30	29,35	1,23
50	0,80	na	PC3	0,05	19,81	0,06
50	0,80	na	PC3	0,15	19,80	0,07
50	0,80	na	PC3	0,30	19,50	0,37
50	0,80	t	PC3	0,05	20,07	0,20
50	0,80	t	PC3	0,15	19,74	0,13
50	0,80	t	PC3	0,30	20,27	0,40
50	0,80	ln	PC3	0,05	19,97	0,10
50	0,80	ln	PC3	0,15	21,44	1,57
50	0,80	ln	PC3	0,30	20,96	1,09
100	0,50	na	PC1	0,05	1,95	0,09
100	0,50	na	PC1	0,15	2,22	0,18
100	0,50	na	PC1	0,30	1,88	0,16

100	0,50	t	PC1	0,05	2,05	0,01
100	0,50	t	PC1	0,15	1,82	0,22
100	0,50	t	PC1	0,30	2,21	0,17
100	0,50	ln	PC1	0,05	1,68	0,36
100	0,50	ln	PC1	0,15	1,42	0,62

Tabela 6, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
100	0,50	ln	PC1	0,30	1,16	0,88
100	0,50	na	PC2	0,05	32,58	0,22
100	0,50	na	PC2	0,15	32,94	0,58
100	0,50	na	PC2	0,30	32,93	0,57
100	0,50	t	PC2	0,05	32,57	0,21
100	0,50	t	PC2	0,15	31,75	0,61
100	0,50	t	PC2	0,30	32,57	0,21
100	0,50	ln	PC2	0,05	31,94	0,42
100	0,50	ln	PC2	0,15	32,08	0,28
100	0,50	ln	PC2	0,30	29,47	2,89
100	0,50	na	PC3	0,05	18,16	0,06
100	0,50	na	PC3	0,15	17,77	0,45
100	0,50	na	PC3	0,30	18,06	0,16
100	0,50	t	PC3	0,05	18,03	0,19
100	0,50	t	PC3	0,15	18,64	0,42
100	0,50	t	PC3	0,30	17,92	0,30
100	0,50	ln	PC3	0,05	18,77	0,55
100	0,50	ln	PC3	0,15	18,83	0,61
100	0,50	ln	PC3	0,30	21,23	3,01
100	0,80	na	PC1	0,05	1,20	0,01
100	0,80	na	PC1	0,15	1,18	0,03
100	0,80	na	PC1	0,30	1,22	0,01
100	0,80	t	PC1	0,05	1,18	0,03
100	0,80	t	PC1	0,15	1,26	0,05
100	0,80	t	PC1	0,30	1,21	0,00
100	0,80	ln	PC1	0,05	1,11	0,10
100	0,80	ln	PC1	0,15	1,08	0,13
100	0,80	ln	PC1	0,30	0,98	0,23

100	0,80	na	PC2	0,05	32,19	0,13
100	0,80	na	PC2	0,15	32,27	0,21
100	0,80	na	PC2	0,30	32,96	0,90
100	0,80	t	PC2	0,05	31,47	0,59
100	0,80	t	PC2	0,15	31,77	0,29

Tabela 6, continua

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
100	0,80	t	PC2	0,30	31,89	0,17
100	0,80	ln	PC2	0,05	31,69	0,37
100	0,80	ln	PC2	0,15	29,91	2,15
100	0,80	ln	PC2	0,30	29,24	2,82
100	0,80	na	PC3	0,05	18,99	0,08
100	0,80	na	PC3	0,15	18,95	0,12
100	0,80	na	PC3	0,30	18,65	0,42
100	0,80	t	PC3	0,05	19,31	0,24
100	0,80	t	PC3	0,15	19,19	0,12
100	0,80	t	PC3	0,30	19,20	0,13
100	0,80	ln	PC3	0,05	19,31	0,24
100	0,80	ln	PC3	0,15	20,79	1,72
100	0,80	ln	PC3	0,30	21,48	2,41
200	0,50	na	PC1	0,05	1,68	0,27
200	0,50	na	PC1	0,15	2,03	0,08
200	0,50	na	PC1	0,30	1,81	0,14
200	0,50	t	PC1	0,05	1,95	0,00
200	0,50	t	PC1	0,15	1,92	0,03
200	0,50	t	PC1	0,30	2,16	0,21
200	0,50	ln	PC1	0,05	1,55	0,40
200	0,50	ln	PC1	0,15	1,41	0,54
200	0,50	ln	PC1	0,30	1,19	0,76
200	0,50	na	PC2	0,05	33,06	0,06
200	0,50	na	PC2	0,15	32,90	0,22
200	0,50	na	PC2	0,30	33,34	0,22
200	0,50	t	PC2	0,05	31,83	1,29
200	0,50	t	PC2	0,15	32,45	0,67
200	0,50	t	PC2	0,30	32,94	0,18

200	0,50	ln	PC2	0,05	32,84	0,28
200	0,50	ln	PC2	0,15	33,68	0,56
200	0,50	ln	PC2	0,30	32,09	1,03
200	0,50	na	PC3	0,05	18,16	0,22
200	0,50	na	PC3	0,15	17,85	0,09

Tabela 6, conclusão

n	ρ	Distribuição	PC	γ	$\hat{a}$	Distância
200	0,50	na	PC3	0,30	17,95	0,01
200	0,50	t	PC3	0,05	18,58	0,64
200	0,50	t	PC3	0,15	18,17	0,23
200	0,50	t	PC3	0,30	17,88	0,06
200	0,50	ln	PC3	0,05	18,34	0,40
200	0,50	ln	PC3	0,15	18,27	0,33
200	0,50	ln	PC3	0,30	19,14	1,20
200	0,80	na	PC1	0,05	1,18	0,03
200	0,80	na	PC1	0,15	1,22	0,07
200	0,80	na	PC1	0,30	1,17	0,02
200	0,80	t	PC1	0,05	1,10	0,05
200	0,80	t	PC1	0,15	1,03	0,12
200	0,80	t	PC1	0,30	1,14	0,01
200	0,80	ln	PC1	0,05	1,12	0,03
200	0,80	ln	PC1	0,15	1,13	0,02
200	0,80	ln	PC1	0,30	1,34	0,19
200	0,80	na	PC2	0,05	33,19	0,17
200	0,80	na	PC2	0,15	32,60	0,42
200	0,80	na	PC2	0,30	32,50	0,52
200	0,80	t	PC2	0,05	33,22	0,20
200	0,80	t	PC2	0,15	32,82	0,20
200	0,80	t	PC2	0,30	32,66	0,36
200	0,80	ln	PC2	0,05	31,80	1,22
200	0,80	ln	PC2	0,15	31,61	1,41
200	0,80	ln	PC2	0,30	28,49	4,53
200	0,80	na	PC3	0,05	18,68	0,10
200	0,80	na	PC3	0,15	18,76	0,02
200	0,80	na	PC3	0,30	18,83	0,05

200	0,80	t	PC3	0,05	18,66	0,12
200	0,80	t	PC3	0,15	18,87	0,09
200	0,80	t	PC3	0,30	18,85	0,07
200	0,80	ln	PC3	0,05	19,81	1,03
200	0,80	ln	PC3	0,15	19,38	0,60
200	0,80	ln	PC3	0,30	21,50	2,72

ANEXO B – Gráficos P-P Plot da distribuição das distâncias

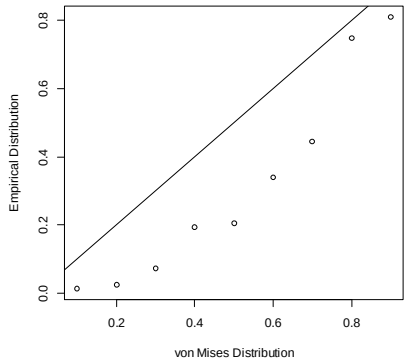


Figura 1 P-P Plot da distribuição Von-mises estrutura AR(1), n=50, ρ =0,5 e CP 1

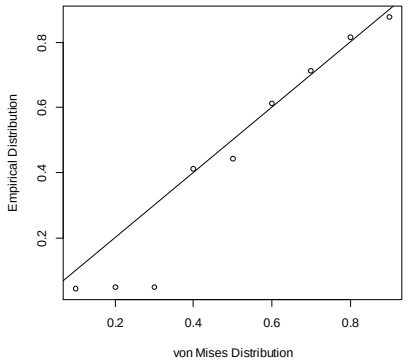
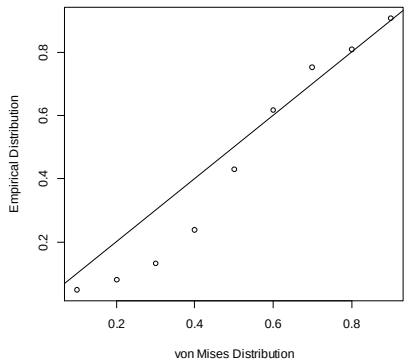


Figura 2 P-P Plot da distribuição Von-mises estrutura AR(1), n=100, ρ =0,5 e CP 1



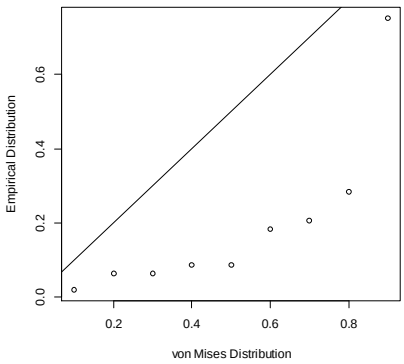


Figura 3 P-P Plot da distribuição Von-mises estrutura AR(1), n=200, ρ =0,5 e CP 1

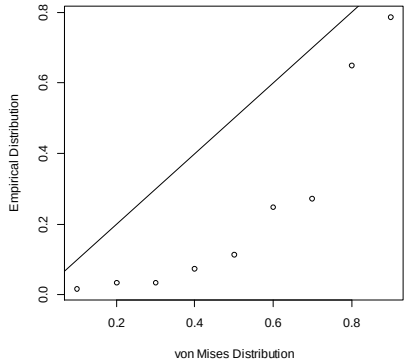


Figura 4 P-P Plot da distribuição Von-mises estrutura AR(1), n=50, ρ =0,8 e CP 1

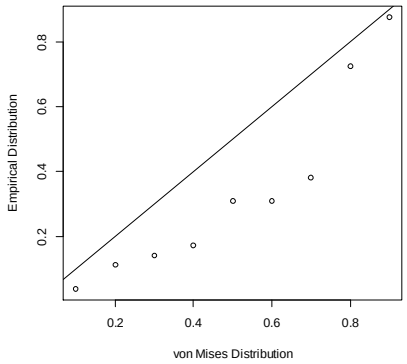


Figura 5 P-P Plot da distribuição Von-mises estrutura AR(1), n=100, ρ =0,8 e CP 1

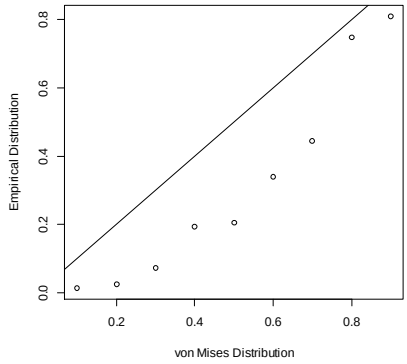


Figura 6 P-P Plot da distribuição Von-mises estrutura AR(1), n=200, ρ =0,8 e CP 1

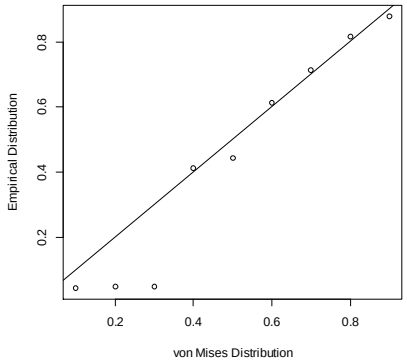


Figura 7 P-P Plot da distribuição Von-mises estrutura Toeplitz,

Figura 8 P-P Plot da distribuição Von-mises estrutura Toeplitz,

n=50, ρ =0,5 e CP 1

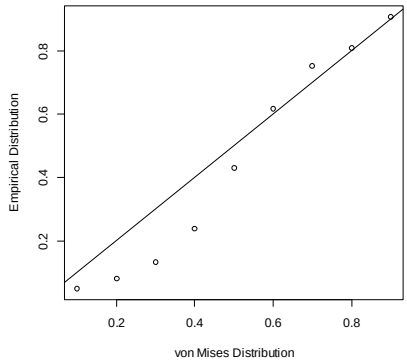


Figura 9 P-P Plot da distribuição Von-mises estrutura Toeplitz, *n*=200, ρ =0,5 e CP 1

n=100, ρ =0,5 e CP 1

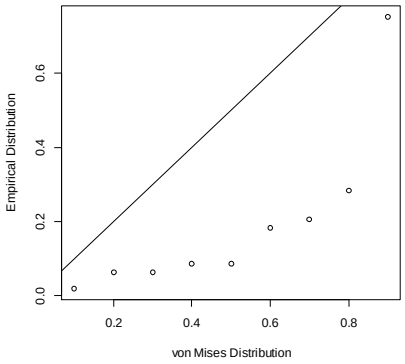


Figura 10 P-P Plot da distribuição Von-mises estrutura Toeplitz, *n*=50, ρ =0,8 e CP 1

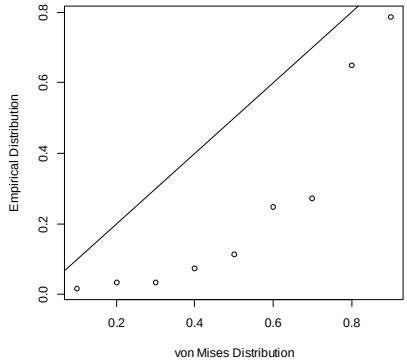


Figura 11 P-P Plot da distribuição Von-mises estrutura Toeplitz, *n*=100, ρ =0,8 e

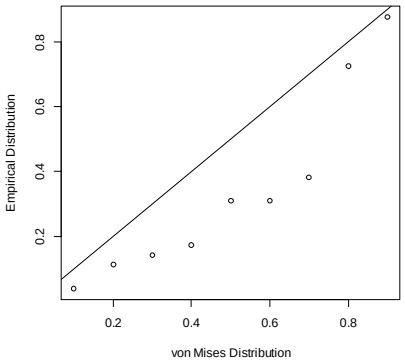


Figura 12 P-P Plot da distribuição Von-mises estrutura Toeplitz, *n*=200, ρ =0,8 e

CP 1 CP 1

ANEXO C – Rotinas computacionais para obtenção das médias angulares

```
library(fMultivar)
library(mvtnorm)
library(circular)

# ##################### Parametros ##################### #
p=3    # n. de variáveis (fixo)
nsim=1000 # número de simulaçoes (fixo)
pho=0.8  # grau de correlaçao entre variáveis #
n=201    # tamanho amostral (lembrarque n + 1 ) #
gama=0.30 # probabilidade de mistura

################## Matriz de covariância ############################ #

AR=matrix(0,p,p) ; eco=matrix (0,p,p) ; mi1=c(rep(0,p)) ; X=matrix(0,1,p)
for (i in 1:p)
 {
 for (j in 1:p)
  {
  if (i==j) AR[i,j]=1
  if (i!=j) AR[i,j]=pho^(abs(i-j))

  if (i==j) eco[i,i]=1
  if (i!=j) eco[i,j]=pho
  }
 }

# ########## Estrutura de correlaçao ###################### #
#covp <- AR        # ##### Matriz AR(1) #### #
#covp <- eco       # ##### Matriz Eco #### #
covp <- toeplitz(c(0.6, 0.5, 0.4))  # ## Matriz circular ## #

# ########### constantes normalizadoras ########## #

#c1p=1/sqrt(1) ; c1n=-c1p
#c2p=1/sqrt(2) ; c2n=-c2p
c3p=1/sqrt(3) ; c3n=-c3p
```

```
vetcp=c(c3p,c3n,c3p)


# #### Funçao de redução de dimensão interpretável ##### #

RDI=function(vetcp,coef)
{
 # vetcp: vetor das constantes normalizadoras #
 # coef: vetor linha dos coeficientes do PCA #
 alfacor=c(0,0,0)  ;  alfa=c(0,0,0)  ;  matalfa=matrix(0,1,3)  ;
                  nor_alfa=matrix(0,500,1) ; vetcp=as.vector(vetcp)
 for (cb in 1:500)
 {
  for (z in 1:3)
  {
   u=round(runif(1,1,3))
   alfa[z]=vetcp[u]
  }
  nor_alfa[cb,1]=as.matrix(t(alfa)%*%alfa)
  matalfa=rbind(matalfa,t(as.matrix(alfa)))
 }
 resalfa=cbind(matalfa[2:nrow(matalfa),],nor_alfa)
 frameres=as.data.frame(resalfa)
 sel=resalfa[frameres$V4=='1',]


 # #### correspondência de sinais entre coef dos PCA e interpretáveis ###### #
 a=sign(coef)
 for (k in 1:nrow(sel))
 {
  b=sign(sel[k,1:3])
  if (a[1]==b[1] && a[2]==b[2] && a[3]==b[3])
  {
   alfacor[1]=sel[k,1] ; alfacor[2]=sel[k,2] ;alfacor[3]=sel[k,3]
  }
 }
 theta=acos(coef%*%alfacor)
 return (list(ang=theta, alfsel=alfacor,dad=sel))
}

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

```
resp_ang=matrix(0,nsim,3) ; resp_alfa=c(0,0,0) ; nulo=0
res_coef_PC1=c(0,0,0) ; res_coef_PC2=c(0,0,0) ; res_coef_PC3=c(0,0,0)


###############INICIO DA SIMULAÇAO ################### #


for (s in 1:nsim)
 {
  ### Normal contaminada #####
  for (r in 1:n)
  {
  u=runif(1)
   obs=rmvnorm(1, mean=mi1, sigma=covp)

  # ################# Distribuiçao de referencia ################# #
   if (u>=gama)  Xaux <- obs

  # ########## geraçao dos outliers ########################### #
  #if  (u<gama)    Xaux  <-  rmvsnorm(1,3,mu=rep(0,3),Omega=covp,alpha=rep(-
                    20,3)) # Normal Assimétrica #
  if (u<gama)  Xaux <-exp(obs)          # log-normal #
  #if (u<gama)  Xaux <- rmvt(1,sigma=covp, df=5)  # t-student
  X=rbind(X,Xaux)
 }
 X=X[2:n,1:p]


 # ######## Obtençao dos PCA ############# #
 comp=princomp(X,cor=T)
 coef=comp$loadings
 mcoef=as.matrix(coef)
 RDI_PCA1=RDI(vetcp,mcoef[,1])
 RDI_PCA2=RDI(vetcp,mcoef[,2])
 RDI_PCA3=RDI(vetcp,mcoef[,3])
 res_coef_PC1=rbind(res_coef_PC1,mcoef[,1])
 res_coef_PC2=rbind(res_coef_PC2,mcoef[,2])
 res_coef_PC3=rbind(res_coef_PC3,mcoef[,3])
  resp_alfa=cbind(resp_alfa,RDI_PCA1$alfsel,RDI_PCA2$alfsel,RDI_PCA3$alf
                    sel)
 resp_ang[s,1]=RDI_PCA1$ang
 resp_ang[s,2]=RDI_PCA2$ang
```

```
 resp_ang[s,3]=RDI_PCA3$ang
  }

########## medidas angulares ######################### #
###### Obtenção das estatísticas angulares
#### Média para o primeiro vetor de ângulos
ângulos1=circular(resp_ang[,1])
média1=summary(angulos1)


#### Média para o segundo vetor de ângulos
ângulos2=circular(resp_ang[,2])
média2=summary(angulos2)

#### Média para o segundo vetor de ângulos
ângulos3=circular(resp_ang[,3])
média3=summary(ângulos3)
```