DOI: http://dx.doi.org/10.5902/2179-460X856
Revista do Centro do Ciências Naturais e Exatas - UFSM
Ciência e Natura, Santa Maria,
ISSN: 2179-460X, v. 35 n. 2 dezembro, 2013, p. 095-104

Efeito de diferentes estruturas de correlação nos ângulos formados entre componentes principais e interpretáveis em amostras com presença de pontos discrepantes

Effect of different correlation structures in angles formed between principal and interpretable components in samples witch presences of outliers

Augusto Maciel da Silva[1], Augusto Ramalho De Morais[2], Marcelo Angelo Cirillo[3]

[1]Universidade Federal de Santa Maria, [2]Universidade Federal de Lavras, Lavras, Minas Gerais,
[3]Universidade Federal de Lavras, Lavras, Minas Gerais.

Resumo

Análise de Componentes Principais (ACP) tem como objetivo descrever a estrutura de covariâncias de um vetor aleatório utilizando combinações lineares das variáveis originais. Em algumas situações, os coeficientes dos Componentes Principais (CP) podem não ser facilmente interpretados devido ao número de variáveis ou presença de pontos discrepantes. Assim foram introduzidos os Componentes Interpretáveis (CI), os quais são avaliados através do ângulo formado entre os mesmos e os Componentes Principais. O presente trabalho tem como objetivo avaliar os efeitos de diferentes estruturas de correlação via Simulação de Monte Carlo e estatística circular na distribuição dos ângulos formados entre os componentes em amostras com e sem contaminação. Foi verificado que as estruturas de correlação atuam de forma diferente nos ângulos, sendo a estrutura de Simetria Composta a que apresenta menores médias angulares para os primeiros componentes em situações de maior coeficiente de correlação. Foi verificado também que a contaminação da amostra não atua diretamente na magnitude dos valores esperados dos ângulos.

***Palavras-chave**: estatística circular, simulação de Monte Carlo, direção média, pontos discrepantes.*

Abstract

The principal component analysis aims to explain the variance structure of a random vector consisting of p variables, using linear combinations of the original variables. In some situations, the coefficients of the principal components may not be easily interpreted because the number of variables or the presence of outliers. Thus were introduced interpretable components, which are measured by the angle formed between the Principal and Interpretable Component. This paper aims to evaluate the effects of different correlation structures via Monte Carlo simulation and circular statistics on the angles formed between the components in samples with and without contamination. It was found that the structures act differently on the angles, and the CS structure which has the smallest expected angle for the first components in situations of higher correlation coefficient. Still, it was found that the contamination of the sample does not act directly on the magnitude of the expected values of the angles.

***Keywords**: Circular Statistics, Monte Carlo Simulation, Mean Direction, Contamination.*

Recebido em: 2013-08-26, Aceito em: 2013-11-24

1. Introdução

Análises estatísticas envolvendo muitas variáveis têm interpretações nem tanto triviais, podendo assumir um alto grau de complexidade. As variáveis envolvidas em determinado processo podem frequentemente apresentar algum tipo de relação entre si. As técnicas de análise multivariada permitem a utilização de modelos mais simplificados, que explorem entre outras características, estas possíveis relações.

A análise de Componentes Principais tem por característica explicar a estrutura de variância e covariância de um conjunto de variáveis através de poucas combinações lineares destas variáveis. Assim, pode-se citar dois objetivos que são a redução da dimensionalidade dos dados e a interpretação (JOHNSON e WICHERN, 2007), sendo a garantia da explicação da variabilidade pela redução da dimensão, o objetivo mais comumente observado na análise.

Apesar da facilidade de aplicação da técnica de Componentes Principais (CP), estes podem apresentar coeficientes de difícil interpretação. Assim, Chipman e Gu (2005) introduziram algumas restrições aos componentes de forma a torná-los mais interpretáveis, restringindo os coeficientes a um número reduzido e obtendo assim os chamados Componentes Interpretáveis (CI). Outros estudos sobre interpretação de componentes podem ser encontrados em Vines (2000) e mais recentemente em Enki et al. (2013), que considera a interpretabilidade dos componentes conjuntamente com análise de agrupamentos.

A avaliação dos CI pode ser feita através da obtenção do ângulo entre o eixo formado pelo CI e o eixo formado pelo CP, que deve ser o menor possível, a fim de garantir a representatividade. Dessa forma torna-se necessário o conhecimento desses ângulos, que formam um conjunto de dados circulares.

Dados circulares 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 uma direção. A periodicidade dos dados circulares os caracteriza de forma diferente de observações na reta, sendo necessárias algumas restrições ao se trabalhar com esse tipo de dados, que possuem definições apropriadas de medidas de posição bem como modelos probabilísticos adequados, que são tratados pela estatística circular (FISHER, 1993).

Os dados circulares estão sujeitos aos mesmos fenômenos que os dados lineares, como por exemplo, ocorrência de pontos discrepantes. A ocorrência de pontos discrepantes em dados lineares tem sido amplamente pesquisada envolvendo os mais diversos modelos, como pode ser observado em Silva e Cirillo (2009) em estudo sobre estimadores robustos em modelos binomiais sob contaminação com excesso de zeros, fonte causadora de pontos discrepantes. Em se tratando de ocorrência de pontos discrepantes em dados circulares, alguns métodos de análise são tratados por Ibrahim (2013) e Collet (1980), propondo testes para a identificação de observações discrepantes em dados provenientes da distribuição Von-Mises, que é apropriada a dados circulares (MARDIA, 1972).

Particularmente em casos multivariados, Filzmoser et al.(2008) propuseram um método computacional para se identificar tais pontos em altas dimensões. Computacionalmente podem-se obter amostras multivariadas com *pontos discrepantes*, através de variáveis com distribuição normal multivariada contaminada (JOHNSON, 1987), sendo necessário para tal estabelecer diferentes vetores de médias e matrizes de correlação ou covariâncias para as variáveis. Um estudo sobre matrizes de covariâncias e utilização de diferentes graus de correlação entre as variáveis pode ser encontrado em Cirillo et al. (2006).

De acordo com Diggle et al. (2002) e Diggle (1988), uma matriz de correlação deve apresentar flexibilidade para englobar diferentes variações entre as variáveis, tais como: fontes de variação devida aos efeitos aleatórios; variação explicada por correlação serial, em que se espera que as observações mais próximas sejam fortemente correlacionadas e ainda variação devido a erros de medida. Para tal, no processo de simulação foram utilizadas duas estruturas que assumem correlações diferentes entre as variáveis e uma estrutura que assume a mesma correlação entre as variáveis, afim de que se possa observar possíveis diferenças nos ângulos em tais situações.

Dessa forma, este trabalho tem como objetivo avaliar computacionalmente a influência de diferentes estruturas de correlação na distribuição dos ângulos formados entre os Componentes Principais e Interpretáveis provenientes de dados na ausência e presença de pontos discrepantes. Foram consideradas ainda, variações nos coeficientes de correlação nas probabilidades de mistura utilizadas na contaminação e também diferentes tamanhos amostrais. Outro aspecto a ser observado é a difusão da estatística circular para obtenção dos valores esperados dos ângulos obtidos entre os componentes no processo de simulação, bem como meios de representação gráfica desses ângulos.

2. Conceitos preliminares

Para um melhor entendimento e compreensão do trabalho, serão apresentados nesta seção alguns conceitos e notações referentes à obtenção da direção média angular, distribuição normal assimétrica multivariada, mistura de distribuições e Componentes Interpretáveis. Estes conceitos são essenciais para a estruturação do processo de simulação.

2.1. Esperança matemática paradados angulares

A representatividade dos CI é avaliada mediante o valor do ângulo formado entre os CI e CP e assim a definição de esperança matemática é necessária pela necessidade de sua utilização na obtenção dos valores esperados dos ângulos através do processo de simulação.

Para justificativa de utilização do método, considere como exemplo três direções dadas pelos ângulos θ_1=80°, θ_2=350° e θ_3=50°. Em um círculo trigonométrico é fácil visualizar que o ângulo médio assuma um valor entre 0° e 50°. Porém, ao se calcular a média aritmética $(\theta_1+\theta_2+\theta_3)/3$ obtém-se o valor 160°, que não corresponde a situação. É necessária, então, a utilização de conceitos de estatística circular para obtenção do valor esperado.

Seja uma amostra circular $\theta_1,\ldots,\theta_p$ associada aos vetores unitários correspondentes $\overrightarrow{OP}_1,\overrightarrow{OP}_2,\ldots,\overrightarrow{OP}_p$, partindo da origem do círculo até dado ponto *P*. Na ocorrência de vários vetores, algebricamente calculam-se as médias com base nas coordenadas do sistema, utilizando-se das seguintes equações (ABUZAID et al., 2009):

$$\bar{\theta}=\begin{cases}\tan^{-1}\left(\dfrac{S}{C}\right) & \text{se } S>0, C>0\\ 180^{\circ}+\tan^{-1}\left(\dfrac{S}{C}\right) & \text{se } C<0\\ 360^{\circ}+\tan^{-1}\left(\dfrac{S}{C}\right) & \text{se } S<0, C>0\end{cases}. \quad (1)$$

Em que:

$$\sum_{i=1}^{n}\cos(\theta_i)=C \text{ e } \sum_{i=1}^{n}sen(\theta_i)=S \cdot \quad (2)$$

2.2. Distribuição normal assimétrica multivariada

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:

$$f_p(\mathbf{z})=2\varphi_p(\mathbf{z},\boldsymbol{\Sigma})\Phi(\boldsymbol{\alpha}^T\mathbf{z}),\ \text{com } \mathbf{z}\in \mathrm{R}^p, \quad (3)$$

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.

2.3. Mistura de distribuições

Um modelo de mistura é importante na geração de amostras com pontos discrepantes e pode ser descrito como:

$$f(\mathbf{x})=(1-\gamma)f_1(\mathbf{x})+\gamma f_2(\mathbf{x}), \quad (4)$$

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 $f_2(\mathbf{x})$, em que $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ representam diferentes distribuições. Dessa forma, tem-se um modelo composto por observações predominantes de uma dada distribuição $f_1(\mathbf{x})$ e alguns pontos pertencentes a uma segunda distribuição $f_2(\mathbf{x})$.

2.4. Componentes interpretáveis – restrição de homogeneidade

Considerando um primeiro CP $\mathbf{e}_1^T\mathbf{X}=e_{11}x_1+e_{12}x_2+e_{13}x_3$, o CI associado é dado por $\boldsymbol{\alpha}_1^T\mathbf{X}=\alpha_{11}x_1+\alpha_{12}x_2+\alpha_{13}x_3$ Os CI são chamados de $\boldsymbol{\alpha}_i$, $i=1,\ldots,p$. Ao passo que os coeficientes $\mathbf{e_i}$ podem assumir valores dispersos, $\boldsymbol{\alpha}_i$ assumem poucos e distintos valores, como 0 ou $\pm c$, considerando um valor de c que permita a restrição $\boldsymbol{\alpha}_i^T\boldsymbol{\alpha}=1$. Assim a i-ésima direção de $\boldsymbol{\alpha}_i$ pode ser mais interpretável. Para isso, a restrição de homogeneidade fixa $\pm c$ como a quantidade $\pm\dfrac{1}{\sqrt{k}}$,com $k=1,\ldots,p$ variáveis, em que *k* é uma constante normalizadora (CHIPMAN e GU, 2005). O ângulo entre $\boldsymbol{\alpha}_i$ e $\mathbf{e_i}$ deve ser o menor possível e é obtido pelo $\cos^{-1}(\mathbf{e}_i^T\boldsymbol{\alpha}_i)$.

Como um exemplo de ilustração, suponha um vetor de coeficientes $\mathbf{e}_1$ de um primeiro CP, com p =4. Seja então, $\mathbf{e}_1=[0{,}41\ \ -0{,}03\ \ -0{,}42\ \ 0{,}81]^T$. 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 0 e $\pm\dfrac{1}{\sqrt{k}}$, obedecendo $\boldsymbol{\alpha}_i^T\boldsymbol{\alpha}=1$ tem-se as opções 0, $\pm\dfrac{1}{\sqrt{1}}$, $\pm\dfrac{1}{\sqrt{2}}$, $\pm\dfrac{1}{\sqrt{3}}$ ou $\pm\dfrac{1}{\sqrt{4}}$.

Alguns possíveis candidatos são

$$\boldsymbol{\alpha}_1=\frac{[0\ \ 0\ \ 0\ \ 1]^T}{\sqrt{1}},\ \boldsymbol{\alpha}_1=\frac{[0\ \ 0\ \ -1\ \ -1]^T}{\sqrt{2}},$$

$$\boldsymbol{\alpha}_1=\frac{[1\ \ 0\ \ -1\ \ 1]^T}{\sqrt{3}} \text{ e } \boldsymbol{\alpha}_1=\frac{[1\ \ -1\ \ -1\ \ 1]^T}{\sqrt{4}}.$$

Neste caso, o $\boldsymbol{\alpha}_i$ mais próximo de $\mathbf{e}_1 = \begin{bmatrix} 0,41 & -0,03 & -0,42 & 0,81 \end{bmatrix}^T$ é

$\boldsymbol{\alpha}_1 = \dfrac{\begin{bmatrix} 1 & 0 & -1 & 1 \end{bmatrix}^T}{\sqrt{3}}$, com um ângulo de 18,8 graus.

Observa-se ainda que existe uma correspondência de sinais de elementos não próximos a zero. Ressalta-se que como a constante é inerente a todos os termos, pode ser omitida para efeito de comparação dos coeficientes.

Um exemplo prático de aplicação dos CI foi apresentado em Chipman e Gu (2005), em um estudo sobre 17 características físicas de carros vendidos nos Estados Unidos em 1993. Para os CI apresentados no estudo, ressaltou-se que o primeiro componente referiu-se ao tamanho dos carros, com coeficientes positivos relacionados positivamente com o tamanho dos carros e coeficientes negativos seguindo a lógica contrária. O segundo CI pôde ser interpretado como um contraste entre carros baratos, fracos e grandes versus carros caros, potentes e pequenos.

3. Metodologia

Para o processo de obtenção das amostras com pontos discrepantes, foi utilizado no processo de simulação o modelo de mistura $f(\mathbf{x}) = (1-\gamma) f_1(\mathbf{x}) + \gamma f_2(\mathbf{x})$ de onde foram compostas as amostras sob contaminação que formaram um vetor $\mathbf{X}$ de $p = 3$ variáveis aleatórias. Foram assumidos dois valores de probabilidade de mistura, γ =0,05 e γ =0,30 obtendo assim amostras com 5% de contaminação e 30% de contaminação. Tem-se ainda $f_1(\mathbf{x})$ como uma distribuição de referência, normal multivariada em que $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ e $f_2(\mathbf{x})$ como uma distribuição normal multivariada assimétrica tal que:

$$\mathbf{X} \sim NA_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = \begin{bmatrix} -20 & -20 & -20 \end{bmatrix}^T . \quad (5)$$

Para as distribuições utilizadas na obtenção das amostras foi utilizado um vetor de médias $\boldsymbol{\mu} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$ e para as matrizes de covariâncias $\boldsymbol{\Sigma}$ de ordem *p*=3 foram consideradas três diferentes estruturas de correlação: autoregressiva de ordem 1, AR(1), Simetria Composta (CS) e Toeplitz, nomeadas $\mathbf{R}_1$, $\mathbf{R}_2$ e $\mathbf{R}_3$, respectivamente. As estruturas de correlação são apresentadas abaixo conforme Littell et al. (2000):

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

$$\mathbf{R}_2 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \quad \text{e}$$

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

Foram assumidos como coeficientes de correlação $\rho = 0,5$ e $\rho = 0,8$ para $\mathbf{R}_1$ e $\mathbf{R}_2$. Para $\mathbf{R}_3$ utilizou-se um vetor de correlações $\boldsymbol{\rho}_1 = \begin{bmatrix} 0,9 & 0,8 & 0,7 \end{bmatrix}^T$ e um vetor $\boldsymbol{\rho}_2 = \begin{bmatrix} 0,6 & 0,5 & 0,4 \end{bmatrix}^T$ 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).

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.

Salienta-se que para gerar amostras sem presença de pontos discrepantes caracterizados pela mistura basta declarar $\gamma = 0$ no processo de simulação, garantindo somente a ocorrência de $f_1(\mathbf{x})$. Deste modo, foram comparados os ângulos formados em amostras com e sem contaminação provenientes unicamente de uma distribuição normal multivariada. A comparação é feita sempre entre o mesmo ρ e a mesma estrutura de correlação.

A partir do vetor $\mathbf{X}$ de $p = 3$ variáveis aleatórias, foram obtidos os três CP através das combinações lineares não correlacionadas dos elementos de $\mathbf{X}$. A transformação linear utilizada é dada por $\mathbf{e}_i^T \mathbf{X}$, $i = 1, 2, 3$, tal que:

$$\begin{aligned} \mathbf{e}_1^T \mathbf{X} &= e_{11} x_1 + e_{12} x_2 + e_{13} x_3 \\ \mathbf{e}_2^T \mathbf{X} &= e_{21} x_1 + e_{22} x_2 + e_{23} x_3 \\ \mathbf{e}_3^T \mathbf{X} &= e_{31} x_1 + e_{32} x_2 + e_{33} x_3 \end{aligned} \quad . \quad (6)$$

Partindo das equações apresentadas em (6) para o cálculo dos CP, procedeu-se então a obtenção dos CI $\boldsymbol{\alpha}_i$, $i = 1, \ldots, p$, considerando a restrição de homogeneidade em que $\boldsymbol{\alpha}_i$ assumiu os valores $\pm c$,

sendo $\pm c$ proposto como $\pm\frac{1}{\sqrt{k}}$, $k=1,2,...,p$.

Esse processo reduziu ainda mais a quantidade dos coeficientes a serem assumidos pelos CI, visto que não foi utilizado o valor zero. Dessa forma foram obtidos os CI:

$$\begin{aligned}\boldsymbol{\alpha}_1^T\mathbf{X}&=\alpha_{11}x_1+\alpha_{12}x_2+\alpha_{13}x_3\\ \boldsymbol{\alpha}_2^T\mathbf{X}&=\alpha_{21}x_1+\alpha_{22}x_2+\alpha_{23}x_3\\ \boldsymbol{\alpha}_3^T\mathbf{X}&=\alpha_{31}x_1+\alpha_{32}x_2+\alpha_{33}x_3\end{aligned}. \quad (7)$$

Para avaliar os CI em relação aos CP, deve-se procurar o ângulo mínimo entre cada um dos dois componentes. Desta forma os CP podem ser substituídos pelos CI fornecendo um número menor de coeficientes.

O algoritmo de procura pelos CI executa os seguintes passos:

1)Fixe os elementos de $\boldsymbol{\alpha}_i$ em $\pm\frac{1}{\sqrt{k}}$, $k=1,2,...,p$, assim para p=3, $\alpha_i=\pm\frac{1}{\sqrt{3}}$;

2)Verifique a correspondência dos sinais dos coeficientes α_{ij} com os coeficientes e_{ij};

3)Considere a restrição $\boldsymbol{\alpha}_i^T\boldsymbol{\alpha}_i=1$;

4)Obtenha o ângulo entre $\boldsymbol{\alpha}_i$ e $\mathbf{e}_i$ através do $\cos^{-1}\left(\mathbf{e}_i^T\boldsymbol{\alpha}_i\right)$.

Desse modo, a avaliação dos ângulos foi feita através dos valores médios angulares obtidos em todas as replicações de Monte Carlo, em que cada replicação foi obtido um valor $a=\cos^{-1}\left(\mathbf{e}_i^T\boldsymbol{\alpha}_i\right)$ correspondente ao ângulo.

Uma rotina computacional para obtenção dos CI foi desenvolvida utilizando o software R (R DEVELOPMENT CORE TEAM, 2013). Ao todo foram realizadas 2000 simulações de Monte Carlo para cada experimento. Os resultados computados são os ângulos formados entre os componentes, chamados $\overline{a}_1$ (direção média entre os primeiros CP e primeiros CI), $\overline{a}_2$ (direção média entre os segundos componentes) e $\overline{a}_3$ (direção média entre os terceiros componentes). Para a obtenção dos valores médios angulares resultantes da simulação, foi utilizado o conceito de direção média, a fim de obter uma estimativa de Monte Carlo da direção média. Ainda, para a representação dos ângulos estimados foram utilizados os gráficos circulares apresentados a seguir.

4 Resultados e discussão

Os resultados descritos nas Tabelas 1 a 4 correspondem aos valores esperados angulares entre os eixos formados pelos CP e CI obtidos por meio da distribuição empírica resultante das realizações de Monte Carlo.

Tabela 1. Média dos ângulos em graus considerando a distribuição Normal Multivariada para $\rho=0,50$

Estrutura de Correlação	n	$\overline{a}_1$	$\overline{a}_2$	$\overline{a}_3$
AR(1)	50	5,91	31,00	18,25
	100	4,93	31,04	17,97
	200	4,36	32,85	16,32
CS	50	2,55	27,55	23,10
	100	1,87	26,47	24,51
	200	1,38	25,85	25,09
		$\boldsymbol{\rho}_2=[0,6\ 0,5\ 0,4]$		
Toeplitz	50	2,01	30,56	19,48
	100	2,04	32,36	18,22
	200	1,95	33,12	17,94

Tabela 2. Média dos ângulos em graus considerando a distribuição Normal Multivariada para $\rho = 0,80$

Estrutura de Correlação	n	$\overline{a}_1$	$\overline{a}_2$	$\overline{a}_3$
AR(1)	50	1,95	30,59	19,74
	100	1,97	31,38	19,02
	200	1,96	33,07	17,91
CS	50	0,62	25,19	23,80
	100	0,43	25,35	24,32
	200	0,32	25,22	23,76
		$\boldsymbol{\rho}_1 = [0,9\ 0,8\ 0,7]$		
Toeplitz	50	1,23	30,58	19,87
	100	1,21	32,06	19,07
	200	1,15	33,02	18,78

Os resultados encontrados nas Tabelas 1 e 2 referem-se aos ângulos considerando amostras sem contaminação e evidenciaram que os valores esperados dos ângulos entre os eixos dos Componentes Principais e Interpretáveis são menos influenciados pelo efeito do tamanho amostral, visto que os ângulos não apresentam grandes variações em relação a *n*.

Considerando a estrutura CS e $\rho = 0,50$ observou-se que os valores das médias angulares entre os primeiros componentes ($\overline{a}_1$) foram 2,55, 1,87 e 1,38, para *n* igual a 50,100 e 200, respectivamente. Já para o caso de $\rho = 0,80$, os valores foram 0,62, 0,43 e 0,32 nas mesmas situações de tamanho amostral. Com isso, 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 as demais estruturas. A estrutura AR (1) por exemplo, apresentou as maiores médias angulares para os primeiros componentes, apresentando seu maior valor em *n* igual a 50 (5,91) para o caso de $\rho = 0,50$. Assim, a estrutura CS foi a que apresentou os menores ângulos entre os primeiros componentes.

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 CP 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 não muito dispersos para o primeiro componente.

Contextualizando com os resultados observados quanto a estrutura CS, a afirmação de Morrison (1990) justifica os valores encontrados para $\overline{a}_1$ na estrutura CS. Se os coeficientes do primeiro CP são aproximadamente os mesmos, $\boldsymbol{\alpha}_i$ apresentará valores bem próximos de $\mathbf{e}_1$, já que a restrição de homogeneidade é obtida por $\pm\frac{1}{\sqrt{k}}$. Isso faz com que $\left(\mathbf{e}_i^T\boldsymbol{\alpha}_i\right)$ seja bem próximo de 1, minimizando o valor de $\cos^{-1}\left(\mathbf{e}_i^T\boldsymbol{\alpha}_i\right)$. Quanto menores os ângulos entre os CP e CI, melhor a representatividade dos CI.

Quanto a estrutura Toeplitz, esta apresenta ângulos menores que AR(1), como pode ser observado em $\overline{a}_1$ que apresentou valores esperados correspondentes a 2,01, 2,04 e 1,95 para $\rho = 0,50$ e 1,23, 1,21 e 1,15 para $\rho = 0,80$.

De forma geral, para os primeiros componentes, CS, apresenta menores ângulos seguido de Toeplitz e AR(1). Situação que não é observada para os terceiros componentes que apresentaram em CS, os valores de 23,10, 24,51 e 25,09 para $\rho = 0,50$ e 23,80, 24,32 e 23,76 para $\rho = 0,80$.

Tabela 3 Média dos ângulos em graus considerando a distribuição Normal Assimétrica com $\gamma = 0,05$, $\gamma = 0,30$ e $\rho = 0,50$.

Estrutura de Correlação	γ	n	$\overline{a}_1$	$\overline{a}_2$	$\overline{a}_3$
AR(1)	0,05	50	5,98	31,03	19,05
		100	5,94	32,06	16,91
		200	5,15	32,84	15,82
AR(1)	0,30	50	6,37	30,93	19,12
		100	5,79	31,77	17,31
		200	5,42	32,03	16,81
CS	0,05	50	3,08	27,27	24,16
		100	1,56	26,67	24,02
		200	1,22	25,76	25,20
CS	0,30	50	2,85	27,41	23,66
		100	1,85	25,95	24,94
		200	1,22	26,67	23,87
				$\tilde{\mathbf{n}}_2 = [0,6\ 0,5\ 0,4]$	
Toeplitz	0,05	50	1,92	31,42	18,86
		100	1,95	32,58	18,16
		200	1,68	33,06	18,16
Toeplitz	0,30	50	2,20	30,52	19,22
		100	1,88	32,93	18,06
		200	1,81	33,34	17,95

Tabela 4 Média dos ângulos em graus considerando a distribuição Normal Assimétrica com $\gamma = 0,05$, $\gamma = 0,30$ e $\rho = 0,80$

Estrutura de Correlação	γ	n	$\overline{a}_1$	$\overline{a}_2$	$\overline{a}_3$
AR(1)	0,05	50	1,91	30,38	19,79
		100	1,95	31,84	18,55
		200	2,02	32,71	18,01
AR(1)	0,30	50	2,07	30,19	19,77
		100	2,29	31,96	18,34
		200	2,07	32,41	18,17
CS	0,05	50	0,72	25,28	25,46
		100	0,46	25,69	25,00
		200	0,29	25,12	25,53
CS	0,30	50	0,65	25,71	25,05
		100	0,52	25,90	24,84
		200	0,34	25,03	25,67
				$\tilde{\mathbf{n}}_1 = [0,9\ 0,8\ 0,7]$	
Toeplitz	0,05	50	1,21	30,78	19,81
		100	1,20	32,19	18,99
		200	1,18	33,19	18,68
Toeplitz	0,30	50	1,42	31,06	19,50
		100	1,22	32,96	18,65
		200	1,17	32,50	18,83

Em concordância com os resultados obtidos em amostras não contaminadas (Tabelas 1 e 2), os resultados descritos nas Tabelas 3 e 4 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 $\overline{a}_1$, formado entre os eixos representados pelo primeiro CI e o primeiro CP, assumiu menor valor quando as variáveis foram altamente correlacionadas ($\rho = 0,80$, Tabela 4), assumindo os valores 1,91, 1,95 e 2,02. Ao comparar os ângulos entre os eixos formados pelos segundos CP e CI, com suas respectivas parametrizações e nos dois coeficientes de correlação, notou-se uma variação mínima entre as médias angulares.

Em se tratando da estrutura CS, nas situações de presença de pontos discrepantes (Tabelas 3 e 4) as menores médias angulares foram identificadas nos primeiros componentes, com valores $\overline{a}_1$ 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 mesmo em dimensões maiores, porem não consideraram pontos discrepantes ou diferentes estruturas de correlação em seus estudos.

Ainda sobre a estrutura CS, a média angular manteve-se inferior para os segundos componentes ($\overline{a}_2$). Para o terceiro componente, apresentou elevação nas médias em relação à estrutura AR(1) e Toeplitz em situações de $\rho = 0,80$.

Ao assumir a estrutura de correlação Toeplitz nas Tabelas 3 e 4, observou-se que os valores angulares esperados para o primeiro CI ($\overline{a}_1$), 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 com as demais estruturas. O terceiro componente apresentou valores esperados menores, próximos a 19°, nas estruturas AR(1) e Toeplitz em ambos os graus de correlação.

Quanto a variação de γ, verificam-se mínimas variações nos valores médios dos ângulos principalmente nos ângulos $\overline{a}_1$ da estrutura AR(1). As variações são pequenas, não excedendo 1°. As maiores variações continuam acontecendo na mudança de $\rho = 0,80$ para $\rho = 0,50$, no caso AR(1) e CS.

De forma geral, os CI podem ser aplicados a situações de ocorrência de pontos discrepantes. Em se tratando de CP, a maior explicação da variabilidade ocorre no primeiro componente e como apresentado nas tabelas, os ângulos entre os componentes sempre foram baixos em tais situações, especialmente na estrutura CS. Assim, em casos onde há dificuldades de interpretabilidade de CP, a utilização dos CI é de importante aplicação.

A fim de se obter uma melhor visualização da ocorrência dos ângulos, as Figuras 1 a 6 apresentam a distribuição dos ângulos entre os componentes no gráfico circular, considerando conjuntamente os dois coeficientes de correlação, $\rho = 0,80$ e $\rho = 0,50$.

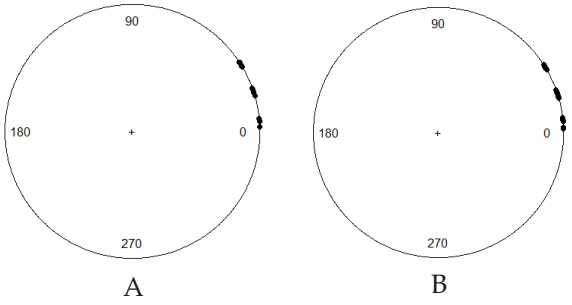

Figura 1 Gráfico circular considerando AR(1) sem contaminação (A) e sob contaminação (B)

Observa-se na Figura 1 (A e B) a presença de três agrupamentos distintos. Os grupos são formados pelas médias angulares observadas para os três CP em relação aos respectivos CI, conforme pode ser observado nas Tabelas 1 a 4.

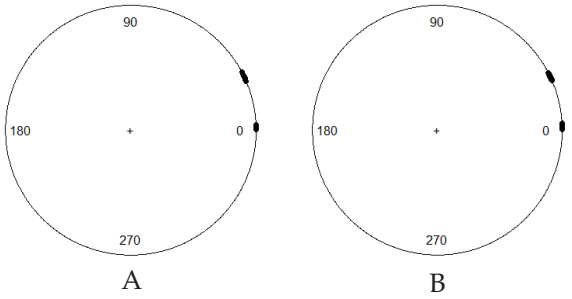

Figura 2 Gráfico circular considerando CS sem contaminação (A) e sob contaminação (B)

No que se refere à estrutura de correlação de Simetria Composta (CS), encontram-se, representados na Figura 2 (A e B), os valores médios angulares para tal estrutura. Verifica-se a presença de dois grupos distintos de pontos, que também correspondem a dispersão no círculo, dos ângulos $\overline{a}_1$, $\overline{a}_2$ e $\overline{a}_3$. A estrutura CS, apresentou valores médios angulares menores que 1° para o primeiro componente e em torno de 25° para o segundo e terceiro componentes, o que caracteriza a visualização de somente dois grupos distintos no diagrama circular.

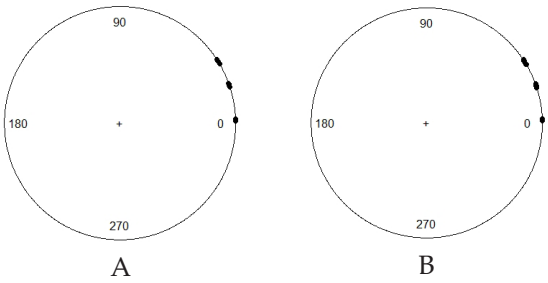

Figura 3 Gráfico circular considerando Toeplitz sem contaminação (A) e sob contaminação (B)

Em relação à estrutura de correlação Toeplitz, os ângulos e distâncias encontram-se representados na Figura 3 (A e B). Na figuram, visualizam-se 3 agrupamentos de valores médios angulares, como na estrutura AR(1). Os valores de $\overline{a}_1$, $\overline{a}_2$ e $\overline{a}_3$, estão em torno de 1°, 20° e 30°, respectivamente, caracterizando os 3 agrupamentos.

A distribuição dos ângulos foi diretamente afetada pela escolha das diferentes estruturas de correlação, que apresentam valores diferenciados entre os ângulos formados. Assim, a forma de estruturação dos dados influencia nos valores dos ângulos. Por outro lado, os primeiros CP e CI sempre apresentam ângulos mínimos em qualquer estrutura. Conjuntamente à estrutura de correlação, situações de maior correlação entre as variáveis fornecem os menores ângulos entre os componentes.

5. Conclusões

O trabalho apresentou um estudo da influência da estrutura de correlação em situações de ocorrência de pontos discrepantes na obtenção de CI. Tais pontos foram inseridos em amostras simuladas através de um modelo de mistura de distribuições. Para tal, foram consideradas diferentes estruturas de correlação e níveis de contaminação, englobando distintas situações que possam vir a ocorrer em dados. Um importante resultado verificado é que os pontos discrepantes não apresentam efeitos perturbadores nos ângulos obtidos entre os CI e os CP. Assim, os CI são robustos a ocorrência de pontos discrepantes nos dados, visto que os ângulos para os primeiros componentes não se alteram de forma significativa em situações na presença e ausência de pontos discrepantes.

Quanto às estruturas de correlação, em situações onde as variáveis possuem a mesma correlação uma com as outras (CS), os ângulos assumem os menores valores nos dois primeiros componentes. Assim, em análise de dados onde a correlação entre as variáveis é homogênea, o método dos CI pode ser aplicado com bom desempenho mesmo em amostras com pontos discrepantes.

Por fim, ressalta-se a disseminação de métodos de estatística circular, que é de suma importância para obtenção de medidas corretas em situações de dados angulares, garantindo inferências e interpretações corretas de resultados.

6. Referências

ABUZAID, A. H.; MOHAMED, I. B.; HUSSIN, A.G. A New Test of Discordancy in Circular Data. **Communications in Statistics - Simulation and Computation.** v.38, n.4, p. 682-691, 2009.

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

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

CIRILLO, M. A. ; FERREIRA, D. F. ; SAFADI, T. . Estudo do poder e tamanho do teste de Levene multivariado via simulação Monte Carlo e bootstrap. **Acta Scientiarum. Technology** , v. 28, p. 105-112, 2006.

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

DIGGLE, P. J. An Approach to the Analysis of Repeated Measurements. **Biometrics,** v.44, n.4, p. 959-971, 1988.

DIGGLE, P. J.; HEAGERTY, P.; LIANG, K. Y.; ZEGER, S. L. **Analysis of longitudinal data**, 2 ed, Oxford: Oxford University Press, 2002

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

FILZMOSER, P.; MARONNA, R.; WERNER, M. Outlier identification in high dimensions. **Computational Statistics and Data Analysis**, v.52, n.3, p.1694–1711, Jan. 2008.

FISHER, N. I. **Statistical Analysis of Circular Data**. 1. ed. Cambridge: University Press, 1993. 296 p.

IBRAHIM, S.; RAMBLI, A.; HUSSIN, A. G.; MOHAMED, I. Outlier Detection in a Circular Regression Model Using COVRATIO Statistic. **Communications in Statistics - Simulation and Computation**, v. 42, n. 10, p. 2272-2280, 2013

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, N.J.: Pearson Prentice Hall, 2007, 773 p.

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

MARDIA, K.V.; **Statistics of Directional Data**, Academic Press, 1972.

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

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

SILVA, A.M.; CIRILLO, M.A. Estudo por simulação Monte Carlo de um estimador robusto utilizado na inferência de um modelo binomial contaminado. **Acta Scientiarum**. Technology , v. 32, p. 303-307, 2010.

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