Revista da Estatística UFOP, Vol III(2), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS

AVALIAÇÃO MONTE CARLO DE MÉTRICAS PARA FALTA DE AJUSTE EM ANÁLISE DE ARQUÉTIPOS

José Márcio Martins Júnior[1,4], Elcio do Nascimento Chagas[2,4], Denismar Alves Nogueira[1,4], Daniel Furtado Ferreira[3,4], Eric Batista Ferreira[1,4]

Resumo: *A Análise de Arquétipos é uma técnica multivariada que busca reduzir a dimensão de dados por meio de combinações lineares de arquétipos, que são selecionados pela minimização de alguma métrica que represente o erro cometido ao se reconstruir os dados. Este trabalho tem como objetivo, utilizando simulação Monte Carlo, avaliar a qualidade de ajuste da análise de arquétipos com as diferentes métricas citadas na literatura (norma quadrática, norma de Frobenius e normal espectral) propostas (determinante e soma de quadrados e produtos de resíduos). Os resultados mostram que as métricas estudadas apresentam o mesmo comportamento: à medida que a correlação aumenta a qualidade de ajuste melhora, quando a perturbação causada nos erros aumenta, a qualidade do ajuste piora. Assim conclui-se que as métricas estudadas são equivalentes. Dessa forma, este trabalho indica o uso da mais simples, ou seja, a soma de quadrados de resíduos.*

Palavras-chave: *Arquétipos, análise multivariada, simulação Monte Carlo, soma de quadrado de resíduos.*

Abstract: *Archetypal analysis represents reduces data dimensionality by finding archetypes able to reconstruct the original data, minimizing an error metric. This work aims to evaluate different metrics. Evaluated metrics showed similar behavior, what makes one recommend the simplest: residual sum of squares.*

Keywords: *Archetypes, multivariate analysis, Monte Carlo simulation, residual sum of squares.*

1 Introdução

A Análise de Arquétipos (AA) é uma técnica multivariada utilizada em diversas áreas do conhecimento, como medicina, economia, marketing, aprendizado de máquinas, reconhecimento de padrões, astrofísica e psicologia (RIEDESEL, 2014; THOGERSEN et al., 2014; MORUP; HANSEN, 2011; CHAN et al., 2002). Foi introduzida por Cutler e Breiman (1994) e tem como propósito simplificar a estrutura de covariâncias, sendo utilizada para reduzir a dimensão de dados por meio de combinações lineares dos seus elementos mais representativos. Os arquétipos

[1]ICEX - UNIFAL-MG. e-mail: *jmmjunifal@gmail.com, denismar.nogueira@unifal-mg.edu.br, eric.ferreira@unifal-mg.edu.br.*
[2]IFES - *Campus* de Alegre. e-mail: *enchagas@ifes.edu.br.*
[3]DEX - UFLA. e-mail: *danielff@dex.ufla.br.*
[4]Agradecimentos: a FAPEMIG, a CAPES e ao CNPq pelo apoio financeiro.

são selecionados pela minimização da soma de quadrado de resíduos (SQR) ao representar cada observação como uma combinação dos arquétipos ou como um dos arquétipos (também denominado arquétipo puro) e estão na fronteira do fecho convexo dos dados. Portanto, são geralmente valores extremos que melhor representam os dados.

O cálculo para encontrar os arquétipos é um problema de quadrados mínimos não-linear, que pode ser resolvido por um algoritmo de otimização iterativo que converge em todos os casos, mas não necessariamente para o minimo global. Por isso, o algoritmo deve ser iniciado várias vezes com arquétipos iniciais diferentes. A cada passo o algoritmo diminui a SQR entre a combinação linear dos arquétipos e o verdadeiro valor dos dados. O algoritmo deve parar quando a SQR for um valor suficientemente pequeno (CUTLER; BREIMAN, 1994).

Para dados multivariados $(\mathbf{x}_i, i = 1, \ldots, n)$ em que cada $\mathbf{x}_i$ é um vetor p-dimensional $\mathbf{x}_i = (\mathbf{x}_{1i}, \ldots, \mathbf{x}_{pi})'$, o padrão arquétipo de uma massa de dados caracteriza o problema de encontrar vetores p-dimensionais $\mathbf{z}_1, \ldots, \mathbf{z}_k$ com $1 < k < N$, sendo N o número de elementos na fronteira (BAUCKHAGE; THURAU, 2012).

$$\mathbf{z}_j = \sum_{i=1}^{n} \mathbf{x}_i b_{ij} \tag{1}$$

em que $j = 1, \ldots, k$ e os coeficientes $b_{ij} \geq 0$ e $\sum_{i=1}^{n} b_{ij} = 1$. Assim, para uma dada escolha de arquétipos, AA minimiza

$$||\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{z}_j a_{ji}||^2 \tag{2}$$

Sabe-se que quanto maior o número de arquétipos selecionados menor é a SQR, pois menos informação é perdida, e por consequência, menor é a redução da dimensão dos dados. Então, fica a cargo do pesquisador decidir quantos arquétipos deve-se usar em um determinado conjunto de dados, desde que $1 < k < N$.

Para determinar os coeficientes a_{ji} que permitam que os dados $\mathbf{x}_i$ sejam bem representados pelos arquétipos, AA impõe a condição que $a_{ji} \geq 0$, de modo que cada elemento pertencente aos dados seja reescrito como combinação linear dos arquétipos para recompor as informações originais e $\sum_j a_{ji} = 1$.

A fim de selecionar qual será o melhor conjunto de arquétipos, a minimização da seguinte equação é que gera este conjunto de arquétipos

$$SQR = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{z}_j a_{ji}||^2 = \sum_{i=1}^{n} ||\mathbf{x}_i - \sum_{j=1}^{k} \sum_{l=1}^{n} \mathbf{x}_l b_{lj} a_{ji}||^2 \tag{3}$$

Bauckhage e Thurau (2012) descrevem (3) em forma matricial. Assim, o conjunto de dados $\mathbf{x}_i \in \mathbb{R}^p$ compõe uma matriz $\mathbf{X}_{(p \times n)}$ e os arquétipos $\mathbf{z}_j \in \mathbb{R}^p$ compõem uma matriz $\mathbf{Z}_{(p \times k)}$.

$$SQR = ||\mathbf{X} - \mathbf{ZA}||^2 = ||\mathbf{X} - \mathbf{XBA}||^2 \tag{4}$$

em que $A \in \mathbb{R}^{k \times n}$ e $B \in \mathbb{R}^{n \times k}$.

Outras métricas são propostas para o cálculo da soma de quadrados dos resíduos, como por exemplo pode-se citar Eugster e Leisch (2009), que utilizaram a Norma Espectral (NE) e D'esposito et al. (2012) que utilizaram a Norma de Frobenius (NF).

Assim, este trabalho tem como objetivo, utilizando estudos de simulação Monte Carlo, avaliar a qualidade de ajuste da análise de arquétipos com as diferentes métricas citadas na literatura: Norma quadrática, Norma de Frobenius e Normal Espectral. Objetivou-se também, propor a utilização de outras duas métricas: Determinante (DET) e Soma de Quadrados e Produtos de Resíduos (SQPR), pois também são medidas sumarizantes de matrizes, como variâncias generalizadas.

2 Material e métodos

Com o simples objetivo de identificar onde os arquétipos estão situados, uma simulação proposta por Cutler e Breiman (1994), foi reproduzida da seguinte forma: foi gerado uma amostra de tamanho 1000 de uma distribuição Normal bivariada com vetor de médias e matriz de correlação

$$\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0,8 \\ 0,8 & 1 \end{bmatrix}.$$

Em seguida, foram descartados todos os pontos com distância de Mahalanobis superiores a uma cota dada por uma elipse de 95% de confiança, ou seja, $d_M \geq \chi^2_2(0,95)$.

Posteriormente, para cada amostra foram ajustados 4 arquétipos e estes foram armazenados. O processo foi repetido 100 vezes.

Por fim, o todos os $4 \times 100 = 400$ arquétipos foram plotados para que o comportamento fronteiriço fosse destacado. Esse primeiro estudo de simulação foi realizado apenas para fins de compreensão e clareza do método.

O principal estudo de simulação deste trabalho foi feito para se avaliarem as métricas que indicam falta de ajuste. Para isso, desenvolveu-se uma rotina de simulação utilizando o software R (R CORE TEAM, 2013).

Uma matriz $\mathbf{M}$ foi composta por dois vetores linearmente independentes ($\mathbf{a_1}$ e $\mathbf{a_2}$) e por quatro vetores ($\mathbf{a_3}, \mathbf{a_4}, \mathbf{a_5}$ e $\mathbf{a_6}$) definidos como uma combinação linear dos dois primeiros vetores, com coeficientes p, q, r e s, respectivamente, pertencentes ao intervalo (0,1) e com a restrição de que o somatório fosse igual a um.

Os vetores $\mathbf{a_1}$ e $\mathbf{a_2}$ foram fixados na forma $\mathbf{a_1} = (1,2,3)$ e $\mathbf{a_2} = (7,7,8)$. Assumindo os coeficientes $p = 0,4$; $q = 0,1$; $r = 0,9$ e $s = 0,5$; foram obtidos os vetores $\mathbf{a_3} = (0,4 \times \mathbf{a_1} + 0,6 \times \mathbf{a_2})$, $\mathbf{a_4} = (0,1 \times \mathbf{a_1} + 0,9 \times \mathbf{a_2})$, $\mathbf{a_5} = (0,9 \times \mathbf{a_1} + 0,1 \times \mathbf{a_2})$ e $\mathbf{a_6} = (0,5 \times \mathbf{a_1} + 0,5 \times \mathbf{a_2})$, resultando em $\mathbf{M} = (\mathbf{a_1}\ \mathbf{a_2}\ \ldots\ \mathbf{a_6})'$.

Os dados foram simulados de uma distribuição normal tri-variada com vetor de médias nulo (sem perda de generalidade) e matriz de covariâncias equicorrelacionada, dada por

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \tag{5}$$

em que as variâncias (σ^2) foram fixadas em $0,1$ (pequena); $0,5$ (média); $1,0$ (grande) , as correlações (ρ) fixadas em $0; 0,25; 0,5; 0,75$ e $0,95$. As diferentes variâncias, representam diferentes graus de perturbação no ajuste. Por outro lado, estudar covariâncias crescentes tem como objetivo verificar se as métricas são influenciadas por multicolinearidade. O número de repetições Monte Carlo foi fixado em 1000, totalizando 15000 cenários simulados.

Depois de agregado o erro/perturbação nos vetores, o algoritmo capaz de encontrar os arquétipos foi executado utilizando a função `archetypes()` do pacote de mesmo nome do software R (R CORE TEAM, 2013). Uma das informações que é recuperada após a execução do algoritmo é a matriz de resíduos ($\mathbf{E}$), que nada mais é que a diferença entre os dados originais e os dados reconstruídos pelos arquétipos, ou seja, $\mathbf{E} = \mathbf{X} - \mathbf{XBA}$. Então é aplicada uma função sumarizante nessa matriz de resíduos. As métricas estudadas neste trabalho são apresentadas na Tabela 1.

3 Resultados e discussões

Os primeiros resultados referem-se à simulação ilustrativa proposta por Cutler e Breiman (1994), que foi reproduzida.

A Figura 1 contém a representação de uma amostra da normal correlacionada, ressaltando-se os pontos com distância de Mahalanobis menor ou igual a $\chi^2_2(0,95)$. De amostras como essa, os

Tabela 1: Métricas utilizadas no cálculo da falta de ajuste da análise de Arquétipos

Métricas	Notação	
Soma de Quadrados de Resíduos	$\|\|.\|\|^2$	$tr(\mathbf{EE}^{'})$
Norma de Frobenius	$\|\|.\|\|_F$	$\sqrt{tr(\mathbf{EE}^{'})}$
Norma Espectral	$\|\|.\|\|_2$	$\sqrt{\lambda_{max}(\mathbf{EE}^{'})}$
Soma de Quadrados e Produtos de Resíduos	$\|\|.\|\|^{QP}$	$\mathbf{1}^{'}\text{abs}(\mathbf{EE}^{'})\mathbf{1}$
Determinante	$\|.\|$	$\|\mathbf{EE}^{'}\|$

em que: $tr(.)$ é o traço, $\lambda_{max}(.)$ é o maior autovalor e $abs(.)$ é o valor absoluto dos elementos de uma matriz.

pontos externos à elipse foram desprezados. Nesse sentido, os arquétipos ajustados se prestariam a representar os os pontos pertencentes a esse fecho (fecho da elipse de 95% de confiança).

Após colecionados todas as amostras de 4 arquétipos, a Figura 2 mostra-os em posição próxima à fronteira da referida elipse. É por esse motivo a Análise de Arquétipos é dita eleger os pontos extremos, capazes de reconstruir todos os dados, ou seja, os arquétipos.

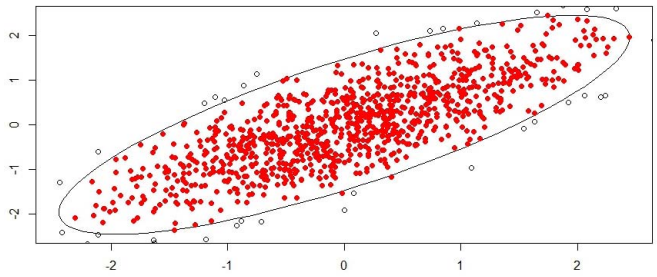

Figura 1: Elipse de confiança com 95% dos dados sorteados. Pontos preenchidos foram mantidos, pontos vazios foram descartados.

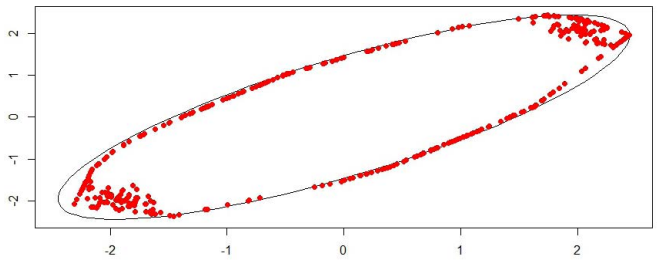

Figura 2: Arquétipos selecionados em 100 iterações da simulação.

Na Tabela 2, são apresentados os resultados obtidos no estudo de simulação.

Por meio dos resultados descritos, observa-se que todas as métricas estudadas apresentam o mesmo comportamento. À medida que a correlação ρ aumenta a qualidade de ajuste melhora, ou seja, o valor da diferença entre os dados originais e os dados recompostos diminui. Quando a perturbação σ^2 causada nos erros aumenta, a qualidade do ajuste piora, ou seja, o valor da diferença aumenta.

Usar o determinante como métrica não apresentou bons resultados, pois o resultado é sempre

Tabela 2: Valores médios da SQR e respectivos erros-padrão considerando variâncias (σ^2) 0,1 (pequena); 0,5 (média); 1,0 (grande) dos erros (ϵ), correlação (ρ) $0,0; 0,25; 0,5; 0,75; 0,95$ e número de repetições Monte Carlo fixado em 1000, para as diferentes métricas.

		Correlação				
	Variância	ρ =0,00	ρ =0,25	ρ =0,5	ρ =0,75	ρ =0,95
NQ	Pequena	0,15 ± 0,002	-	-	-	-
	Média	0,69 ± 0,011	0,36 ± 0,006	0,01 ± 0,000	-	-
	Grande	1,28 ± 0,021	0,97 ± 0,017	0,56 ± 0,009	0,36 ± 0,006	0,08 ± 0,001
SQPR	Pequena	1,24 ± 0,010	-	-	-	-
	Média	2,66 ± 0,021	1,91 ± 0,016	0,27 ± 0,003	-	-
	Grande	3,59 ± 0,030	3,11 ± 0,027	2,35 ± 0,020	1,88 ± 0,017	0,87 ± 0,008
Frob	Pequena	0,38 ± 0,003	-	-	-	-
	Média	0,81 ± 0,006	0,58 ± 0,005	0,08 ± 0,001	-	-
	Grande	1,10 ± 0,009	0,95 ± 0,008	0,72 ± 0,006	0,58 ± 0,005	0,27 ± 0,002
Espec	Pequena	0,33 ± 0,003	-	-	-	-
	Média	0,70 ± 0,006	0,50 ± 0,004	0,08 ± 0,001	-	-
	Grande	0,95 ± 0,008	0,82 ± 0,007	0,62 ± 0,006	0,50 ± 0,005	0,23 ± 0,002
Det	Pequena	$10^{-47} \pm 10^{-47}$	-	-	-	-
	Média	$10^{-42} \pm 10^{-43}$	$10^{-44} \pm 10^{-44}$	-	-	-
	Grande	$10^{-41} \pm 10^{-41}$	$10^{-42} \pm 10^{-42}$	$10^{-43} \pm 10^{-43}$	$10^{-44} \pm 10^{-44}$	$10^{-49} \pm 10^{-49}$

muito próximo de zero. Isso era esperado, pois ele é aplicado em uma matriz positiva semidefinida de soma de quadrados e produtos de resíduos.

Uma possível explicação para a qualidade do ajuste melhorar à medida que a correlação aumenta, é que fica mais fácil obter os coeficientes da mistura dos arquétipos, implicando em uma melhor adaptação dos arquétipos aos dados e consequentemente na diminuição do erro da diferença entre os dados originais e dos dados recuperados a partir dos arquétipos.

4 Conclusões

De acordo com os resultados apresentados, pode-se concluir que todas as métricas são equivalentes. Portanto, este trabalho indica o uso da mais simples, ou seja, a soma de quadrados de resíduos.

Referências

[1] BAUCKHAGE, C.; THURAU, C. Making Archetypal Analysis Practical. **Lecture Notes in Computer Science** v.5748, pp 272-281, 2009.

[2] CHAN, B. H. P.; MITCHELL, D. A.; CRAM, L. E. **Archetypal analysis of galaxy spectra**. Astrophysics Department, School of Physics, A28, University of Sydney, NSW 2006, Australia. 2002.

[3] CUTLER, A.; BREIMAN, L. Archetypal analysis. **Technometrics**, v.36, pages 338-347, 1994.

[4] D'ESPOSITO R. M., PALUMBO F. RAGOZINI. G. Interval Archetypes: A New Tool for Interval Data Analysis. **Statistical Analysis and Data Mining** 5(4):322-335, 2012.

[5] EUGSTER, M.J. A.; LEISCH, F. From Spider Man to Hero - Archetypal Analysis. **Journal of Statistical Software**, v.30, pages 1-23, 2009.

[6] MORUP M.; HANSEN, L. K. **Archetypal analysis for machine learning and data mining**. Section for Cognitive Systems, Technical University of Denmark, Richard Petersens Plads, bld321, 2800 Lyngby, Denmark. 2011

[7] R DEVELOPMENT CORE TEAM. **R**: A Language and Environment for Statistical Computing. Vienna, Austria, 2013. Disponível em:<http://www.R-project.org>. Acesso em 20 fev. 2014.

[8] RIEDESEL, P. **Archetypal Analysis in Marketing Research:** A New Way of Understanding Consumer Heterogeneity. Disponível em:<http://www.action-research.com/archtype.html.>. Acesso em 24 fev. 2014.

[9] THOGERSEN, J. C.; MORUP, M.; DAMKIÆR, S.; MOLIN, S.; JELSBAK, L. Archetypal analysis of diverse Pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. ***BMC Bioinformatics*** **2013**. Disponível em:<http://http://www.biomedcentral.com/1471-2105/14/279.>. Acesso em 24 fev. 2014.