Detecção de *outliers* multivariados

Josino José Barbosa
Programa de Pós-Graduação em Estatística Aplicada e Biometria, Universidade Federal de Viçosa,
36.570-000 - Viçosa - MG, Brasil
josinojba@gmail.com

Tiago Martins Pereira
Departamento de Estatística, Universidade Federal de Ouro Preto,
35.400-000 - Ouro Preto - MG, Brasil
tiago.martin@gmail.com

Fernando Luiz Pereira de Oliveira
Departamento de Estatística, Universidade Federal de Ouro Preto,
35.400-000 - Ouro Preto - MG, Brasil
fernandoluiz@iceb.ufop.br

RESUMO

A detecção de *outliers* desempenha um papel importante na análise estatística, pois tais observações podem conter informações importantes em relação aos dados. Em situações práticas, os próprios *outliers* são muitas vezes os pontos especiais de interesse e sua identificação pode ser o principal objetivo da investigação. Por isso, o objetivo desse trabalho é propor uma técnica de detecção de *outliers* multivariados, baseada em análise agrupamento e comparar essa técnica com o método de identificação de *outliers* via distância de Mahalanobis. Para geração dos dados utilizou-se simulação através do Método de Monte Carlo e a técnica de mistura de distribuições normais multivariadas. Os resultados apresentados nas simulações mostram que o método proposto foi superior ao método de Mahalanobis tanto para sensibilidade quanto para especificidade, ou seja, ele apresenta maior capacidade de diagnosticar corretamente os indivíduos *outliers* e os não *outliers*.

PALAVRAS CHAVE. ***Outlier*, Análise de agrupamento, Método de Monte Carlo.**

Estatística (EST), Análise Multivariada.

ABSTRACT

The detection of outliers plays an important role in statistical analysis, as such observations may contain important information regarding the data. In practical situations, the outliers themselves are often the special points of interest and their identification may be the main objective of the investigation. Therefore, the objective of this work is to propose a technique of detection of multivariate outliers, based on cluster analysis and to compare this technique with the method of identification of outliers via Mahalanobis distance. For data generation, the Monte Carlo method was used to simulate the multivariate normal distribution. The results presented in the simulations show that the proposed method was superior to the Mahalanobis method for both sensitivity and specificity, that is, it presents greater capacity to correctly diagnose outliers and non-outliers individuals.

KEYWORDS. Outliers, Cluster analysis, Monte Carlo Method.

Statistics, Multivariate Analysis.

1. Introdução

Outlier é uma observação, ou um subconjunto de observações, que parece ser inconsistente quando comparada ao restante do conjunto [Hawkins, 1980]. Segundo [Bamnett e Lewis, 1994], *outlier* é uma observação que desvia muito de outras observações e desperta suspeitas de que é gerada por um mecanismo diferente. Estas observações são também designadas por observações anormais, contaminantes, estranhas, extremas ou aberrantes.

Em se tratando do espaço multivariado, uma observação é considerada anormal se está muito distante das outras no espaço p-dimensional definido pelas variáveis. Uma observação pode não ser um *outlier* em nenhuma das variáveis originais estudadas isoladamente e ainda ser na análise multivariada, por não se conformar com a estrutura de correlação do restante dos dados [Jolliffe, 2002].

A identificação de *outliers* desempenha um papel importante na análise estatística. Se modelos estatísticos clássicos são cegamente aplicados a dados contendo valores atípicos, os resultados podem ser enganosos e decisões equivocadas podem ser tomadas. Além disso, em situações práticas, os próprios *outliers* são muitas vezes os pontos especiais de interesse e sua identificação pode ser o principal objetivo da investigação.

A detecção de *outliers* tem sido extensivamente utilizada em diversas aplicações. Segundo [Aggarwal, 2013], na maioria das aplicações, os dados são criados por um ou mais processos de produção. Quando o processo de geração se comporta de uma maneira incomum resulta na criação de *outliers*. Portanto, um *outlier* muitas vezes contém informações úteis sobre as características anormais dos sistemas e entidades, que impactam no processo de geração de dados. O reconhecimento de tais características incomuns fornece uma série de aplicações úteis. Alguns exemplos de aplicações são os seguintes: sistemas de detecção de intrusão, fraude de cartão de crédito, sensores de eventos e diagnóstico médico.

Em função de sua ampla gama de aplicações, muitas técnicas têm sido desenvolvidas para a detecção de *outliers*. [Veloso e Cirillo, 2016] propuseram uma técnica para detecção de *outliers* baseada em componentes principais com amostras corrigidas por distância do tipo qui-quadrado. [Critchley, 1985] discutiu a influência dos *outliers* pela Curva de Influência baseada na Influência Global, que envolve a deleção de algumas observações utilizando a técnica de componentes principais. Um método alternativo para avaliar o efeito local de pequenas perturbações nos dados foi proposto por [Cook, 1986], tendo por base a curvatura normal, estruturada na verossimilhança.

Buscando superar as limitações dos procedimentos clássicos na identificação de *outliers*, [Filzmoser et al., 2008] propuseram um método de fácil implementação computacional, capaz de identificar *outliers* em altas dimensões. Contudo, vale ressaltar que o método proposto por esses autores consiste na descrição de um procedimento no qual se aplicou uma reescalonagem dos dados por meio da mediana (med) e do Desvio Absoluto da Mediana (Median Absolute Deviation - MAD).

[Berton et al., 2010] apresenta o desenvolvimento de um método baseado em redes complexas para detecção de diferentes tipos de *outliers* que utiliza a caminhada aleatória e um índice de dissimilaridade. Já [Valadares et al., 2012] propuseram um trabalho que apresenta uma análise, via detecção de *outliers*, sobre dados multivariados proveniente de rede de sensores.

Diante da importância da identificação de *outliers*, o objetivo desse trabalho é propor uma técnica de detecção de *outliers* multivariados, baseada em análise agrupamento. Além disso, estimar o poder da metodologia proposta para vários cenários hipotéticos e comparar com o método de identificação de *outliers* via Distância de Mahalanobis.

2. Referencial Teórico

2.1. Método de Monte Carlo

O Método de Monte Carlo é um método numérico que permite resolver problemas físicos ou matemáticos através da simulação de processos aleatórios [Sobol, 1994]. A criação deste método está ligada aos matemáticos norte-americanos J. von Neumann e S. Ulam, que foram os principais

responsáveis pela grande utilização do método de Monte Carlo em Física e Engenharia modernas, sem a necessidade de fundamentos sofisticados da teoria estatística [Sobol, 1994].

A geração de números aleatórios é feita através de algoritmos e esses valores gerados normalmente seguem as distribuições estatísticas das respectivas variáveis de interesse. O Método de Monte Carlo tem um algoritmo de estrutura relativamente simples. Elabora-se primeiro um programa para a realização de um evento aleatório e depois esse evento se repete N vezes de modo que cada experiência seja independente das outras.

2.2. Distribuição Normal Multivariada Contaminada

Dado o vetor aleatório $\boldsymbol{X}' = [X_1, X_2, ..., X_p] \in \Re^p$ com distribuição normal multivariada contaminada, sua função densidade de probabilidade será:

$$f(\boldsymbol{x}) = (1-\delta)(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma_1}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})'|\boldsymbol{\Sigma_1}|^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})\right] +$$

$$+\delta(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma_2}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_2})'|\boldsymbol{\Sigma_2}|^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2})\right] \quad (1)$$

em que $(1-\delta)$ é a probabilidade de que o processo tem de ser realizado por $N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$, δ é a probabilidade que o processo tem de ser realizado por $N_p(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$, $\boldsymbol{\Sigma_i}$ é uma matriz positiva definida, $\boldsymbol{\mu_i} \in \Re^p$ é o vetor de médias, $i = 1, 2$ e $0 \leq \delta \leq 1$.

Segundo [Johnson, 2011], a geração de variáveis estatísticas a partir da equação 1 é fácil e pode ser realizada como a seguir:

I. Gerar um valor u de uma distribuição uniforme contínua, com valores entre 0 e 1. Se $u \geq \delta$, avance para o passo II. Caso contrário, execute o passo III.

II. Gerar $\boldsymbol{X} \sim N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$.

III. Gerar $\boldsymbol{X} \sim N_p(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$.

2.3. Método de Identificação de *Outliers* via Distância de Mahalanobis

O uso da distância de Mahalanobis (MD) é sugerido por muitos autores como um método para detectar *outliers* em dados multivariados.

Pode-se definir a distância de Mahalanobis amostral como:

$$MD_i = \sqrt{(\boldsymbol{x_i}-\bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})} \quad (2)$$

em que $\bar{\boldsymbol{x}}$ é o vetor de médias amostrais do conjunto $\boldsymbol{X}$, e

$$\boldsymbol{S} = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})(\boldsymbol{x_i}-\bar{\boldsymbol{x}})'}{n-1} \quad (3)$$

é a matriz de variâncias e covariâncias amostrais de $\boldsymbol{X}$.

As medidas de distância, em especial a de Mahalanobis, são muito sensíveis à presença de *outliers*. Valores extremos, ou grupo de valores aberrantes, podem influenciar severamente estas medidas de distância.

Entretanto, como uma distância facilmente influenciada por *outliers*, pode ser capaz de identificá-los? A resposta é simples. Basta tratar das partes mais sensíveis desta medida, a média e a matriz de variâncias, calculando-as de forma robusta, em que a expressão 'robusta' significa resistência a observações atípicas.

O determinante mínimo da variância estimada (MCD) é provavelmente o método mais utilizado na prática para a construção de estimadores robustos, por se tratar de um algoritmo computacionalmente rápido [Rousseeuw e Driessen, 1999]. O estimador MCD é determinado por um

subconjunto de tamanho h, que minimize o determinante da matriz de covariâncias da amostra, calculado apenas sob os h pontos. A estimativa de dispersão é a média destes pontos, enquanto que o estimador de dispersão é proporcional à sua matriz de covariância, em que a escolha do tamanho de h determina a robustez do estimador.

Para indicar possíveis candidatos a *outliers*, baseados em MD_i, [Rousseeuw e van Zomeren, 1990] sugerem determinar aquelas observações cuja distância quadrática de Mahalanobis (MD^2) seja maior que $\chi^2_p(\alpha)$, em que p são os graus de liberdade e o número de variáveis consideradas, com valor de α sugerido igual a $0,975$.

3. Metodologia

3.1. Geração dos Dados

A geração de populações normais multivariadas com a presença de *outliers* foi realizada através da técnica de mistura de distribuições normais via simulação pelo Método de Monte Carlo.

A abordagem metodológica foi concebida em termos computacionais e os valores paramétricos assumidos nas simulações foram definidos nos vetores de médias $\boldsymbol{\mu_1}$ e $\boldsymbol{\mu_2}$ de dimensões $(p \times 1)$, da seguinte forma:

$$\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \boldsymbol{\mu_2} = \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{bmatrix} \tag{4}$$

Considerou-se, também, a matriz de covariância $\boldsymbol{\Sigma}$, de ordem p, definida da seguinte forma:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix} \tag{5}$$

em que ρ é o coeficiente de correlação assumido.

Nas simulações realizadas recorreram-se a:

- Tamanhos de amostras: $n = 50$, 100, 200 e 500;
- Número de variáveis: $p = 5$ e 30;
- Taxas de misturas: $\delta = 0$; 0,05 e 0,10;
- Coeficientes de correlação: $\rho = 0$; 0,2; 0,5; 0,7 e 0,9;
- Número de réplicas em cada caso: $nr = 100$.

Alternando os valores paramétricos descritos anteriormente, diferentes populações de distribuições normais multivariadas contaminadas foram geradas, a partir do seguinte processo:

I. Gerar um valor u de uma distribuição uniforme contínua, com valores entre 0 e 1.

II. Se $u \geq \delta$, então os dados assumirão valores de uma distribuição normal p-variada com a configuração $\boldsymbol{X} \sim N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$.

III. Se $u < \delta$, então os dados assumirão valores de uma distribuição normal p-variada com a configuração $\boldsymbol{X} \sim N_p(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$.

3.2. Descrição do Método Proposto

Uma vez obtida a população de interesse com a presença de *outliers*, utilizou-se o método de análise de agrupamento k-médias, com o objetivo de agrupar os indivíduos semelhantes. O número de grupos (k) da análise de agrupamento foram definidos com base no tamanho das amostras, sendo $k = \frac{n}{10}$ grupos. Uma peculiaridade do método de agrupamento k-médias é que, para iniciar seu processo, escolhe-se aleatoriamente k valores como centroides. Como esta escolha é aleatória, o método pode produzir partições diversas, ocasionando respostas diferentes em uma mesma análise. Para excluir essa aleatoriedade do método proposto, fixou-se a semente do processo aleatório do método de agrupamento k-médias. Essa fixação pode ser realizada, no software [R Core Team, 2014], por meio da função "set.seed(1)", que é inserida antes da função "kmeans". Dessa forma, o método de agrupamento k-médias produzirá sempre a mesma partição para um mesmo conjunto de dados.

Em seguida, calculou-se o centroide de cada grupo, assim como a mediana dos dados e através da distância euclidiana obteve-se a distância entre o centroide de cada grupo e a mediana dos indivíduos gerados na população. Para testar se um determinado grupo de indivíduos é um grupo de *outliers*, utilizou-se como critério uma medida baseada no desvio padrão amostral (s) das distâncias entre os centroides dos grupos e a mediana dos dados. Portanto, caso a distância euclidiana entre o centroide de um grupo e a mediana dos dados seja superior a $2,5s$, este grupo é definido como *outlier*, conforme ilustra a Figura 1. No caso ilustrado, o grupo 4 seria definido como um grupo outlier, uma vez que $d(C4, Md) > 2,5s$.

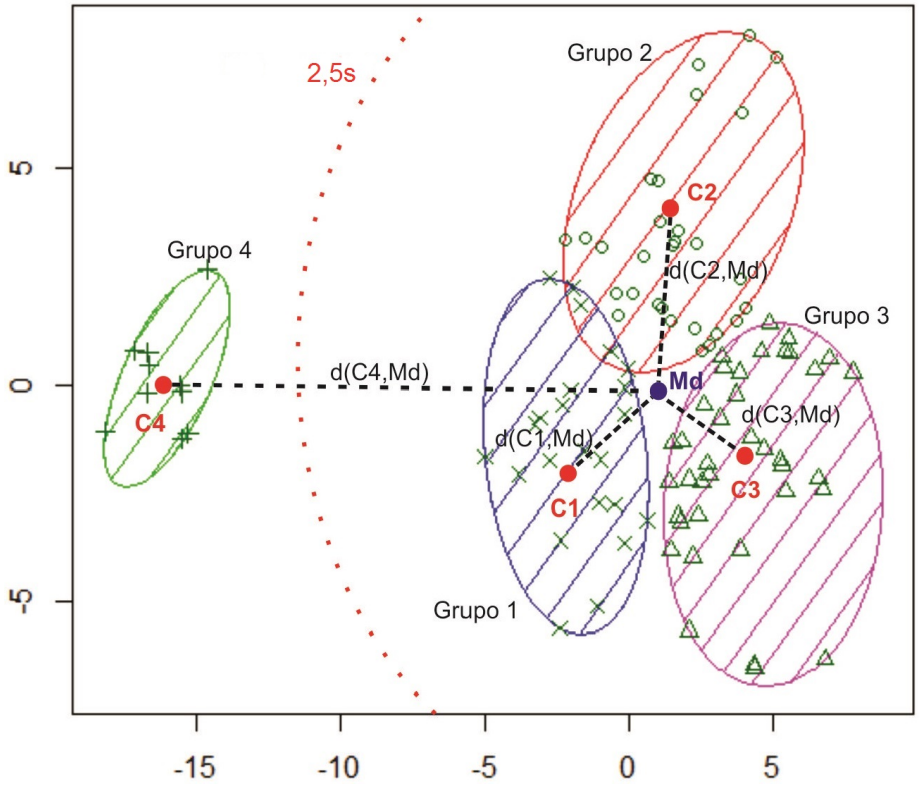

Figura 1: Representação gráfica do método

Para analisar a qualidade da técnica proposta neste trabalho, foram utilizadas as medidas sensibilidade (S) e especificidade (E), descritas com base na Tabela 1. A sensibilidade é a capacidade que o teste apresenta de detectar os indivíduos verdadeiramente positivos, ou seja, de diagnosticar corretamente os *outliers*, enquanto que a especificidade é a capacidade que o teste tem de detectar os verdadeiros negativos, isto é, de diagnosticar corretamente os indivíduos não *outliers*.

Tabela 1: Medidas de eficiência

Método \ *Outlier*	Sim	Não	Total
Positivo	a (verdadeiros positivos)	b (falsos positivos)	$a+b$ (positivos)
Negativo	c (falsos negativos)	d (verdadeiros negativos)	$c+d$ (negativos)
Total	$a+c$ (*outliers*)	$b+d$ (não *outliers*)	$a+b+c+d$ (n)

A partir da Tabela 1 podemos definir as medidas de sensibilidade e especificidade da seguinte forma:

$$S = \frac{a}{a+c} \tag{6}$$

e

$$E = \frac{d}{b+d} \tag{7}$$

Para que o método proposto seja considerado eficiente na detecção de *outliers*, espera-se que os valores dos verdadeiros positivos estejam próximos do total de *outlier* e os valores dos verdadeiros negativos estejam próximos do total de não *outlier*, ou seja, espera-se que os valores de S e E estejam próximos de 1.

4. Resultados das Simulações e Comparação dos Métodos

Considerando as possíveis variações de n, p, δ e ρ foram simulados 120 cenários hipotéticos e para cada cenário foram realizadas 100 réplicas. Para efeito de comparação, cada caso foi submetido ao método proposto nesse trabalho e também ao método de identificação de *outliers* via distância de Mahalanobis. Em cada caso, obtiveram-se as médias pontual e intervalar para as medidas de sensibilidade e especificidade, considerando as 100 réplicas, bem como foi realizado um teste para verificar se as médias são estatisticamente iguais, ao nível de 5% de significância. Como as amostras foram submetidas aos dois métodos, utilizou-se o teste t de Student pareado, com $nr-1$ graus de liberdade, sendo nr o número de réplicas. A Tabela 2 apresenta os resultados das simulações para o caso hipotético em que $p=30$, $n=500$ e $\delta=0,05$.

Tabela 2: Comparação entre os dois métodos considerando $p=30$, $n=500$, $\delta=0,05$ e os coeficientes de correlação (ρ), contendo intervalo de confiança inferior (IC inf.), média, intervalo de confiança superior (IC sup.), estatística de teste (t) e p-valor tanto para sensibilidade (S) quanto para especificidade (E)

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	***p*-valor**
$\rho=0$	S	1	1	1	1	1	1	-	-
	E	0,70226	0,77229	0,84232	0,89531	0,89772	0,90014	-3,52768	0,00064
$\rho=0,2$	S	0,95837	0,96777	0,97718	0,53809	0,56547	0,59285	29,01197	$<0,001$
	E	0,87559	0,89415	0,91271	0,88888	0,89142	0,89396	0,28944	0,77285
$\rho=0,5$	S	0,79326	0,8126	0,83194	0,292	0,31154	0,33107	36,6249	$<0,001$
	E	0,91076	0,91845	0,92614	0,88524	0,88777	0,8903	7,19116	$<0,001$
$\rho=0,7$	S	0,65727	0,68133	0,70539	0,23784	0,2567	0,27556	27,20583	$<0,001$
	E	0,91542	0,92293	0,93044	0,88306	0,88584	0,88863	9,51673	$<0,001$
$\rho=0,9$	S	0,5197	0,5438	0,5679	0,19591	0,21239	0,22888	22,42424	$<0,001$
	E	0,93168	0,93721	0,94275	0,88265	0,88516	0,88766	17,62282	$<0,001$

Analisando a Tabela 2, pode-se observar que para $\rho = 0$ ambos os métodos obtiveram média 1 (100% de acerto) para a sensibilidade, enquanto que para especificidade o método proposto obteve média de 0,77229 e o método de Mahalanobis de 0,89772. O teste t de Student apresentou um p-valor de 0,00064, portanto, para especificidade e correlação nula, pode-se concluir que o método de Mahalanobis obteve uma média de acertos estatisticamente superior ao método proposto, com significância superior a $0,1\%$. Considerando $\rho = 0,2$, o método proposto apresentou média superior ao de Mahalanobis para sensibilidade, enquanto que para especificidade as médias foram estatisticamente iguais. Já para $\rho = 0,5$, $\rho = 0,7$ e $\rho = 0,9$ o método proposto apresentou médias superiores tanto para sensibilidade quanto para especificidade.

Em relação às médias para a sensibilidade, nota-se que à medida que ρ aumenta, a qualidade de ambos os métodos reduz, principalmente quando $\rho \geq 0,7$. Entretanto, em uma situação prática, caso um conjunto de dados apresente variáveis correlacionadas, uma possível solução seria primeiramente aplicar uma técnica multivariada de redução de dimensões, tais como análise fatorial ou componentes principais, e posteriormente fazer a análise de *outliers*.

As Tabelas 3 e 4 apresentam um resumo dos resultados das comparações das médias de sensibilidade e especificidade obtidas pelo método proposto e pelo método de identificação de *outliers* via distância de Mahalanobis.

Tabela 3: Resultado das comparações das médias de sensibilidade entre os dois métodos

Resultado	**Frequência Absoluta**	**Frequência Relativa**
Iguais	12	15%
Mahalanobis supera	4	5%
Proposto supera	64	80%
Total	**80***	**100%**

*Não se calcula sensibilidade quando $\delta = 0$, pois nesse caso não há verdadeiros positivos e nem *outliers*. Portanto, temos 80 cenários.

Tabela 4: Resultado das comparações das médias de especificidade entre os dois métodos

Resultado	**Frequência Absoluta**	**Frequência Relativa**
Iguais	8	6,67%
Mahalanobis supera	29	24,17%
Proposto supera	83	69,17%
Total	**120**	**100%**

Os resultados apresentados na Tabela 3 mostram que para a sensibilidade o método proposto foi superior em 80% dos casos e o método de Mahalanobis em apenas 5%. Já os resultados apresentados na Tabela 4 mostram que para a especificidade o método proposto foi superior em 69,17%, enquanto que o método de Mahalanobis em aproximadamente 24%. Esses resultados mostram que, nos cenários simulados, de um modo geral, o método proposto foi superior ao método de Mahalanobis.

Para avaliar os resultados das simulações de acordo com as variações de ρ, δ , n e p, foram construídas as tabelas de contingência dos resultados tanto para a sensibilidade quanto para a especificidade.

As Tabelas 5, 6, 7 e 8 apresentam os resultados das simulações e das comparações das médias de sensibilidade entre os dois métodos levando-se em consideração a correlação (ρ) existente entre as variáveis, a taxa de mistura (δ) atribuída, o tamanho da amostra (n) e o número de variáveis (p), respectivamente.

Na Tabela 5 pode-se verificar que, considerando os 16 cenários hipotéticos em que $\rho = 0$, em 8 cenários as médias de sensibilidade dos dois métodos foram consideradas estatisticamente

Tabela 5: Resultado das comparações das médias de sensibilidade entre os dois métodos levando-se em consideração a correlação (ρ) existente entre as variáveis

Resultado \ ρ	**0**	**0,2**	**0,5**	**0,7**	**0,9**	**Total**
Iguais	8	1	0	1	2	12
Mahalanobis supera	4	0	0	0	0	4
Proposto supera	4	15	16	15	14	64
Total	**16**	**16**	**16**	**16**	**16**	**80**

iguais, enquanto que nos demais níveis de ρ o método proposto foi superior em pelo menos 14 cenários.

Tabela 6: Resultado das comparações das médias de sensibilidade entre os dois métodos levando-se em consideração a taxa de mistura (δ)

Resultado \ δ	**0**	**0,05**	**0,1**	**Total**
Iguais	0	8	4	12
Mahalanobis supera	0	1	3	4
Proposto supera	0	31	33	64
Total	**0**	**40**	**40**	**80**

Tabela 7: Resultado das comparações das médias de sensibilidade entre os dois métodos levando-se em consideração o tamanho da amostra (n)

Resultado \ n	**50**	**100**	**200**	**500**	**Total**
Iguais	1	6	2	3	12
Mahalanobis supera	0	1	2	1	4
Proposto supera	19	13	16	16	64
Total	**20**	**20**	**20**	**20**	**80**

Tabela 8: Resultado das comparações das médias de sensibilidade entre os dois métodos levando-se em consideração o número de variáveis (p)

Resultado \ p	**5**	**30**	**Total**
Iguais	7	5	12
Mahalanobis supera	4	0	4
Proposto supera	29	35	64
Total	**40**	**40**	**80**

A Figura 2 mostra como os resultados das simulações se comportam levando-se em consideração as relações existentes entre as taxas de acerto das médias da sensibilidade para os dois métodos e os valores paramétricos ρ, δ, n e p.

A partir da análise das Tabelas 5, 6, 7, 8 e da Figura 2 é possível verificar que o método proposto apresenta resultados ainda melhores quando $\rho > 0$, ou seja, quando há a presença de correlação entre as variáveis, assim como quando o número de variáveis é maior.

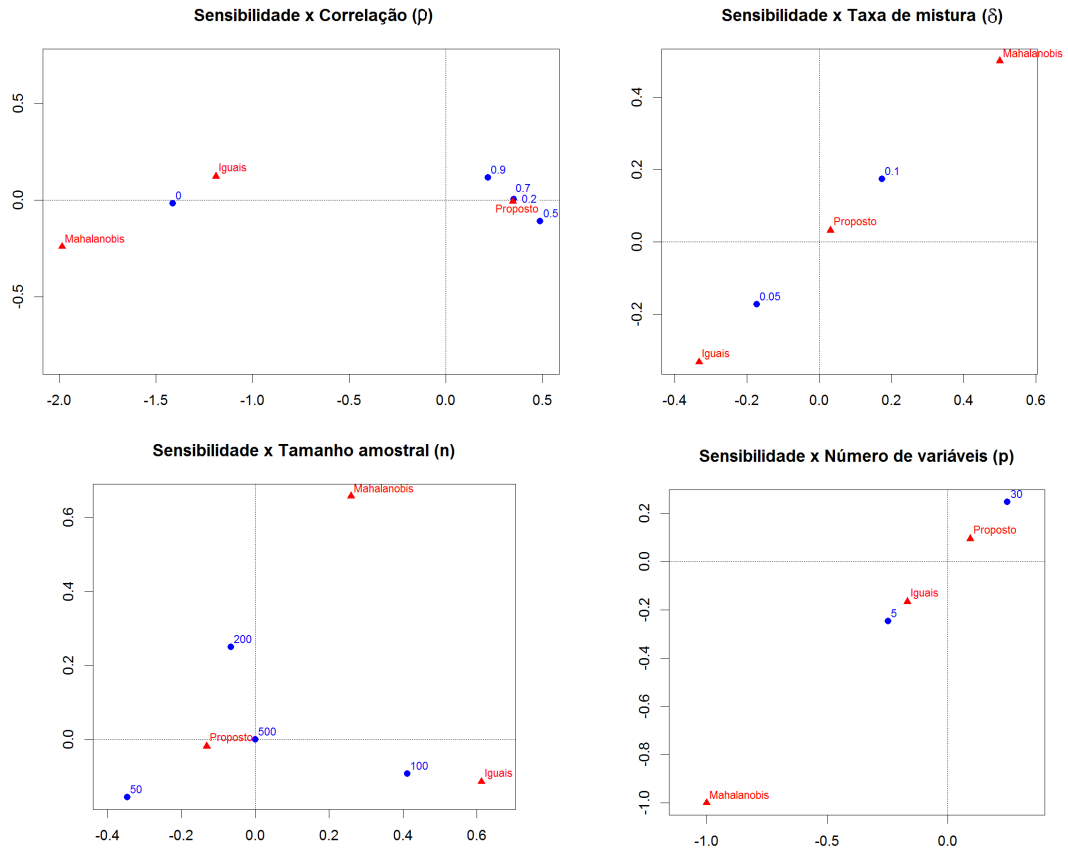

Figura 2: Gráficos de análise de correspondência: sensibilidade X valores paramétricos

As Tabelas 9, 10, 11 e 12 apresentam os resultados das comparações das médias de especificidade entre os dois métodos levando-se em consideração ρ, δ, n e p, respectivamente.

Tabela 9: Resultado das comparações das médias de especificidade entre os dois métodos levando-se em consideração a correlação (ρ) existente entre as variáveis

Resultado \ ρ	**0**	**0,2**	**0,5**	**0,7**	**0,9**	**Total**
Iguais	2	5	1	0	0	8
Mahalanobis	13	10	4	2	0	29
Proposto	9	9	19	22	24	83
Total	**24**	**24**	**24**	**24**	**24**	**120**

Tabela 10: Resultado das comparações das médias de especificidade entre os dois métodos levando-se em consideração a taxa de mistura (δ)

Resultado \ δ	**0**	**0,05**	**0,1**	**Total**
Iguais	1	5	2	8
Mahalanobis	22	5	2	29
Proposto	17	30	36	83
Total	**40**	**40**	**40**	**120**

Tabela 11: Resultado das comparações das médias de especificidade entre os dois métodos levando-se em consideração o tamanho da amostra (n)

Resultado \ n	50	100	200	500	Total
Iguais	3	2	1	2	8
Mahalanobis	4	5	8	12	29
Proposto	23	23	21	16	83
Total	**30**	**30**	**30**	**30**	**120**

Tabela 12: Resultado das comparações das médias de especificidade entre os dois métodos levando-se em consideração o número de variáveis (p)

Resultado \ p	5	30	Total
Iguais	7	1	8
Mahalanobis	18	11	29
Proposto	35	48	83
Total	**60**	**60**	**120**

A Figura 3 mostra como os resultados das simulações se comportam levando-se em consideração as relações entre as taxas de acerto das médias da especificidade para os dois métodos e os valores paramétricos ρ, δ, n e p.

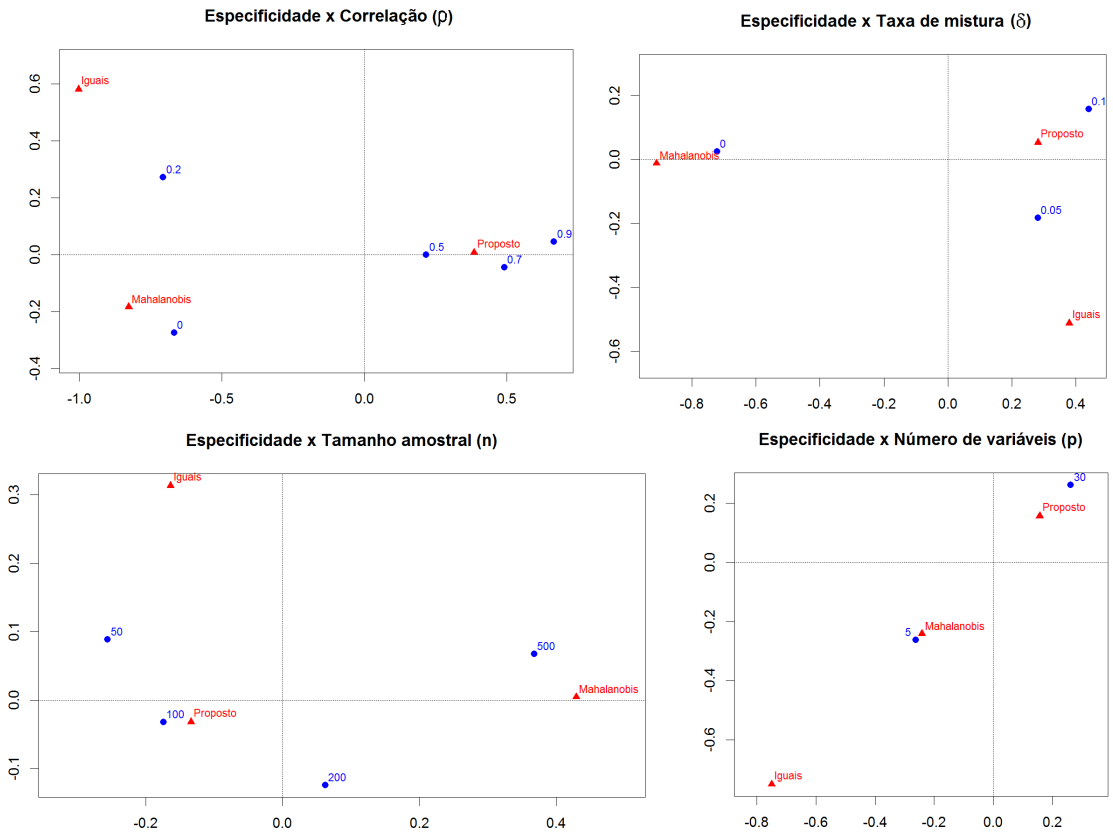

Figura 3: Gráficos de análise de correspondência: especificidade X valores paramétricos

Assim como ocorreu para a sensibilidade, analisando as Tabelas 9, 10, 11, 12 e a Figura 3, é possível verificar que para a especificidade o método proposto apresenta resultados ainda melhores quando há a presença de correlação, principalmente quando $\rho \geqslant 0,5$, assim como quando o número de variáveis é maior e também quando a taxa de mistura é diferente de 0.

De um modo geral o método proposto foi superior ao método de Mahalanobis tanto para sensibilidade quanto para especificidade, mas vale destacar que quando há correlação entre as variáveis analisadas, o que normalmente ocorre na prática, o método proposto apresenta resultados ainda melhores.

5. Considerações Finais

Os resultados apresentados nas simulações mostram que o método proposto nesse trabalho foi superior ao método de identificação de *outliers* via distância de Mahalanobis tanto para a sensibilidade quanto para a especificidade, ou seja, ele apresenta maior capacidade de diagnosticar corretamente os indivíduos *outliers* e os não *outliers*. Além disso, foi possível verificar que o método proposto apresenta resultados ainda melhores, comparado ao método de Mahalanobis, quando há a presença de correlação, assim como quando o número de variáveis em estudo é maior.

Agradecimentos

Esta pesquisa é parcialmente apoiada pelo CNPq (processo 300825 / 2016-1), FAPEMIG (processo CEX-PPM-00427-17), CAPES e Universidade Federal de Ouro Preto (processo Edital PROPP 09/2016).

Referências

Aggarwal, C. C. (2013). An introduction to outlier analysis. In *Outlier Analysis*, p. 1–40. Springer.

Bamnett, V. e Lewis, T. (1994). Outliers in statistical data.

Berton, L., Huertas, J., Araújo, B., e Zhao, L. (2010). Identifying abnormal nodes in complex networks by using random walk measure. In *IEEE Congress on Evolutionary Computation*, p. 1–6. IEEE.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, p. 133–169.

Critchley, F. (1985). Influence in principal components analysis. *Biometrika*, 72(3):627–636.

Filzmoser, P., Maronna, R., e Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Chapman and Hall.

Johnson, M. E. (2011). Multivariate statistical simulation. In *International Encyclopedia of Statistical Science*, p. 930–932. Springer.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Rousseeuw, P. J. e Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223. URL `http://www.tandfonline.com/doi/abs/10.1080/00401706.1999.10485670`.

Rousseeuw, P. J. e van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639. ISSN 01621459. URL `http://www.jstor.org/stable/2289995`.

Sobol, I. M. (1994). *A Primer for the Monte Carlo Method*. CRC PRESS.

Valadares, F. G., Aquino, A. L. L., e Junior, A. R. P. (2012). Detecção de outliers multivariados em redes de sensores. In *XLIV Simpósio Brasileiro de Pesquisa Operacional*. SBPO.

Veloso, M. V. S. e Cirillo, M. A. (2016). Principal components in the discrimination of outliers: A study in simulation sample data corrected by pearson's and yates's chi-square distance. *Acta Scientiarum. Technology*, 38(2):193–200.