Artigo Original

DOI:10.5902/2179460X29535

Ciência e Natura, Santa Maria v.40, e40, 2018
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307
ISSN on-line: 2179-460X

CIÊNCIA e NATURA

Recebido: 13/10/2017 Aceito: 16/05/2018

Uma proposta para identificação de *outliers* multivariados

A proposal for identifying multivariate outliers

Josino José Barbosa[1], Tiago Martins Pereira[2] e Fernando Luiz Pereira de Oliveira[2]

[1]Universidade Federal de Viçosa, MG, Brasil
[2]Universidade Federal de Ouro Preto, MG, Brasil

Resumo

A identificação de outliers desempenha um papel importante na análise estatística, pois tais observações podem conter informações importantes em relação às hipóteses do estudo. 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. Desta forma, objetivou-se 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 via método de Monte Carlo e a técnica de mistura de distribuições normais multivariadas. Os resultados apresentados nas simulações mostraram que o método proposto foi superior ao método de Mahalanobis tanto para sensibilidade quanto para especificidade, ou seja, ele apresentou maior capacidade de diagnosticar corretamente os indivíduos outliers e os não outliers. Além disso, a metodologia proposta foi ilustrada com uma aplicação em dados reais provenientes da área de saúde.

Palavras-chave: *Outlier, Análise de Agrupamento, Método de Monte Carlo.*

Abstract

The identification of outliers plays an important role in the statistical analysis, since such observations may contain important information regarding the hypotheses of the study. If classical statistical models are blindly applied to data containing atypical values, the results may be misleading and mistaken decisions can be made. Moreover, in practical situations, the outliers themselves are often the special points of interest and their identification may be the main objective of the investigation. In this way, it was proposed 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, Monte Carlo method simulation and the mixed multivariate normal distribution technique were used. The results presented in the simulations showed that the proposed method was superior to the Mahalanobis method for both sensitivity and specificity, that is, it presented greater ability to correctly diagnose outliers and non-outliers individuals. In addition, the proposed methodology was illustrated with an application in real data from the health area.

Keywords: *Outlier, Grouping Analysis, Monte Carlo Method.*

1 Introdução

Outlier é uma observação, ou um subconjunto de observações, que parecem ser inconsistentes quando comparados ao restante do conjunto (Hawkins, 1980). Segundo Bamnett e Lewis (1994), *outlier* é uma observação que desvia muito de outras observações despertando suspeitas de que são geradas por um mecanismo diferente. Estas observações são também designadas por anormais, discrepantes, 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 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 dessas características incomuns fornece uma série de aplicações úteis, tais como em sistemas de detecção de intrusão, fraude de cartão de crédito, sensores de eventos, diagnóstico médico, entre outros (Aggarwal, 2013).

Em função de sua ampla gama de aplicações, muitas técnicas têm sido desenvolvidas para a detecção de *outliers*. Oliveira et al. (2006) e Filzmoser (2004) trabalharam na identificação de valores discrepantes por meio da distância de Mahalanobis, técnica sugerida por diversos autores para detectar outliers em dados multivariados. 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 (MAD). Berton et al. (2010) apresentam o desenvolvimento de um método baseado em redes complexas para detecção de diferentes tipos de *outliers* que utiliza 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. Veloso e Cirillo (2016) propuseram uma técnica para detecção de *outliers* baseada em componentes principais com amostras corrigidas por distância do tipo qui-quadrado.

Diante da importância da identificação de *outliers* e buscando ampliar as possibilidades de detecção dos mesmos, os objetivos desse trabalho são propor uma técnica de detecção de outliers multivariados baseada em análise agrupamento, estimar o poder da metodologia proposta para vários cenários hipotéticos, comparar essa técnica com o método de identificação de outliers via Distância de Mahalanobis e apresentar uma aplicação da técnica desenvolvida em uma base de dados reais.

2 Material e Métodos

A geração de populações normais multivariadas com a presença de *outliers* pode ser realizada por meio da mistura de distribuições normais multivariadas via simulação pelo método de Monte Carlo. Essa mistura gera populações cuja distribuição é também conhecida como 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:

$$f(\boldsymbol{x}) = (1-\delta)(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma_1}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})'|\boldsymbol{\Sigma_1}|^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})\right] + \delta(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma_2}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_2})'|\boldsymbol{\Sigma_2}|^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2})\right]$$

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 $0 \leq \delta \leq 1$.

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 $(p \times 1)$ e na matriz de covariância $\boldsymbol{\Sigma} = \boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2}$, de ordem p, definidos da seguinte forma:

$$\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \boldsymbol{\mu_2} = \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix}$$

em que ρ é o coeficiente de correlação entre as variáveis.

Nas simulações realizadas recorreram-se a: n (tamanho de amostra) = 50, 100, 200 e 500; p (número de variáveis) = 5 e 30; δ (taxa de mistura) = 0, 0,05 e 0,10; ρ (coeficiente de correlação) = 0, 0,2, 0,5, 0,7 e 0,9; nr (número de réplicas por caso) = 100.

Alternando os valores paramétricos descritos anteriormente, diferentes populações de distribuições normais multivariadas com a presença de *outliers* foram geradas, a partir dos seguintes passos:

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

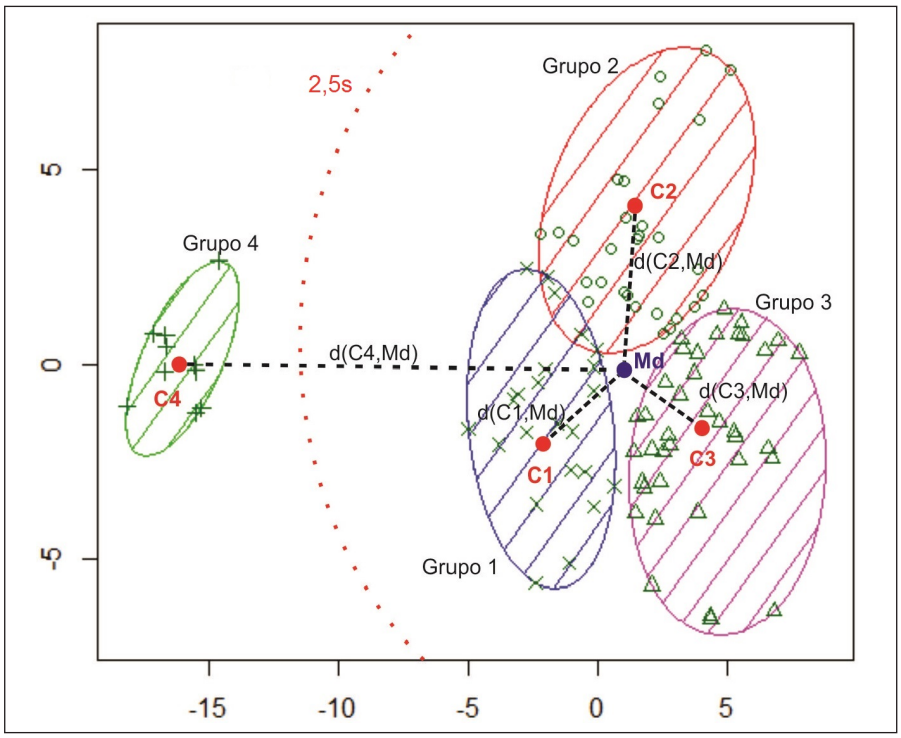
2) Se $u \geq \delta$, então os dados assumirão valores de uma distribuição normal p-variada com a configuração $\boldsymbol{X} \sim N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$. As observações geradas por esse processo serão os indivíduos não *outliers*, ou seja, as observações comuns;

3) Se $u < \delta$, então os dados assumirão valores de uma distribuição normal p-variada com a configuração $\boldsymbol{X} \sim N_p(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$. As observações geradas por esse processo serão os indivíduos *outliers*. Na definição dos *outliers* procurou-se utilizar vetores de médias $\boldsymbol{\mu_1}$ e $\boldsymbol{\mu_2}$ com valores não muito distantes para não tornar óbvia a identificação dos *outliers*.

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. No método k-médias o número de grupos (k) é definido à priori, logo, para captar melhor a variação e a estrutura de correlação dos dados, o número de grupos foi definido com base no tamanho da amostra, 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 (2017), 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 (C) de cada grupo, assim como a mediana (Md) dos dados e através da distância euclidiana (d) obteve-se a distância entre o centroide de cada grupo e a mediana dos indivíduos gerados. 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 for superior a $2{,}5s$, este grupo é definido como *outlier*, conforme ilustra a Figura 1. Foram realizados também testes utilizando 2 e 3 desvios como critério, mas os melhores resultados foram obtidos com 2,5 desvios.

Figura 1: Representação gráfica do método proposto. A curva em linha pontilhada vermelha representa o limite de 2,5 desvios padrão, definido a partir da mediana dos dados. A d(C4,Md) representa a distância euclidiana entre o centroide do grupo 4 e a mediana dos dados



No caso hipotético ilustrado na Figura 1 temos $k = 4$ grupos e justamente o grupo 4 seria definido como um grupo de indivíduos outlier, uma vez que a distância euclidiana entre o centroide desse grupo e a mediana dos dados foi superior a $2{,}5s$, ou seja, $d(C4{,}Md) > 2{,}5s$.

Para analisar a qualidade da técnica proposta, utilizaram-se as medidas sensibilidade (S) e especificidade (E). Nas simulações, o indivíduo pode ser um *outlier* ou não. Além disso, o método pode atestar positivo ou negativo para identificação de *outliers*. A sensibilidade é a razão entre o número de indivíduos *outliers* que o método classificou como positivo e o número total de *outliers*. Já a especificidade é a razão entre o número de indivíduos não *outliers* que o método classificou como negativo e o número total de não *outliers*. Portanto, a sensibilidade é a capacidade que o método apresenta de detectar os indivíduos verdadeiramente positivos, ou seja, de diagnosticar corretamente os *outliers*, enquanto que a especificidade é a capacidade que o método tem de detectar os verdadeiros negativos, isto é, de diagnosticar corretamente os indivíduos não *outliers*. Para que o método seja considerado eficiente na detecção de *outliers*, espera-se que os valores de S e E estejam próximos de 1.

Para detectar *outliers* em dados multivariados muitos autores sugerem o uso da distância de Mahalanobis (MD). Para efeito de comparação, cada cenário foi submetido ao método proposto nesse trabalho e também ao método de identificação de *outliers* via distância de Mahalanobis, utilizando o estimador robusto MCD (Rousseeuw e Driessen, 1999). Para indicar possíveis candidatos a *outliers* 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 de uma distribuição qui-quadrado e o número de variáveis consideradas, com $\alpha = 0{,}975$.

Considerando as possíveis variações de n, p, δ e ρ foram simulados 120 cenários hipotéticos e para cada cenário foram realizadas 100 réplicas. Em cada caso, obtiveram-se a média pontual e o intervalo de confiança 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.

Em seguida, a metodologia proposta foi ilustrada com uma aplicação em dados reais, obtidos através de coleta realizada em mineradoras da Região dos Inconfidentes, Minas Gerais, no ano de 2015. O projeto de pesquisa referente à coleta dos dados foi submetido e aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de Ouro Preto (CAAE: 39682014.7.0000.5150), sob o parecer de número 1.381.376. A amostra foi composta de 214 operadores de caminhão fora de estrada, do sexo masculino, com idade média de 34 anos, que trabalhavam em regime de turnos alternantes, com jornada de trabalho de 6 horas por turno e descanso de 12 horas entre turnos. A coleta dos dados foi realizada nos ambulatórios das minas e foi realizado um teste piloto para a realização e validação dos procedimentos.

O conjunto de dados apresenta 12 variáveis, distribuídas em 9 fatores de risco, cujos limites foram definidos exclusivamente para esse estudo e estão descritos a seguir:

1) Pressão Arterial Sistólica (PAS) e Pressão Arterial Diastólica (PAD): A pressão arterial é a pressão que o sangue exerce na parede das artérias. Indivíduos com pressão acima de 140 x 90 mmHg são considerados hipertensos. Nesse caso, tem-se duas variáveis e somente um fator de risco que é a pressão arterial;

2) Glicose: A glicemia é a glicose que circula pela corrente sanguínea. Para o bom funcionamento do organismo e para o equilíbrio de um estado de saúde, é necessário que seus níveis estejam estáveis. Valores inferiores a 60 mg/dL correspondem a problemas relacionados com a hipoglicemia, enquanto que superiores a 100 a problemas relacionados com a hiperglicemia, ou talvez, com a diabetes mellitus;

3) Vitamina D: A vitamina D é fundamental para o equilíbrio do cálcio e do fósforo no organismo e para a saúde do esqueleto. Indivíduos com valores de vitamina D inferiores a 30 ng/mL estão propensos a apresentar osteoporose e outras doenças;

4) Colesterol não-HDL: É a soma de todos os tipos de colesterol considerados ruins: IDL+LDL+VLDL. Valores superiores a 200 mg/dL de colesterol não-HDL são considerados ruins;

5) HDL: Promove a retirada do excesso de colesterol das células, inclusive das placas arteriais. Por isso, denomina-se o HDL como colesterol bom. Valores de HDL inferiores a 40 mg/dL são considerados ruins;

6) LDL: Leva colesterol para as células e facilita a deposição de gordura nos vasos sanguíneos. Por isso, denomina-se o LDL como um colesterol ruim. Valores de LDL superiores a 130 mg/dL são considerados elevados;

7) Triglicérides: A hipertrigliceridemia, nome que se dá ao aumento dos triglicerídeos no sangue, também é fator de risco para aterosclerose, principalmente se associados a níveis baixos de HDL. Valores de triglicérides superiores a 150 mg/dL são considerados ruins;

8) Circunferência da cintura (CC), circunferência do quadril (CQ) e relação cintura x quadril (RCQ): Estudos científicos relacionam futuras doenças e risco à saúde com a quantidade de gordura depositada em determinadas partes do corpo, como na região abdominal. Quanto mais alto for o valor da RCQ maior é o risco à saúde. Para esse estudo, o índice de corte para risco cardiovascular é 0,90. Nesse caso, tem-se três variáveis e somente um fator de risco que é a relação cintura x quadril;

9) Circunferência do pescoço (CP): A CP aumentada favorece o desenvolvimento de doenças cardiovasculares. Indivíduos com CP acima de 40 cm são considerados elevados.

Com o objetivo de se identificar grupos de indivíduos que apresentam maior possibilidade de desenvolverem doenças cardiovasculares e que consequentemente apresentam maior chance de ocorrência de acidentes de trabalho, o conjunto de dados foi submetido ao método proposto nesse artigo.

3 Resultados e Discussão

3.1 Resultados das Simulações e Comparação dos Métodos

A Tabela 1 apresenta os resultados das comparações das médias de sensibilidade e especificidade obtidas pelos dois métodos, considerando os 120 cenários hipotéticos. Entretanto, vale destacar que não se calcula sensibilidade quando $\rho = 0$, pois nesse caso não há verdadeiros positivos e nem *outliers*. Logo, nesse caso, tem-se 80 cenários. Para a sensibilidade, o método proposto foi superior em 80% dos casos e o método de Mahalanobis em apenas 5%. Já para especificidade, o método proposto foi superior em 69,17%, enquanto que o método de Mahalanobis em aproximadamente 24%.

Tabela 1: Resultado das simulações e comparação entre os dois métodos tanto para sensibilidade quanto para especificidade

Medida	**Resultado**	**Frequência Absoluta**	**Frequência Relativa**
Sensibilidade	Iguais	12	15,00%
	Mahalanobis supera	4	5,00%
	Proposto supera	64	80,00%
	Total	80	100,00%
Especificidade	Iguais	8	6,67%
	Mahalanobis supera	29	24,17%
	Proposto supera	83	69,17%
	Total	120	100,00%

Para ilustrar como as simulações ocorreram, a Tabela 2 apresenta os resultados das simulações para 5 cenários hipotéticos em que $p = 30$, $n = 500$ e $\delta = 0{,}05$. Pode-se observar que para $\rho = 0$ (correlação nula) ambos os métodos obtiveram média 1 (100% de acerto) para sensibilidade, enquanto que para especificidade o método proposto obteve média de 0,772 e o método de Mahalanobis de 0,898. Pelo teste teste t de Student, para esse cenário, o método de Mahalanobis apresentou média de especificidade superior àquela apresentada pelo método proposto (p-valor<0,001%). 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 (p-valor<0,001%). Em relação às médias de acerto para 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 com correlação mais forte, uma possível solução seria primeiramente aplicar uma técnica multivariada de redução de dimensões, tais como análise fatorial ou componentes principais.

Tabela 2: Resultado das simulações e comparação entre os dois métodos considerando $p = 30$, $n = 500$, $\delta = 0{,}05$ e os coeficientes de correlação (ρ), contendo limite inferior do intervalo de confiança (IC inf.), média, limite superior do intervalo de confiança (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	0,702	0,772	0,842	0,895	0,898	0,900	-3,528	<0,001
$\rho = 0{,}2$	S	0,958	0,968	0,977	0,538	0,565	0,593	29,012	<0,001
	E	0,876	0,894	0,913	0,889	0,891	0,894	0,289	0,773
$\rho = 0{,}5$	S	0,793	0,812	0,832	0,292	0,312	0,331	36,625	<0,001
	E	0,911	0,918	0,926	0,885	0,888	0,890	7,191	<0,001
$\rho = 0{,}7$	S	0,657	0,681	0,705	0,238	0,257	0,276	27,206	<0,001
	E	0,915	0,923	0,930	0,883	0,886	0,889	9,517	<0,001
$\rho = 0{,}9$	S	0,520	0,544	0,568	0,196	0,212	0,229	22,424	<0,001
	E	0,932	0,937	0,943	0,883	0,885	0,888	17,623	<0,001

Para avaliar os resultados das simulações de acordo com as variações do coeficiente de correlação (ρ) foram construídas tabelas de contingência dos resultados tanto para a sensibilidade quanto para a especificidade.

A Tabela 3 apresenta os resultados das simulações e comparação entre os dois métodos para a sensibilidade levando-se em consideração a correlação existente entre as variáveis. Considerando os 16 cenários em que $\rho = 0$ (correlação nula), pode-se verificar que em 8 cenários as médias de sensibilidade dos dois métodos foram consideradas estatisticamente iguais, enquanto que nos demais níveis de correlação o método proposto foi superior em pelo menos 14 cenários.

Tabela 3: Resultado das simulações para sensibilidade levando-se em consideração a correlação entre as variáreis

ρ / **Resultado**	**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**

A Tabela 4 apresenta os resultados das simulações e comparação entre os dois métodos para a especificidade levando-se em consideração a correlação existente entre as variáveis. Considerando os 24 cenários em cada nível de correlação, pode-se observar que o método de Mahalanobis foi levemente superior quando $\rho = 0$ (correlação nula) e $\rho = 0{,}2$, enquanto que nos demais níveis de correlação o método proposto foi superior em pelo menos 19 cenários.

Tabela 4: Resultado das simulações para especificidade levando-se em consideração a correlação entre as variáreis

ρ / **Resultado**	**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**

Majoritariamente, quando há correlação entre as variáveis analisadas, o método proposto foi superior ao método de Mahalanobis na identificação dos *outliers* tanto para sensibilidade quanto para especificidade. Diante desses resultados, o método proposto torna-se interessante uma vez que em situações práticas normalmente há correlação entre as variáveis.

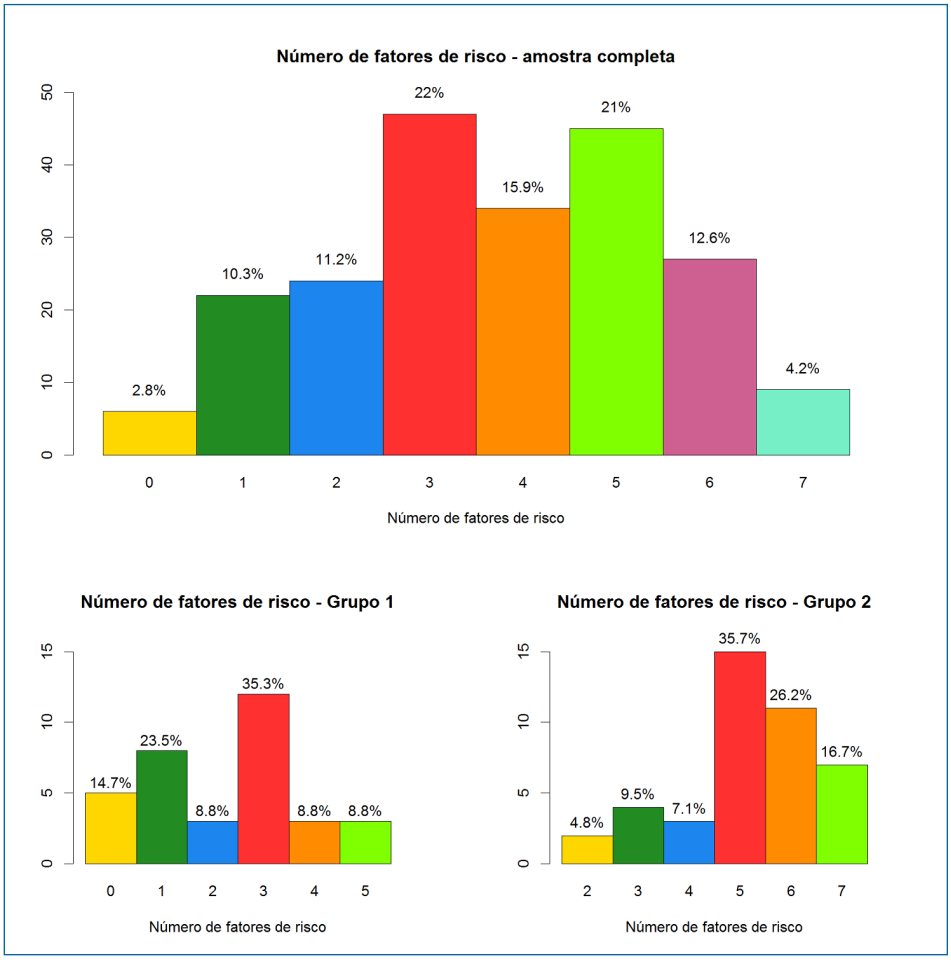
3.2 Análise de Dados Reais

Antes de iniciar a aplicação da metodologia proposta nos dados reais foi verificada a existência de correlação significativa entre a maioria das variáveis, com valores variando entre -0,237 e 0,906. Diante da evidência de correlação entre as variáveis e dos resultados apresentados nas simulações, o método proposto mostrou-se indicado para a análise dos dados. Em função da não homogeneidade das variáveis optou-se por padronizá-las. Após a aplicação do método e a definição do grupo de *outliers*, composto inicialmente por 76 indivíduos, foi possível observar que esse grupo apresentava indivíduos com valores divergentes, sendo uns mais elevados e outros mais baixos. Como o método baseia-se na distância euclidiana entre os centroides dos grupos e a mediana dos dados, indivíduos que apresentam tanto valores altos quanto baixos nas variáveis em estudo estão sujeitos a serem captados pelo método. Portanto, com intuito de se separar esses indivíduos heterogêneos, foi realizada uma análise de agrupamento pelo método k-médias, com k = 2 grupos. Após a análise, o grupo 1 ficou composto por 34 indivíduos, enquanto que o grupo 2 composto por 42. Em seguida, foi realizada uma análise descritiva das variáveis nos dois grupos e foi possível verificar que o segundo grupo apresenta valores muito superiores ao primeiro, com exceção das variáveis Vitamina D e HDL, tanto para mediana quanto para média.

Considerando-se os limites de risco especificados na descrição das variáveis, para facilitar a visualização da distinção dos grupos, somaram-se o número de fatores de risco de cada indivíduo e calcularam-se as frequências dessas somas, que podem ser observadas na Figura 2. Por exemplo, 12 indivíduos (35,3%) do grupo 1 apresentam 3 fatores de risco, enquanto que 15 indivíduos (35,7%) do grupo 2 apresentam 5 fatores de risco.

Conforme pode ser visto na Figura 2, 82,35% dos indivíduos do grupo 1 apresentam risco em no máximo 3 fatores, enquanto que 85,71% dos indivíduos do grupo 2 apresentam risco em pelo menos 4 fatores. Portanto, o grupo 2 é o grupo dos outliers, o que sugere que os indivíduos desse grupo tenham um acompanhamento especial e prioritário, pois os mesmos apresentam maior risco de desenvolverem doenças cardiovasculares, assim como maior chance de ocorrência de acidentes de trabalho.

Figura 2: Gráficos de frequência das somas dos fatores de risco para a amostra completa, o Grupo 1 e o Grupo 2. Em cada gráfico, o eixo y representa o número de indivíduos e o eixo x o número de fatores de risco. Sobre a barra estão os valores percentuais de indivíduos que apresentam determinado fator de risco



4 Conclusões

Os resultados apresentados nas simulações mostram que o método proposto foi superior ao método 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, quando as variáveis são correlacionadas, verificou-se que o método proposto apresenta resultados ainda melhores comparado ao método de Mahalanobis.

Em relação à aplicação da metodologia proposta em dados de saúde, foi possível ilustrar o funcionamento do método e mostrar que o mesmo pode ser utilizado para análise de outliers em dados reais correlacionados.

Referências

Aggarwal, C. C. (2013). An introduction to outlier analysis. Em: *Outlier Analysis*, Springer, pp. 1–40.

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

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

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

Filzmoser, P. A. (2004). A multivariate outlier detection method. Em: *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, vol 1, pp. 18–22.

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

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

Oliveira, P. T. M. S., Santos, J. O., Munita, C. S. (2006). Identificação de valores discrepantes por meio da distância mahalanobis. Em: *XVII Simpósio Nacional de Probabilidade e Estatística*.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`.

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

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

Veloso, M. V. S., 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.

Josino José Barbosa

Universidade Federal de Viçosa, MG, Brasil
E-mail: <josinojba@gmail.com

Participação do autor:

-

Tiago Martins Pereira

Universidade Federal de Ouro Preto, MG, Brasil
E-mail: tiago.martin@iceb.ufop.br

Participação do autor:

-

Fernando Luiz Pereira de Oliveira

Universidade Federal de Ouro Preto, MG, Brasil
E-mail: fernando@iceb.ufop.br

Participação do autor:

-