JOSINO JOSÉ BARBOSA

IDENTIFICAÇÃO DE *OUTLIERS* MULTIVARIADOS - UMA APLICAÇÃO EM DADOS DE SAÚDE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2017

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
B238i
2017

Barbosa, Josino José, 1985-
Identificação de *outliers* multivariados - Uma aplicação em dados de saúde / Josino José Barbosa. – Viçosa, MG, 2017.
xii, 51f. : il. (algumas color.) ; 29 cm.

Inclui anexos.
Orientador: Fernando Luiz Pereira de Oliveira.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Análise multivariada. 2. Análise por agrupamento. 3. Monte Carlo, Método de. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 519.535

JOSINO JOSÉ BARBOSA

IDENTIFICAÇÃO DE OUTLIERS MULTIVARIADOS - UMA APLICAÇÃO EM DADOS DE SAÚDE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 17 de fevereiro de 2017.

Frederico Rodrigues Borges da Cruz

Tiago Martins Pereira
(Coorientador)

Paulo Roberto Cecon

Fernando Luiz Pereira de Oliveira
(Orientador)

Dedico este trabalho à minha esposa Ana Regina e à minha família.

Agradecimentos

Agradeço aos meus pais, Geralda e José, pelo amor, carinho e por toda a dedicação.

Aos meus irmãos, Josãne e Josias, pelo apoio, incentivo e por tudo que fizeram por mim.

À minha esposa, Ana Regina, pelo amor e apoio nos momentos difíceis.

À Universidade Federal de Viçosa, em especial aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, pelos ensinamentos que contribuíram para a minha formação acadêmica.

Aos secretários da pós-graduação Carla Zinato e Júnior Pires, pelo empenho e dedicação em nos atender.

Aos professores Fernando Luiz Pereira de Oliveira e Tiago Martins Pereira, pela confiança depositada durante o desenvolvimento do mestrado. Obrigado pela orientação, dedicação, incentivo e por todo o aprendizado.

Aos professores Frederico Rodrigues Borges da Cruz e Paulo Roberto Cecon, por terem aceitado o convite para participar da banca.

À FAPEMIG pelo apoio financeiro para o desenvolvimento deste trabalho.

Agradeço especialmente a Deus por ter me proporcionado alcançar mais essa conquista e por ter colocado todas estas pessoas no meu caminho.

O sucesso é uma consequência e não um objetivo.

Gustave Flaubert

Lista de figuras

Lista de tabelas

RESUMO

BARBOSA, Josino José, M.Sc., Universidade Federal de Viçosa, fevereiro de 2017. **Identificação de *outliers* multivariados - Uma aplicação em dados de saúde**. Orientador: Fernando Luiz Pereira de Oliveira. Coorientador: Tiago Martins Pereira.

A identificação de *outliers* desempenha um papel importante na análise estatística, pois tais observações podem conter informações importantes em relação aos dados. 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. Por isso, a finalidade 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*. Além disso, a metodologia proposta foi ilustrada com uma aplicação em dados reais provenientes da área de saúde.

ABSTRACT

BARBOSA, Josino José, M.Sc., Universidade Federal de Viçosa, February, 2017. **Outlier Identification Multivariate - An application for health data**. Advisor: Fernando Luiz Pereira de Oliveira. Co-Advisor: Tiago Martins Pereira.

The identification of outliers plays an important role in statistical analysis, as such observations may contain important information regarding the data. 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. Therefore, the purpose 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 identifying outliers via Mahalanobis Distance. For data generation, the Monte Carlo method and the mixed-multivariate normal distribution technique were used. 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. In addition, the proposed methodology was illustrated with an application in real data from the health area.

Sumário

1 Introdução

1.1 Motivação

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: Em muitos sistemas de computador, inseridos em uma única máquina ou em rede, diferentes tipos de dados são recolhidos através das chamadas do sistema operacional, do tráfego da rede, ou de outras atividades no sistema. Estes dados podem mostrar um comportamento anormal devido à atividades maliciosas. A detecção de tais atividades é referida como a detecção de intrusão.

- Fraude de cartão de crédito: a fraude de cartão de crédito é bastante prevalente, devido à facilidade com que informações confidenciais, tal como o número de cartão de crédito, podem ser comprometidas. Isso normalmente leva a utilização não autorizada do cartão de crédito. Em muitos casos, o uso não autorizado pode mostrar padrões diferentes, como uma onda de compras a partir de locais geograficamente obscuros. Tais padrões podem ser utilizados para detectar valores extremos nos dados de transação do cartão de crédito.

- Sensores de eventos: Sensores são muitas vezes utilizados para rastrear vários parâmetros ambientais e locais em muitas aplicações reais. As mudanças bruscas nos padrões subjacentes podem representar eventos de interesse. A detecção de eventos é uma das aplicações que mais motivaram o campo da rede de sensores.

- Diagnóstico médico: Em muitas aplicações médicas os dados são recolhidos a partir de uma variedade de dispositivos, tais como exames de ressonância magnética, tomografia ou eletrocardiograma. Padrões incomuns em tais dados normalmente refletem as condições de uma doença.

- Polícia: A detecção de *outlier* possui várias utilidades para a aplicação da lei, especialmente em casos onde padrões incomuns só podem ser descobertos ao longo do tempo através de várias ações de um indivíduo ou entidade. Determinar fraudes em transições financeiras, atividades comerciais ou reivindicação de seguros normalmente requer a determinação de padrões incomuns nos dados gerados pelas ações da entidade criminal.

- Ciências da Terra: Uma quantidade significativa de dados em um espaço temporal sobre padrões climáticos, mudanças climáticas ou padrões de ocupação do solo é recolhida através de uma variedade de mecanismos, tais como satélites ou sensoriamento remoto. Anomalias em tais dados fornecem uma visão significativa sobre as tendências humanas ou ambientais ocultas que podem ter causado tais anomalias.

Ainda segundo Aggarwal (2013), em todas estas aplicações os dados têm um modelo "normal" e anomalias são reconhecidas como desvios deste modelo normal. Em muitos casos, tais como detecção de intrusão ou fraude, os valores atípicos só podem ser descobertos com uma sequência de várias observações de dados, ao invés de uma observação individual. Por exemplo, um evento de fraude pode muitas vezes refletir as ações de um indivíduo em uma determinada sequência. A especificidade da sequência é relevante para identificar o evento anômalo. Tais anomalias são também referidas como anomalias coletivas, porque só podem ser inferidas a partir de um conjunto ou sequência de dados. Tais anomalias coletivas geralmente representam eventos incomuns que precisam ser descobertos a partir dos dados.

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, Maronna e Werner (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, Aquino e Junior (2012) propuseram um trabalho que apresenta uma análise, via detecção de *outliers*, sobre dados multivariados proveniente de rede de sensores.

1.2 Principais Contribuições

Diante da importância da identificação de *outliers*, 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;
- Apresentar uma aplicação da técnica desenvolvida em uma base de dados reais.

1.3 Organização

O restante desta dissertação está organizada conforme se segue.

Como forma de auxiliar o leitor, a segunda seção apresenta uma breve revisão bibliográfica sobre os assuntos e técnicas que serão abordados no decorrer do trabalho. A

terceira seção apresenta a metodologia utilizada na geração das populações contaminadas pelos *outliers*, assim como uma descrição do método proposto e algumas medidas utilizada para testar a eficiência do método. A quarta seção apresenta os resultados das simulações e a comparação do método proposto com o de Mahalanobis, além de uma breve discussão. Apresenta ainda uma aplicação da metodologia proposta em dados reais provenientes da área de saúde. E por fim, a quinta seção apresenta as considerações finais sobre a dissertação e indica alguns possíveis trabalhos futuros.

2 Revisão Bibliográfica

2.1 Outlier

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.

O tratamento de *outliers*, seja para identificação, remoção ou ambos, tem sido extensivamente pesquisado em diversas áreas do conhecimento, tais como a estatística, a mineração de dados, o aprendizado de máquina e a teoria da informação. Algumas das aplicações que se beneficiam do tratamento de *outliers* são a identificação de fraudes, a intrusão de redes, a análise de desempenho, a previsão do tempo, entre outras (CHANDOLA; BANERJEE; KUMAR, 2009).

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).

Muitas das primeiras propostas para a identificação de *outliers* multivariados referem-se a métodos baseados em análises gráficas. Uma das contribuições mais importantes deve-se a Gnanadesikan e Kettenring (1972) que trabalharam com detecção de *outliers* em dados de resposta múltipla. Entretanto, para grandes valores de p é preciso considerar a possibilidade de que *outliers* irão se manifestar em direções diferentes das que são detectáveis na confecção de um gráfico simples de pares de variáveis originais.

Um método clássico para o reconhecimento de *outliers* é a distância de Mahalanobis, que utiliza como estimadores de locação e dispersão, a média aritmética simples e a matriz

de covariância amostral, respectivamente (HAZEWINKEL, 1995).

2.2 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.

2.3 Distribuição Normal

Introduzida em 1733 pelo matemático Abraham de Moivre e posteriormente redescoberta por Laplace (1774) e Gauss (1809), por isso também conhecida como distribuição gaussiana, a distribuição normal é a mais importante das distribuições de probabilidade.

Uma aplicação da distribuição normal é na aproximação probabilística de variáveis aleatórias binomiais quando o número n de experimentos de Bernoulli é suficientemente grande. Isso decorre do Teorema Central do Limite, o qual afirma que a média de uma amostra de n elementos de uma população aproxima-se de uma distribuição gaussiana. Muitas variáveis contínuas que descrevem fenômenos naturais e sociais apresentam distribuições de probabilidades próximas da distribuição normal.

A distribuição de Gauss serve de base teórica para a estatística inferencial e tem como característica a média e o desvio padrão, em que a média indica a posição central da distribuição e o desvio padrão refere-se à dispersão. Uma variável aleatória contínua X tem distribuição normal se sua função densidade de probabilidade for dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \qquad -\infty < x < \infty. \tag{2.1}$$

em que μ é a média e σ é o desvio padrão da distribuição.

O gráfico da distribuição tem o formato de um sino e depende dos valores dos parâmetros μ e σ. A Figura 1 exemplifica o gráfico da distribuição normal:

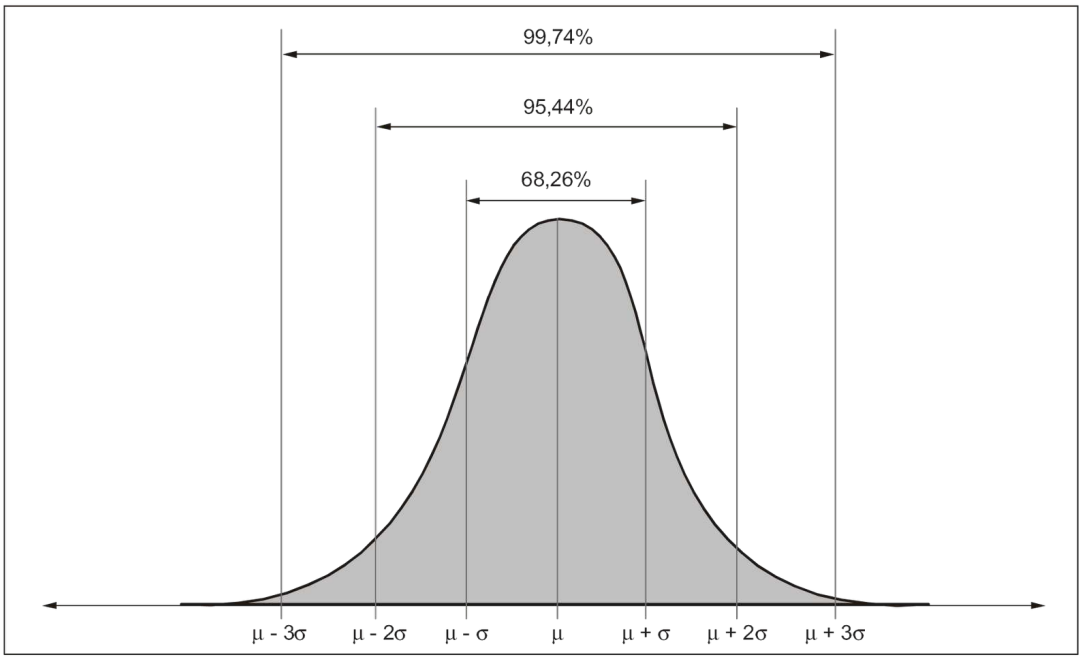

Figura 1: Gráfico da Distribuição Normal

2.4 Distribuição Normal Multivariada

A distribuição normal multivariada é uma generalização para várias dimensões da densidade normal univariada. No campo de estudos da análise multivariada, a distribuição normal para $p \geq 2$ dimensões desempenha um papel muito importante, já que esta distribuição representa uma aproximação adequada de distribuições populacionais e dados experimentais, além de ser utilizada em várias áreas como engenharia, psicologia e economia, e de servir para descrever qualquer conjunto de variáveis aleatórias de valores reais correlacionados.

Dizemos que o vetor $\boldsymbol{X}' = [X_1, X_2, ..., X_p]$ tem distribuição normal multivariada se sua função densidade de probabilidade é dada por:

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'|\boldsymbol{\Sigma}|^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{2.2}$$

em que $\boldsymbol{\mu}$ é o vetor de médias, $\boldsymbol{\Sigma}$ a matriz de covariâncias e p é o índice da dimensão da distribuição normal p-variada e indica o número de variáveis envolvidas.

O termo $\left(\frac{X-\mu}{\sigma}\right)^2$ é generalizado para $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, que é definida como a distância de Mahalanobis. Quando $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, ou seja, tem distribuição normal multivariada, então cada um dos elementos de $\boldsymbol{X}$ tem distribuição normal univariada.

A Figura 2 mostra o gráfico da distribuição normal multivariada para o caso bidimensional, ou seja, $p = 2$.

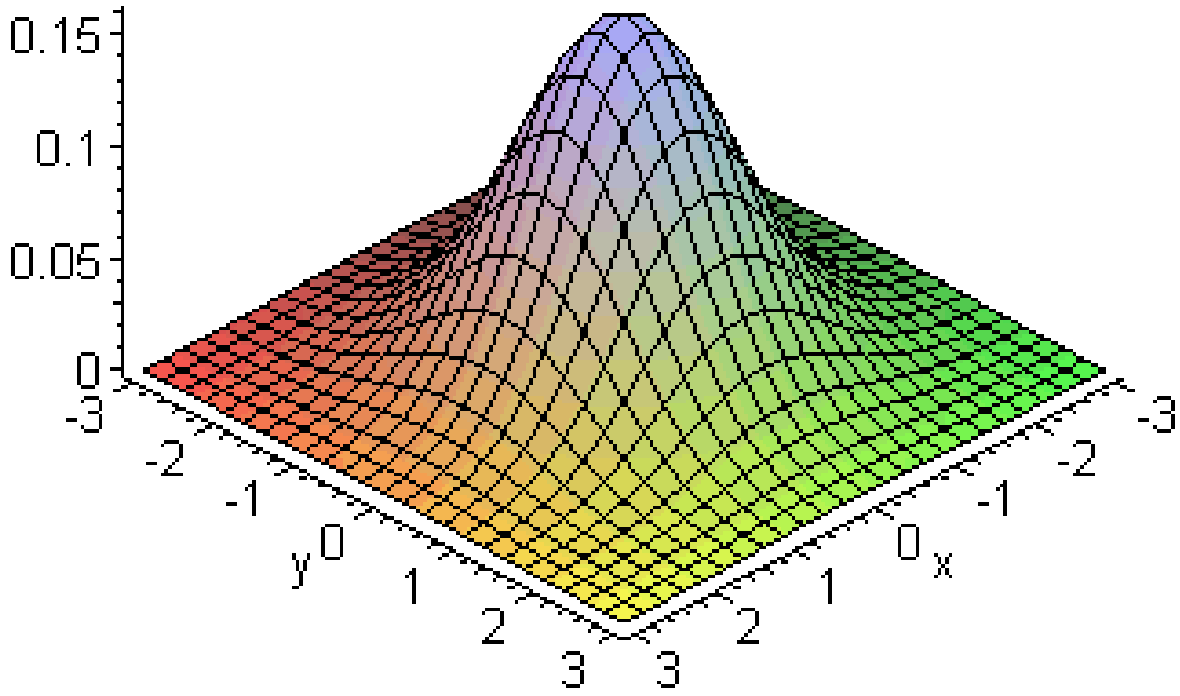

Figura 2: Gráfico da Distribuição Normal Multivariada

2.5 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:

$$f(\boldsymbol{x}) = (1-\delta)(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma_1}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})'|\boldsymbol{\Sigma_1}|^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})\right] +$$
$$+\delta(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma_2}|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_2})'|\boldsymbol{\Sigma_2}|^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2})\right] \quad (2.3)$$

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

2.6 Análise de Agrupamentos (AA)

Segundo Malhotra (2012), a análise de agrupamento, ou análise de *clusters*, é uma técnica usada para classificar objetos ou casos em grupos relativamente homogêneos, chamados de agrupamentos ou conglomerados. Assim, os objetos em cada agrupamento tendem a ser semelhantes entre si, mas diferentes de objetos em outros agrupamentos.

Conforme Hair et al. (2009), as características de cada objeto são combinadas em uma medida de semelhança, que pode ser de similaridade ou dissimilaridade, calculada para todos os pares de objetos, possibilitando a comparação de qualquer objeto com outro pela medida de similaridade e a associação dos objetos semelhantes por meio da análise de agrupamento. As medidas de distância representam a similaridade, que indica a proximidade entre as observações.

De uma maneira geral, os métodos de agrupamento são divididos em:

- Métodos hierárquicos aglomerativos e divisivos: Envolvem a construção de uma hierarquia através de uma estrutura do tipo árvore. Nos métodos hierárquicos aglomerativos os agrupamentos são formados a partir de uma matriz de parecença, que é atualizada a cada união de um par de objetos. Neste procedimento, os objetos individuais vão se juntando sucessivamente e uma vez que dois elementos são unidos eles permanecem juntos até o final do processo. Já os métodos hierárquicos divisivos fazem o caminho oposto aos métodos hierárquicos aglomerativos. Neste método, um

único grupo de objetos é particionado sucessivamente até se obter os objetos individuais. Além disso, uma vez que dois elementos são separados eles jamais voltarão a fazer parte do mesmo agrupamento.

- Métodos não-hierárquicos: Produzem uma partição em um número fixo de classes, sendo necessário definir o número de grupos à priori. Esses métodos tem por objetivo determinar a classificação dos n indivíduos em k grupos que otimize algum critério de homogeneidade interna e heterogeneidade externa.

2.6.1 Método *k*-médias

O método k-médias é um método de partição não-hierárquico que fornece indicações mais precisas sobre o número de conglomerados a ser formado. Este método talvez seja um dos mais utilizados quando se têm muitos objetos para agrupar, com pequenas variações. O critério mais utilizado de homogeneidade dentro do grupo e heterogeneidade entre os grupos é o da soma dos quadrados residuais, baseado na análise de variância. Assim, quanto menor for este valor, mais homogêneos são os elementos dentro de cada grupo e melhor será a partição (BUSSAB, 1990). Nesse método, devemos conhecer a priori o número de grupos k e os n objetos são aglomerados nos k grupos.

Segundo Ferreira (2011), o método k-médias é baseado nos seguintes passos, utilizando uma das mais simples entre as várias possibilidades de se implementar esse algoritmo:

a) Alocar arbitrariamente os n objetos aos k grupos e calcular os centroides. Alternativamente, podem-se gerar os centroides de cada grupo por um processo aleatório qualquer. Se o centroide foi obtido por esse último processo, alocar cada um dos n objetos aos grupos que apresentam a menor distância euclidiana com o respectivo objeto. A distância entre o objeto e o grupo é obtida, em geral, pela distância euclidiana entre o vetor de observações do objeto e o centroide do grupo.

b) Recalcular os centroides de cada grupo.

c) Realocar o primeiro objeto de seu grupo para um outro grupo em que a distância euclidiana seja mínima e, por razões óbvias, menor do que a distância euclidiana desse objeto para o seu próprio grupo de origem.

d) Repetir os passos (b) e (c) até que não ocorram mais mudanças de objetos de um grupo para outro. Nesse processo é realizada apenas a transferência por iteração.

A Figura 3 ilustra o funcionamento do método de agrupamento k-médias para o caso em que $k = 2$.

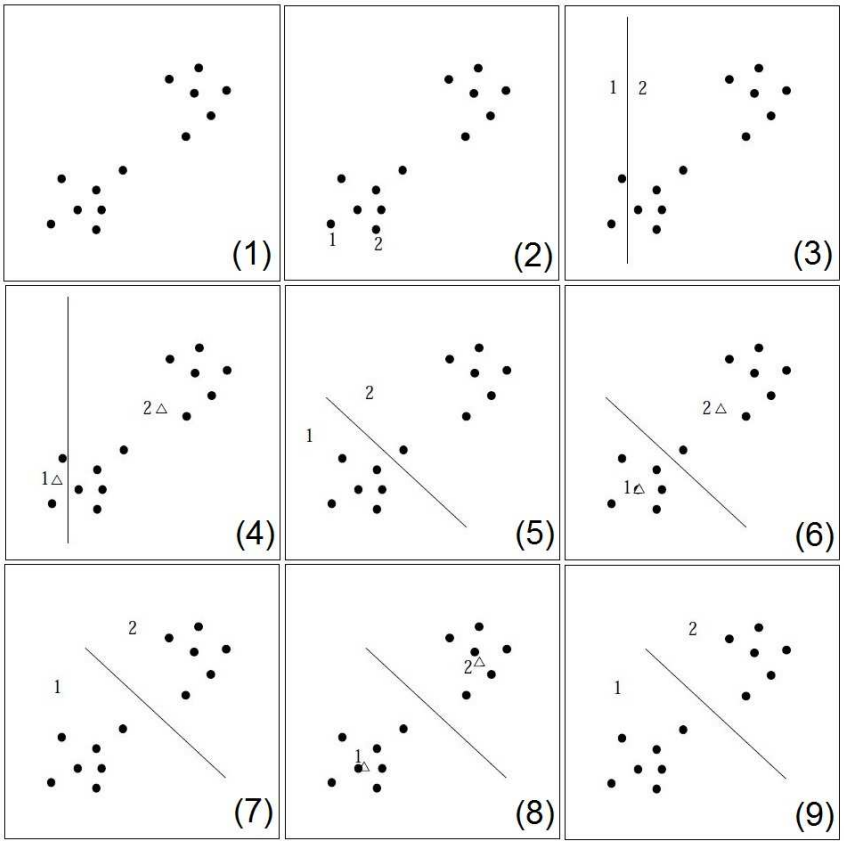

Figura 3: Representação gráfica do método de agrupamento k-médias

Abaixo, segue uma breve descrição da Figura 3:

- Quadro 1: ilustra o conjunto de dados;
- Quadro 2: escolha arbitrária dos dois centroides;
- Quadro 3: agrupa os objetos nos dois grupos;
- Quadro 4: recalcula os centroides;
- Quadro 5: reagrupa os objetos;
- Quadro 6: recalcula os centroides;
- Quadro 7: reagrupa os objetos;
- Quadro 8: recalcula os centroides;
- Quadro 9: partição final.

2.7 Distâncias

A distância entre dois objetos é usada para medir quão semelhantes esses objetos são. Em muitas situações podemos usar distâncias para agrupar objetos ou indivíduos semelhantes.

A distância entre dois vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$ pode ser definida como uma função d que associa um número real positivo $d(\boldsymbol{x}, \boldsymbol{y})$, com as seguintes propriedades:

i. $d(\boldsymbol{x}, \boldsymbol{y}) = 0$, se e somente se, $\boldsymbol{x}$=$\boldsymbol{y}$

ii. $d(\boldsymbol{x}, \boldsymbol{y}) > 0 \; \forall \; \boldsymbol{x} \neq \boldsymbol{y}$

iii. $d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{y}, \boldsymbol{z}) \; \forall \; \boldsymbol{x}, \boldsymbol{y}$ e $\boldsymbol{z}$, onde $\boldsymbol{z}$ é um vetor qualquer do $\Re^p$.

Considere dois vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$ e a matriz positiva definida $\boldsymbol{A}$. O grau de afastamento das observações, ou seja, a distância, chamada de métrica, é utilizada também para detecção de *outliers*. A expressão para a distância quadrática entre os vetores $\boldsymbol{X}$ e $\boldsymbol{Y}$ é dada analiticamente por:

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{X} - \boldsymbol{Y})' \boldsymbol{A} (\boldsymbol{X} - \boldsymbol{Y}) \tag{2.4}$$

As medidas de distância podem ser definidas como medidas de similiridade e dissimilaridade, na qual a primeira é para definir o grau de semelhança e a segunda para medir o grau de diferença. O coeficiente de correlação é uma medida de similaridade e quanto maior for essa medida maior a semelhança entre os indivíduos. Entre outras medidas de dissimilaridade que são encontradas na literatura, como a distância de Manhattan e a distância de Minkowsky, nesse estudo serão dadas ênfase nas distâncias euclidiana e euclidiana ponderada.

2.7.1 Distância Euclidiana

A distância euclidiana é a medida de distância mais frequentemente empregada quando todas as variáveis são quantitativas. A distância euclidiana é utilizada para calcular medidas específicas, assim como a distância euclidiana simples e a distância euclidiana quadrática ou absoluta.

A distância euclidiana é dada por:

$$d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{X} - \boldsymbol{Y})'(\boldsymbol{X} - \boldsymbol{Y}) \tag{2.5}$$

Essa distância é a mais simples entre dois vetores e deve ser utilizada quando a métrica for definida por:

$$\boldsymbol{A} = \boldsymbol{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{2.6}$$

A distância euclidiana entre dois pontos é dada por:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2.7}$$

A Figura 4 representa essa medida de dissimilaridade.

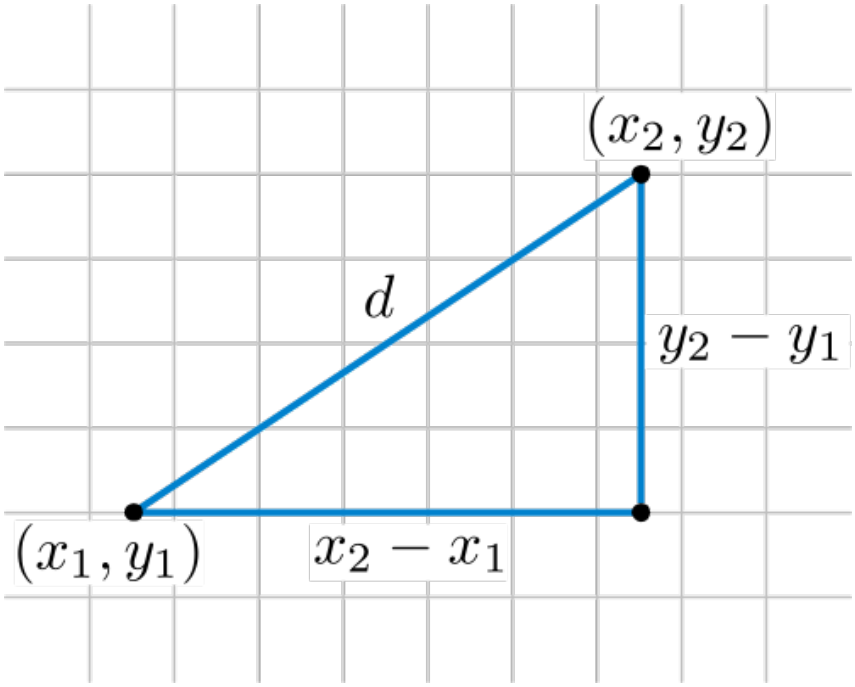

Figura 4: Representação gráfica da distância euclidiana entre dois pontos

2.8 Desvio Padrão

O desvio padrão é uma medida de dispersão dos dados em relação à média. Um baixo desvio padrão indica que os dados tendem a estar próximos da média, enquanto que um desvio padrão alto indica que os dados estão espalhados por uma grande gama de valores.

Dada uma amostra aleatória $x_1, x_2, ..., x_n$, composta por n elementos, podemos calcular o desvio padrão amostral a partir da seguinte expressão:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.8}$$

em que $\bar{x}$ é a média amostral dada por:

$$\bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i \tag{2.9}$$

2.9 Análise de Correspondência

A análise de correspondência, técnica exploratória de simplificação da estrutura da variabilidade de dados multivariados, utiliza de variáveis categóricas dispostas em tabelas de contingência, levando em conta medidas de correspondência entre as linhas e colunas da matriz de dados.

Segundo LUCIO, TOSCANO e ABREU (1999) a análise de correspondência é um método para determinação de um sistema de associação entre os elementos de dois ou mais conjuntos, buscando explicar a estrutura de associação dos fatores em questão. Assim, são construídos gráficos com as coordenadas principais das linhas e das colunas permitindo a visualização da relação entre os conjuntos, em que a proximidade dos pontos referentes à linha e à coluna indica uma associação e o distanciamento desses pontos indica uma repulsão.

2.10 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:

$$MD_i = \sqrt{(\boldsymbol{x_i} - \bar{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})} \tag{2.10}$$

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

$$\boldsymbol{S} = \frac{\sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})'}{n-1} \tag{2.11}$$

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; 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 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 Material e Métodos

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:

$$\boldsymbol{\mu_1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \boldsymbol{\mu_2} = \begin{bmatrix} 2 \\ 2 \\ \vdots \\ 2 \end{bmatrix} \tag{3.1}$$

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

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{bmatrix} \tag{3.2}$$

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 5. No caso ilustrado, o grupo 4 seria definido como um grupo outlier, uma vez que $d(C4, Md) > 2,5s$.

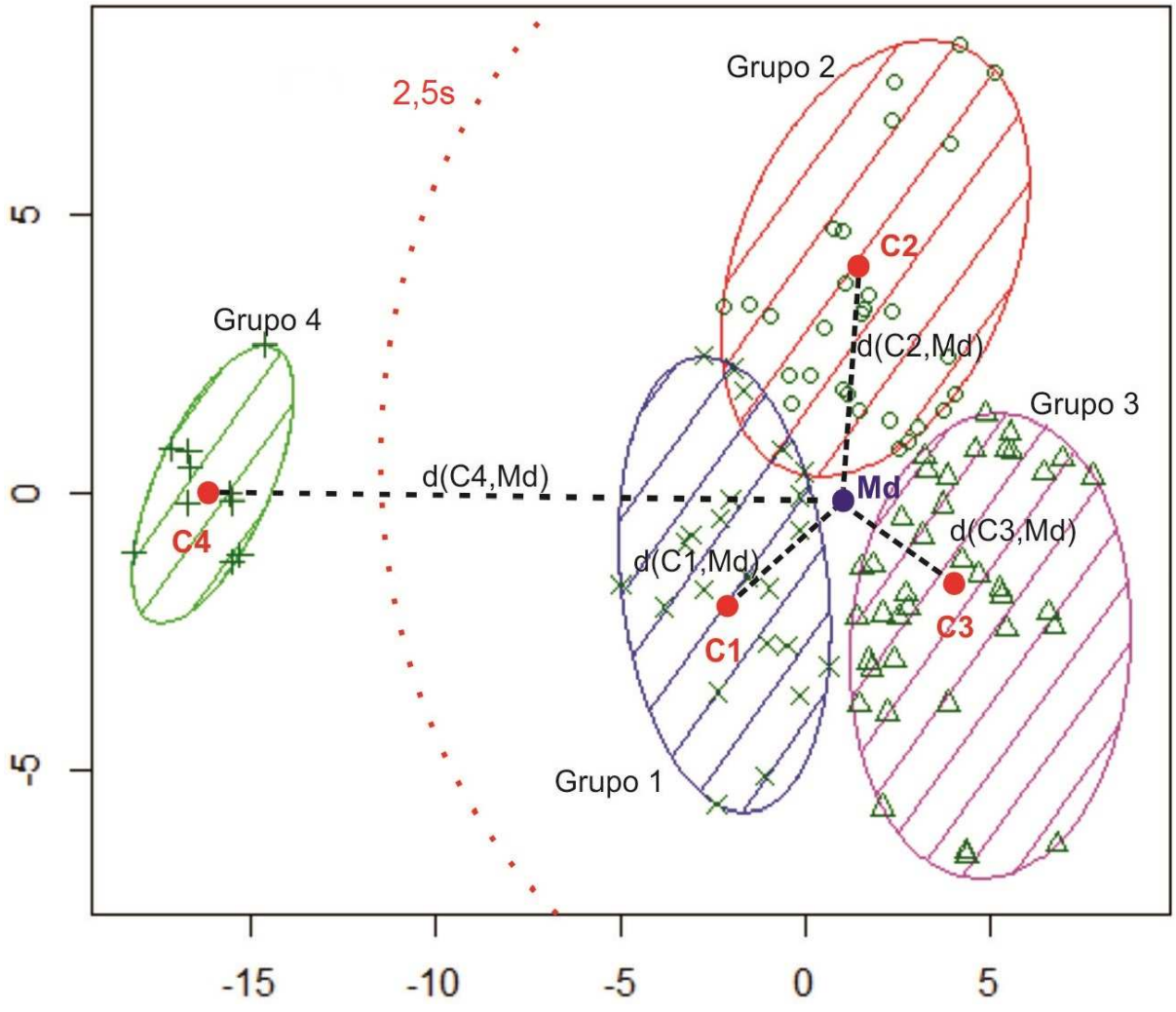

Figura 5: Representação gráfica do método

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 do método

Outlier / Método	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:

$$S = \frac{a}{a+c} \tag{3.3}$$

e

$$E = \frac{d}{b+d} \tag{3.4}$$

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.

Todos os cálculos e simulações foram feitos utilizando o software R Core Team (2014) por meio do desenvolvimento de um programa que se encontra listado nos anexos.

4 Resultados e Discussão

4.1 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: Resultado das simulações e 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
	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
	E	0,91076	0,91845	0,92614	0,88524	0,88777	0,8903	7,19116	0
$\rho = 0,7$	***S***	0,65727	0,68133	0,70539	0,23784	0,2567	0,27556	27,20583	0
	E	0,91542	0,92293	0,93044	0,88306	0,88584	0,88863	9,51673	0
$\rho = 0,9$	***S***	0,5197	0,5438	0,5679	0,19591	0,21239	0,22888	22,42424	0
	E	0,93168	0,93721	0,94275	0,88265	0,88516	0,88766	17,62282	0

Os resultados dos demais cenários hipotéticos encontram-se nos anexos.

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 pelos métodos proposto e de Mahalanobis.

Tabela 3: Resultado das simulações e comparação entre os dois métodos para sensibilidade

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 simulações e comparação entre os dois métodos para especificidade

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 mostram que, 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%. 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 para a sensibilidade levando-se em consideração a correlação existente entre as variáveis, a taxa de mistura atribuída, o tamanho da amostra e o número de variáveis, respectivamente.

Tabela 5: Resultado das simulações e comparação entre os dois métodos para sensibilidade levando-se em consideração a correlação (ρ) entre as variáveis

Resultado \ $\boldsymbol{\rho}$	**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**

Na Tabela 5 pode-se verificar que, considerando os 16 cenários em que $\rho = 0$, em 8 cenários as médias de sensibilidade dos dois métodos foram consideradas estatisticamente iguais, enquanto que nos demais níveis de ρ o método proposto foi superior em pelo menos 14 cenários.

Tabela 6: Resultado das simulações e comparação entre os dois métodos para sensibilidade levando-se em consideração a taxa de mistura (δ)

Resultado \ $\boldsymbol{\delta}$	**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 simulações e comparação entre os dois métodos para sensibilidade levando-se em consideração o tamanho da amostra (n)

Resultado \ $\boldsymbol{n}$	**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 simulações e comparação entre os dois métodos para sensibilidade 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 6 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.

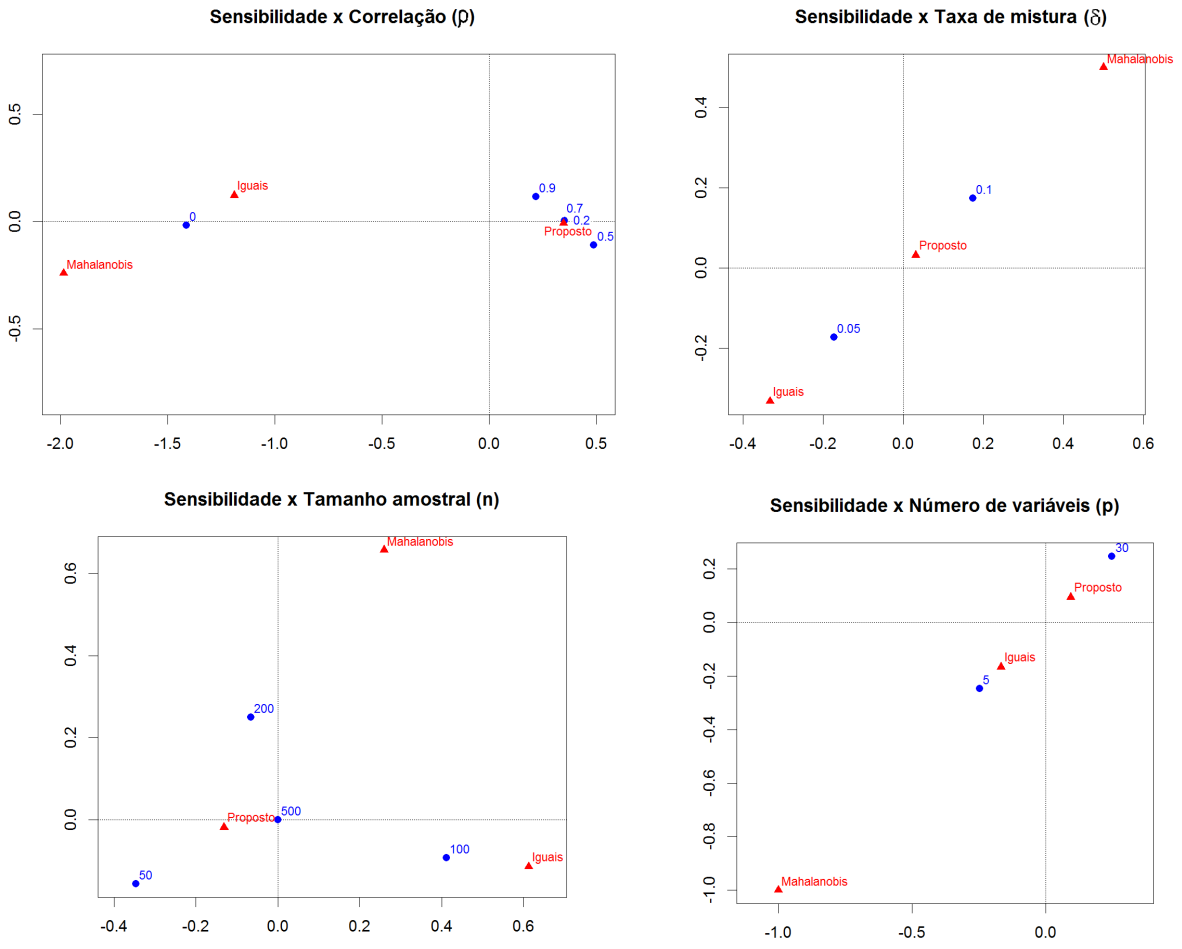

Figura 6: Gráficos de análise de correspondência: sensibilidade X valores paramétricos

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

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 simulações e comparação entre os dois métodos para especificidade levando-se em consideração a correlação (ρ) entre as variáveis

Resultado \ $\boldsymbol{\rho}$	**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 simulações e comparação entre os dois métodos para especificidade levando-se em consideração a taxa de mistura (δ)

Resultado \ $\boldsymbol{\delta}$	**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 simulações e comparação entre os dois métodos para especificidade levando-se em consideração o tamanho da amostra (n)

Resultado \ $\boldsymbol{n}$	**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 simulações e comparação entre os dois métodos para especificidade levando-se em consideração o número de variáveis (p)

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

A Figura 7 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.

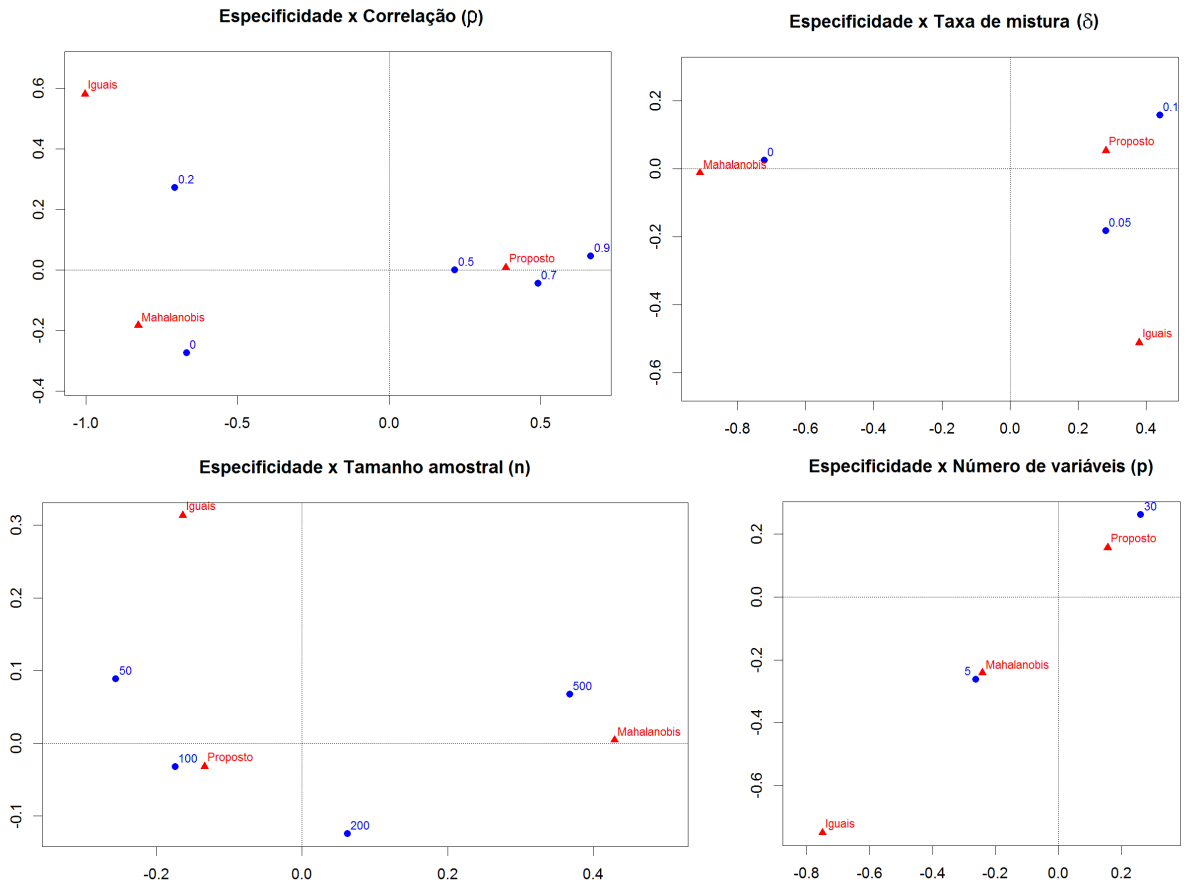

Figura 7: Gráficos de análise de correspondência: especificidade X valores paramétricos

Assim como ocorreu para a sensibilidade, analisando as Tabelas 9, 10, 11, 12 e a Figura 7, 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.

4.2 Análise de Dados Reais

A metodologia proposta, apresentada na seção 3.2, 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. A coleta dos dados foi realizada nos ambulatórios das minas e foi realizado um teste piloto para a realização 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:

- PAS e PAD: A pressão arterial é a pressão que o sangue exerce na parede das artérias. Ela é medida em milímetros de mercúrio (mmHg) e em duas etapas: pressão arterial sistólica (PAS) e pressão arterial diastólica (PAD). A PAS é quando o coração se contrai e é também conhecida como pressão máxima. Já a PAD é quando ele se dilata e é também conhecida como pressão mínima. O valor ótimo de pressão arterial é 120 x $80 mmHg$ (12 por 8) e indivíduos com pressão acima de 140 x $90 mmHg$ são considerados hipertensos.

- 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 normais de glicose no sangue de um indivíduo devem estar entre 60 e $100 mg/dL$. 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.

- Vitamina D: A vitamina D é fundamental para o equilíbrio do cálcio e do fósforo no organismo e para a saúde do esqueleto. A deficiência dessa vitamina prejudica a mineralização óssea em todas as fases da vida, causando o raquitismo em crianças e a osteoporose e outras doenças em adultos. Indivíduos com valores de vitamina D no sangue inferiores a 30 nanogramas por mililitros (ng/mL) estão propensos a apresentar os problemas relatados anteriormente.

- Colesterol não-HDL: O colesterol não-HDL é a soma de todos os tipos de colesterol considerados ruins: IDL+LDL+VLDL. Supõe-se que o colesterol não-HDL seja um

marcador mais sensível de risco de aterosclerose (acumulo de gordura na parede dos vasos sanguíneos) do que o LDL isoladamente. Valores superiores a $200mg/dL$ de colesterol não-HDL são considerados ruins.

- HDL: O 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 $40mg/dL$ são considerados ruins.

- LDL: O 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 $130mg/dL$ são considerados elevados.

- 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 $150mg/dL$ são considerados ruins.

- CC, CQ e 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. A circunferência da cintura (CC) é uma medida obtida aferido-se a circunferência do abdômen na altura do umbigo. Já a circunferência do quadril (CQ) é uma medida obtida aferido-se a circunferência do quadril. A RCQ é a relação cintura - quadril e quanto mais alto for o valor dessa relação maior é o risco à saúde. Para homens, o índice de corte para risco cardiovascular é $0,90$, ou seja, indivíduos com RCQ maior que $0,90$ estão propensos a esse tipo de risco.

- CP: A circunferência do pescoço (CP) tem sido utilizada por ser uma medida simples, que possibilita a identificação de sobrepeso e de obesidade. A CP aumentada leva a um acúmulo de moléculas de gordura na parede das artérias carótidas, favorecendo o desenvolvimento de doenças cardiovasculares. Indivíduos com CP acima de $40cm$ são considerados elevados.

Com o objetivo de se identificar os indivíduos que apresentam maior possibilidade de desenvolverem doenças cardiovasculares, tais como Infarto do Miocárdio e Acidente Vascular Cerebral (AVC), e que consequentemente apresentam maior chance de ocorrência de acidentes de trabalho, o conjunto de dados foi submetido ao método proposto nesse trabalho.

Para verificar se existe correlação entre as variáveis construiu-se a matriz de correlações e, com base na mesma, o gráfico de correlações, conforme ilustra a Figura 8. Nesse gráfico, quanto mais correlacionadas forem as variáveis mais achatada será a elipse. Além disso, quanto mais próximo de 1 for a correlação mais azul escuro será a elipse e quanto mais próximo de -1 mais vermelho a elipse será.

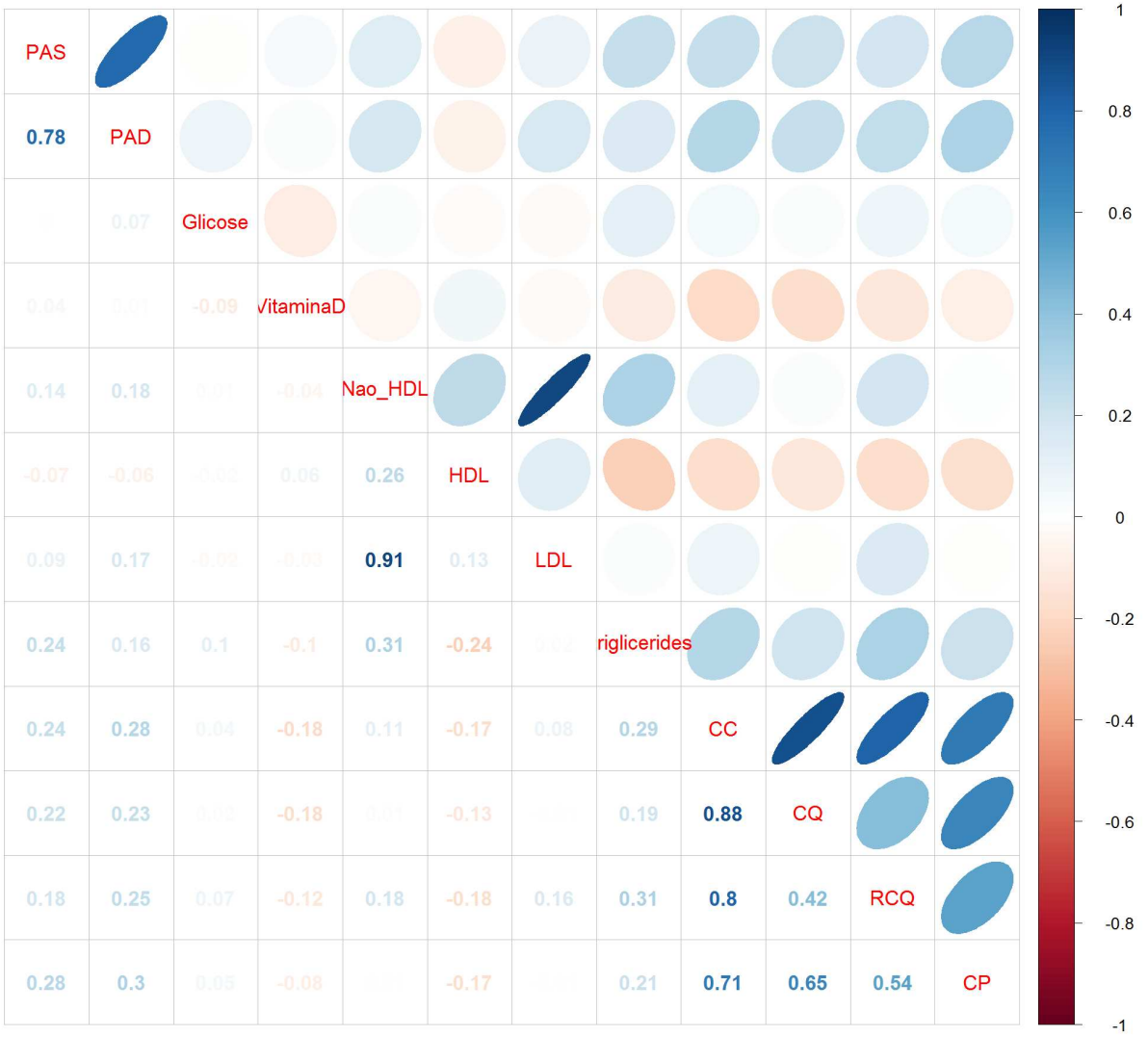

Figura 8: Gráfico de correlações entre as variáveis

Analisando a Figura 8 pode-se notar que existe correlação entre a maioria das variáveis. Diante da evidência de correlação entre as variáveis e dos resultados apresentados nas simulações, o método proposto parece ser indicado para a análise desse conjunto de dados.

A Tabela 13 apresenta uma análise descritiva das variáveis em estudo, contendo valor mínimo, mediana, média, máximo e desvio padrão para cada variável.

Tabela 13: Análise descritiva dos dados

Variáveis	Mínimo	Mediana	Média	Máximo	Desvio Padrão
PAS	102,000	130,500	130,500	195,500	13,110
PAD	59,500	81,000	81,890	117,000	9,519
Glicose	61,500	88,250	89,320	172,600	10,708
Vitamina D	13,900	24,210	25,160	41,080	5,833
Colesterol não-HDL	105,500	207,200	210,700	339,100	43,272
HDL	33,150	53,060	54,940	139,990	13,234
LDL	32,790	116,160	118,210	214,010	37,061
Triglicérides	45,100	168,200	187,700	392,000	79,910
CC	68,330	93,500	93,390	123,670	10,648
CQ	84,770	100,970	101,500	123,530	7,527
RCQ	0,730	0,920	0,918	1,070	0,057
CP	28,200	39,100	39,210	46,900	2,715

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 um pouco divergentes. Como o método baseia-se na distância euclidiana entre os centroides dos grupo e a mediana dos dados, indivíduos que apresentam tanto altos quanto baixos valores 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. As Tabelas 14 e 15 apresentam uma análise descritiva das variáveis dos dois grupos, contendo valor mínimo, mediana, média, máximo e desvio padrão para cada variável.

Tabela 14: Análise descritiva das variáveis do grupo 1

Variáveis	Mínimo	Mediana	Média	Máximo	Desvio Padrão
PAS	102,000	119,500	122,600	150,500	12,343
PAD	59,500	73,500	76,510	100,000	9,178
Glicose	72,200	83,750	89,300	172,600	17,939
Vitamina D	17,480	27,990	28,670	41,080	5,996
Colesterol não-HDL	119,700	204,800	204,800	301,300	51,417
HDL	35,840	59,700	64,500	139,990	19,423
LDL	48,210	111,100	117,180	203,540	40,691
Triglicérides	45,100	123,600	130,200	232,100	42,996
CC	68,330	79,850	79,460	99,200	6,131
CQ	84,770	92,560	92,790	106,100	4,706
RCQ	0,730	0,860	0,855	0,930	0,040
CP	30,500	36,300	36,370	41,100	2,166

Tabela 15: Análise descritiva das variáveis do grupo 2

Variáveis	Mínimo	Mediana	Média	Máximo	Desvio Padrão
PAS	113,500	140,200	143,000	195,500	15,820
PAD	75,000	90,500	91,790	117,000	10,093
Glicose	72,900	86,650	89,150	118,100	9,504
Vitamina D	14,140	22,890	24,620	36,280	5,409
Colesterol não-HDL	176,600	250,900	244,300	339,100	44,577
HDL	36,310	54,300	55,130	86,640	11,422
LDL	89,140	145,690	148,890	214,010	34,290
Triglicérides	96,400	192,800	201,400	351,100	75,851
CC	84,530	102,350	103,250	123,670	9,880
CQ	95,330	107,360	107,790	123,530	7,779
RCQ	0,840	0,960	0,956	1,040	0,046
CP	36,400	40,850	41,050	46,900	2,531

Analisando as Tabelas 14 e 15 pode-se verificar que o segundo grupo apresenta maiores valores, com exceção das variáveis Vitamina D e HDL, tanto para mediana quanto para média nas variáveis em estudo.

Considerando-se os limites de risco especificados na descrição das variáveis, a partir das 12 variáveis em estudo, tem-se 9 fatores de risco. Para facilitar a visualização da distinção dos grupos somou-se o número de fatores de risco de cada indivíduo e calculou-se as frequências dessas somas, que podem ser observadas na Figura 9. 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.

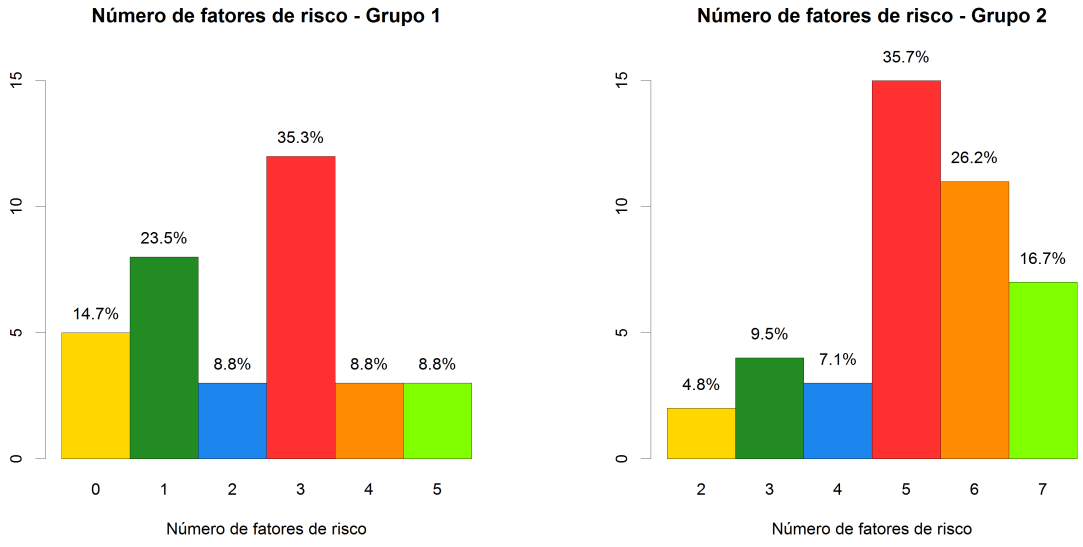

Figura 9: Gráficos de frequência das somas dos fatores de risco

Conforme pode ser visto na Figura 9, 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, pois os mesmos apresentam maior risco de desenvolverem doenças cardiovasculares, assim como maior chance de ocorrência de acidentes de trabalho.

5 Conclusões

5.1 Considerações Finais

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, 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.

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.

5.2 Trabalhos Futuros

Como sugestão para trabalhos futuros, pode-se destacar:

- Estudar a possibilidade de se utilizar a técnica de componentes principais, antes de aplicar o método proposto, para quebrar a estrutura de correlação das variáveis, de modo a converter um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas.
- Além do desvio padrão, buscar outras alternativas a serem utilizadas como critério para identificação dos *outliers*.

Referências

AGGARWAL, C. C. An introduction to outlier analysis. In: *Outlier Analysis*. [S.l.]: Springer, 2013. p. 1–40.

BAMNETT, V.; LEWIS, T. Outliers in statistical data. JSTOR, 1994.

BERTON, L. et al. Identifying abnormal nodes in complex networks by using random walk measure. In: IEEE. *IEEE Congress on Evolutionary Computation*. [S.l.], 2010. p. 1–6.

BUSSAB, W. de O. *Introdução à análise de agrupamentos*. [S.l.]: ABE, 1990. 105 p., 1990.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, ACM, v. 41, n. 3, p. 15, 2009.

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

CRITCHLEY, F. Influence in principal components analysis. *Biometrika*, Biometrika Trust, v. 72, n. 3, p. 627–636, 1985.

FERREIRA, D. F. *Estatística multivariada*. 2. ed. [S.l.]: UFLA, 2011.

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

GNANADESIKAN, R.; KETTENRING, J. R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, JSTOR, p. 81–124, 1972.

HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman Editora, 2009.

HAWKINS, D. M. *Identification of outliers*. [S.l.]: Chapman and Hall, 1980.

HAZEWINKEL, M. *Encyclopaedia of Mathematics*. [S.l.]: Kluwer Academic, 1995.

JOHNSON, M. E. Multivariate statistical simulation. In: *International Encyclopedia of Statistical Science*. [S.l.]: Springer, 2011. p. 930–932.

JOLLIFFE, I. *Principal component analysis*. [S.l.]: Wiley Online Library, 2002.

LUCIO, P. S.; TOSCANO, E. M. M.; ABREU, M. L. Caracterização de séries climatológicas pontuais via análise canônica de correspondência - estudo de caso. *Revista Brasileira de Geofísica*, v. 17, p. 41, 1999.

MALHOTRA, N. K. *Pesquisa de marketing: uma orientação aplicada*. [S.l.]: Bookman Editora, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

ROUSSEEUW, P. J.; DRIESSEN, K. V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, v. 41, n. 3, p. 212–223, 1999. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/00401706.1999.10485670>.

ROUSSEEUW, P. J.; ZOMEREN, B. C. van. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, [American Statistical Association, Taylor & Francis, Ltd.], v. 85, n. 411, p. 633–639, 1990. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2289995>.

SOBOL, I. M. *A Primer for the Monte Carlo Method*. [S.l.]: CRC PRESS, 1994.

VALADARES, F. G.; AQUINO, A. L. L. de; JUNIOR, A. R. P. Detecção de outliers multivariados em redes de sensores. In: SBPO. *XLIV Simpósio Brasileiro de Pesquisa Operacional*. [S.l.], 2012.

VELOSO, M. V. de S.; CIRILLO, M. A. 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*, v. 38, n. 2, p. 193–200, 2016.

ANEXO A – Resultados das simulações

Tabela 16: Resultado das simulações e comparação entre os dois métodos considerando $p = 5$, $n = 50$ e $\delta = 0$ 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***	-	-	-	-	-	-	-	-
	E	0,19346	0,2562	0,31894	0,7881	0,8004	0,8127	-16,48484	0
$\rho = 0,2$	***S***	-	-	-	-	-	-	-	-
	E	0,34557	0,4106	0,47563	0,79294	0,8046	0,81626	-11,22272	0
$\rho = 0,5$	***S***	-	-	-	-	-	-	-	-
	E	0,73979	0,7748	0,80981	0,78656	0,7992	0,81184	-1,26617	0,20842
$\rho = 0,7$	***S***	-	-	-	-	-	-	-	-
	E	0,82427	0,8488	0,87333	0,79173	0,8042	0,81667	3,12241	0,00235
$\rho = 0,9$	***S***	-	-	-	-	-	-	-	-
	E	0,90253	0,9164	0,93027	0,78755	0,7988	0,81005	12,98336	0

Tabela 17: Resultado das simulações para $p = 5$, $n = 50$ e $\delta = 0,05$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	***S***	0,92851	0,95683	0,98516	0,95023	0,97117	0,9921	-0,87607	0,38311
	E	0,72456	0,78914	0,85373	0,81405	0,82867	0,84329	-1,22064	0,22512
$\rho = 0,2$	***S***	0,90539	0,93974	0,97408	0,79438	0,8514	0,90843	2,58201	0,01129
	E	0,74983	0,80766	0,86549	0,80956	0,82213	0,83469	-0,4824	0,63058
$\rho = 0,5$	***S***	0,76936	0,8225	0,87564	0,55758	0,63167	0,70575	4,51524	0,00002
	E	0,81937	0,85253	0,88569	0,80353	0,81584	0,82814	2,16184	0,03304
$\rho = 0,7$	***S***	0,62571	0,70133	0,77695	0,52859	0,60067	0,67274	2,10976	0,0374
	E	0,88278	0,90146	0,92015	0,79581	0,80975	0,82368	8,01273	0
$\rho = 0,9$	***S***	0,47343	0,554	0,63457	0,37021	0,43855	0,50688	2,49828	0,01413
	E	0,92486	0,93634	0,94782	0,79868	0,81074	0,82281	16,00171	0

Tabela 18: Resultado das simulações para $p = 5$, $n = 50$ e $\delta = 0,10$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,91142	0,93769	0,96396	0,80154	0,85066	0,89978	3,35825	0,00111
	E	0,90864	0,93887	0,96911	0,8338	0,84651	0,85923	5,84076	0
$\rho = 0,2$	S	0,82271	0,86152	0,90034	0,61503	0,67402	0,73302	5,75012	0
	E	0,87819	0,91269	0,94719	0,84267	0,85378	0,86488	3,29881	0,00135
$\rho = 0,5$	S	0,71121	0,75869	0,80617	0,48793	0,54621	0,6045	6,5962	0
	E	0,90954	0,92997	0,9504	0,83113	0,84364	0,85614	7,06807	0
$\rho = 0,7$	S	0,54994	0,61251	0,67509	0,43332	0,49588	0,55844	2,90864	0,00448
	E	0,9234	0,93855	0,9537	0,81642	0,82884	0,84127	11,81636	0
$\rho = 0,9$	S	0,43254	0,50204	0,57153	0,3358	0,39296	0,45013	2,57291	0,01157
	E	0,94117	0,95174	0,9623	0,8117	0,82376	0,83581	15,86898	0

Tabela 19: Resultado das simulações para $p = 5$, $n = 100$ e $\delta = 0$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	-	-	-	-	-	-	-	-
	E	0,10922	0,1566	0,20398	0,82772	0,8343	0,84088	-26,65423	0
$\rho = 0,2$	S	-	-	-	-	-	-	-	-
	E	0,27907	0,3306	0,38213	0,82843	0,8347	0,84097	-18,55336	0
$\rho = 0,5$	S	-	-	-	-	-	-	-	-
	E	0,75328	0,7802	0,80712	0,82948	0,8348	0,84012	-3,90273	0,00017
$\rho = 0,7$	S	-	-	-	-	-	-	-	-
	E	0,84291	0,8574	0,87189	0,82316	0,8291	0,83504	3,57088	0,00055
$\rho = 0,9$	S	-	-	-	-	-	-	-	-
	E	0,90211	0,9126	0,92309	0,83542	0,8411	0,84678	12,1247	0

Tabela 20: Resultado das simulações para $p = 5$, $n = 100$ e $\delta = 0,05$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,91711	0,94307	0,96902	0,93048	0,94917	0,96786	-0,47117	0,63855
	E	0,75587	0,81129	0,86671	0,85073	0,85681	0,86289	-1,62593	0,10714
$\rho = 0,2$	S	0,82945	0,86887	0,90829	0,78676	0,82504	0,86332	1,61112	0,11034
	E	0,79706	0,84189	0,88673	0,85149	0,85785	0,86422	-0,69456	0,48896
$\rho = 0,5$	S	0,68832	0,733	0,77768	0,58409	0,63018	0,67627	3,65156	0,00042
	E	0,89201	0,91206	0,93211	0,84623	0,85262	0,85902	5,62423	0
$\rho = 0,7$	S	0,57804	0,62804	0,67803	0,50678	0,56096	0,61514	2,0155	0,04656
	E	0,92632	0,93957	0,95282	0,84308	0,84865	0,85423	12,81923	0
$\rho = 0,9$	S	0,48802	0,54462	0,60123	0,42494	0,48126	0,53757	1,95499	0,0534
	E	0,92813	0,93989	0,95165	0,84111	0,84733	0,85354	14,19543	0

Tabela 21: Resultado das simulações para $p = 5$, $n = 100$ e $\delta = 0,10$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,89285	0,92255	0,95224	0,93464	0,95189	0,96914	-1,99811	0,04845
	E	0,89669	0,92006	0,94343	0,87807	0,88429	0,89052	3,05375	0,0029
$\rho = 0,2$	S	0,76885	0,80664	0,84443	0,72154	0,75672	0,7919	2,12085	0,03643
	E	0,88982	0,91228	0,93474	0,86697	0,87327	0,87957	3,37664	0,00105
$\rho = 0,5$	S	0,57863	0,62176	0,66489	0,47464	0,51497	0,55529	3,88707	0,00018
	E	0,92517	0,93784	0,95052	0,85647	0,86303	0,86959	11,33073	0
$\rho = 0,7$	S	0,46468	0,50925	0,55381	0,41876	0,46279	0,50682	1,61717	0,10902
	E	0,94011	0,95047	0,96083	0,85034	0,85718	0,86401	15,90493	0
$\rho = 0,9$	S	0,40451	0,44533	0,48616	0,38751	0,42404	0,46056	0,79485	0,4286
	E	0,95452	0,96342	0,97232	0,85542	0,86135	0,86729	23,00394	0

Tabela 22: Resultado das simulações para $p = 5$, $n = 200$ e $\delta = 0$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	-	-	-	-	-	-	-	-
	E	0,11635	0,1496	0,18285	0,8357	0,83865	0,8416	-39,35281	0
$\rho = 0,2$	S	-	-	-	-	-	-	-	-
	E	0,26708	0,30865	0,35022	0,83567	0,8386	0,84153	-24,67244	0
$\rho = 0,5$	S	-	-	-	-	-	-	-	-
	E	0,74393	0,7644	0,78487	0,83304	0,83625	0,83946	-6,81619	0
$\rho = 0,7$	S	-	-	-	-	-	-	-	-
	E	0,84408	0,85715	0,87022	0,8314	0,83465	0,8379	3,16469	0,00206
$\rho = 0,9$	S	-	-	-	-	-	-	-	-
	E	0,8983	0,90645	0,9146	0,83527	0,83855	0,84183	15,47447	0

Tabela 23: Resultado das simulações para $p = 5$, $n = 200$ e $\delta = 0,05$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,94652	0,95921	0,9719	0,9657	0,97676	0,98781	-2,53546	0,0128
	E	0,71858	0,75801	0,79745	0,86139	0,86557	0,86975	-5,3505	0
$\rho = 0,2$	S	0,86682	0,89448	0,92215	0,81231	0,84021	0,86811	3,00329	0,00338
	E	0,76439	0,80048	0,83656	0,85905	0,86265	0,86625	-3,36931	0,00108
$\rho = 0,5$	S	0,73504	0,76865	0,80227	0,58175	0,61559	0,64943	8,06119	0
	E	0,88103	0,89292	0,9048	0,85167	0,85543	0,85919	5,82838	0
$\rho = 0,7$	S	0,59529	0,63254	0,66979	0,51957	0,54897	0,57836	3,85122	0,00021
	E	0,90932	0,91921	0,9291	0,84878	0,85283	0,85688	12,10615	0
$\rho = 0,9$	S	0,4789	0,51757	0,55623	0,43193	0,46511	0,49829	2,14708	0,03423
	E	0,94067	0,9474	0,95414	0,84762	0,85133	0,85503	25,6265	0

Tabela 24: Resultado das simulações para $p = 5$, $n = 200$ e $\delta = 0,10$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,92643	0,94115	0,95587	0,94979	0,95983	0,96988	-2,34798	0,02086
	E	0,79911	0,83529	0,87146	0,8835	0,88783	0,89216	-2,85641	0,00522
$\rho = 0,2$	S	0,8195	0,84662	0,87375	0,73165	0,7572	0,78275	5,27145	0
	E	0,87029	0,89069	0,91109	0,87217	0,87626	0,88034	1,39994	0,16466
$\rho = 0,5$	S	0,66471	0,69607	0,72742	0,50888	0,53658	0,56428	8,51967	0
	E	0,90953	0,92081	0,93209	0,86423	0,86829	0,87235	8,74378	0
$\rho = 0,7$	S	0,54652	0,57688	0,60725	0,4123	0,43885	0,4654	6,75469	0
	E	0,9249	0,93412	0,94333	0,85877	0,86289	0,86701	14,57888	0
$\rho = 0,9$	S	0,39826	0,42983	0,4614	0,35773	0,38016	0,40259	2,94847	0,00398
	E	0,95308	0,95922	0,96536	0,85121	0,85555	0,85989	28,15672	0

Tabela 25: Resultado das simulações para $p = 5$, $n = 500$ e $\delta = 0$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	-	-	-	-	-	-	-	-
	E	0,1827	0,20128	0,21986	0,83186	0,83412	0,83638	-64,98063	0
$\rho = 0,2$	S	-	-	-	-	-	-	-	-
	E	0,28125	0,30458	0,32791	0,83077	0,83298	0,83519	-43,5717	0
$\rho = 0,5$	S	-	-	-	-	-	-	-	-
	E	0,70012	0,71388	0,72764	0,83194	0,8339	0,83586	-16,75659	0
$\rho = 0,7$	S	-	-	-	-	-	-	-	-
	E	0,80734	0,81646	0,82558	0,8317	0,83388	0,83606	-3,44683	0,00083
$\rho = 0,9$	S	-	-	-	-	-	-	-	-
	E	0,87092	0,87742	0,88392	0,83077	0,83292	0,83507	12,45174	0

Tabela 26: Resultado das simulações para $p = 5$, $n = 500$ e $\delta = 0,05$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,97045	0,97694	0,98343	0,975	0,98041	0,98582	-0,85889	0,39247
	E	0,63051	0,6656	0,70069	0,86311	0,86581	0,86851	-11,18862	0
$\rho = 0,2$	S	0,92684	0,93824	0,94964	0,83171	0,84766	0,86362	9,8131	0
	E	0,74091	0,7674	0,7939	0,85946	0,86188	0,8643	-7,0044	0
$\rho = 0,5$	S	0,80305	0,82323	0,84341	0,59241	0,61762	0,64284	15,27506	0
	E	0,85425	0,86696	0,87966	0,85097	0,85316	0,85535	2,17412	0,03208
$\rho = 0,7$	S	0,66565	0,68969	0,71373	0,52714	0,54776	0,56838	9,85141	0
	E	0,8954	0,903	0,91061	0,84843	0,85061	0,85278	13,23865	0
$\rho = 0,9$	S	0,56738	0,59159	0,61579	0,45417	0,47128	0,48838	8,20796	0
	E	0,92277	0,92894	0,9351	0,84449	0,84662	0,84874	25,17413	0

Tabela 27: Resultado das simulações para $p = 5$, $n = 500$ e $\delta = 0,10$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,94879	0,95614	0,9635	0,95893	0,96471	0,97049	-2,13103	0,03556
	E	0,78924	0,81518	0,84111	0,88799	0,89054	0,89308	-5,74747	0
$\rho = 0,2$	S	0,87081	0,88391	0,897	0,74326	0,75971	0,77616	12,74185	0
	E	0,85232	0,86716	0,88201	0,87604	0,87858	0,88112	-1,52241	0,13109
$\rho = 0,5$	S	0,70299	0,72024	0,73749	0,50755	0,52442	0,54128	17,30332	0
	E	0,9015	0,91001	0,91852	0,86346	0,86562	0,86777	10,23421	0
$\rho = 0,7$	S	0,60263	0,62583	0,64903	0,4454	0,46211	0,47883	15,00559	0
	E	0,9149	0,92198	0,92906	0,85791	0,86013	0,86235	16,65684	0
$\rho = 0,9$	S	0,4812	0,50066	0,52012	0,3881	0,40096	0,41382	9,17934	0
	E	0,94007	0,94483	0,94959	0,85372	0,85581	0,85791	35,03092	0

Tabela 28: Resultado das simulações para $p = 30$, $n = 50$ e $\delta = 0$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	-	-	-	-	-	-	-	-
	E	0,06882	0,126	0,18318	0,75653	0,7712	0,78587	-21,46419	0
$\rho = 0,2$	S	-	-	-	-	-	-	-	-
	E	0,56087	0,6142	0,66753	0,73024	0,7554	0,78056	-5,16632	0
$\rho = 0,5$	S	-	-	-	-	-	-	-	-
	E	0,79446	0,8204	0,84634	0,70903	0,7392	0,76937	3,93077 0	0,00016
$\rho = 0,7$	S	-	-	-	-	-	-	-	-
	E	0,86633	0,8826	0,89887	0,70096	0,7362	0,77144	7,19824	0
$\rho = 0,9$	S	-	-	-	-	-	-	-	-
	E	0,90218	0,9138	0,92542	0,70785	0,7398	0,77175	10,20095	0

Tabela 29: Resultado das simulações para $p = 30$, $n = 50$ e $\delta = 0,05$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,95242	0,98	1,00758	0,66057	0,73379	0,80701	6,35908	0
	E	0,90908	0,94848	0,98788	0,74018	0,76946	0,79873	7,15813	0
$\rho = 0,2$	S	0,86126	0,90671	0,95217	0,37541	0,45062	0,52582	10,55709	0
	E	0,90037	0,93145	0,96253	0,74307	0,76471	0,78636	8,45999	0
$\rho = 0,5$	S	0,67102	0,74317	0,81532	0,32952	0,40783	0,48615	7,4027	0
	E	0,90015	0,92047	0,94078	0,73264	0,75777	0,7829	9,62735	0
$\rho = 0,7$	S	0,51463	0,596	0,67737	0,26701	0,33883	0,41065	5,3594	0
	E	0,90175	0,92126	0,94077	0,74386	0,76676	0,78967	10,48502	0
$\rho = 0,9$	S	0,45238	0,53381	0,61524	0,29079	0,36602	0,44126	3,64535	0,00043
	E	0,92615	0,93846	0,95077	0,72329	0,75284	0,7824	11,68057	0

Tabela 30: Resultado das simulações para $p = 30$, $n = 50$ e $\delta = 0,10$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,89701	0,94075	0,98449	0,50429	0,5731	0,6419	8,98009	0
	E	0,97765	0,99245	1,00725	0,70608	0,74222	0,77835	12,38558	0
$\rho = 0,2$	S	0,84537	0,88952	0,93367	0,25993	0,31085	0,36178	15,72841	0
	E	0,9606	0,9776	0,9946	0,71209	0,74431	0,77653	14,79781	0
$\rho = 0,5$	S	0,65101	0,70457	0,75812	0,25265	0,30495	0,35724	11,06345	0
	E	0,9496	0,9631	0,97659	0,71226	0,74367	0,77508	12,11213	0
$\rho = 0,7$	S	0,49648	0,55716	0,61784	0,26498	0,31792	0,37085	6,12805	0
	E	0,94693	0,95815	0,96937	0,71174	0,73766	0,76358	15,08525	0
$\rho = 0,9$	S	0,42022	0,48219	0,54416	0,26031	0,31763	0,37495	3,9447	0,00015
	E	0,93984	0,95202	0,96419	0,69354	0,73174	0,76993	11,04131	0

Tabela 31: Resultado das simulações para $p = 30$, $n = 100$ e $\delta = 0$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	-	-	-	-	-	-	-	-
	E	0,0077	0,0466	0,0855	0,70945	0,7146	0,71975	-33,21598	0
$\rho = 0,2$	S	-	-	-	-	-	-	-	-
	E	0,39811	0,4536	0,50909	0,70148	0,7064	0,71132	-8,82913	0
$\rho = 0,5$	S	-	-	-	-	-	-	-	-
	E	0,83104	0,8489	0,86676	0,70841	0,7134	0,71839	14,61413	0
$\rho = 0,7$	S	-	-	-	-	-	-	-	-
	E	0,8816	0,8939	0,9062	0,70628	0,711	0,71572	26,9891	0
$\rho = 0,9$	S	-	-	-	-	-	-	-	-
	E	0,91387	0,9237	0,93353	0,70577	0,7104	0,71503	41,28887	0

Tabela 32: Resultado das simulações para $p = 30$, $n = 100$ e $\delta = 0,05$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	1	1	1	0,97085	0,98786	1,00486	1,39935	0,16483
	E	0,97019	0,98979	1,00939	0,74385	0,75089	0,75792	24,06496	0
$\rho = 0,2$	S	0,8304	0,86529	0,90018	0,50456	0,55683	0,60911	10,40212	0
	E	0,96423	0,97692	0,98962	0,7149	0,72016	0,72542	36,07643	0
$\rho = 0,5$	S	0,62384	0,67463	0,72541	0,39757	0,44826	0,49895	5,98118	0
	E	0,94282	0,95461	0,9664	0,71235	0,7175	0,72266	37,51581	0
$\rho = 0,7$	S	0,50871	0,56041	0,61211	0,36856	0,40837	0,44818	5,43844	0
	E	0,95643	0,96456	0,97268	0,71119	0,7162	0,72121	53,62297	0
$\rho = 0,9$	S	0,43465	0,48871	0,54277	0,31177	0,3614	0,41103	3,70736	0,00035
	E	0,94889	0,95814	0,96739	0,71033	0,71573	0,72114	50,79574	0

Tabela 33: Resultado das simulações para $p = 30$, $n = 100$ e $\delta = 0,10$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	0,95139	0,97612	1,00085	0,87791	0,91621	0,95451	2,80349	0,00608
	E	1	1	1	0,77318	0,78201	0,79085	48,3577	0
$\rho = 0,2$	S	0,78531	0,82208	0,85885	0,4077	0,44335	0,479	14,50476	0
	E	0,98743	0,99162	0,99581	0,72287	0,72904	0,73522	72,67814	0
$\rho = 0,5$	S	0,48827	0,53409	0,57991	0,33163	0,36776	0,4039	5,68832	0
	E	0,97419	0,98084	0,98749	0,71418	0,72014	0,7261	56,40863	0
$\rho = 0,7$	S	0,44272	0,48552	0,52832	0,29898	0,32816	0,35735	6,04964	0
	E	0,95624	0,96421	0,97217	0,71528	0,72139	0,72749	49,87429	0
$\rho = 0,9$	S	0,35344	0,39456	0,43567	0,31309	0,34342	0,37375	2,14466	0,03443
	E	0,96554	0,97279	0,98003	0,70791	0,71337	0,71882	58,68144	0

Tabela 34: Resultado das simulações para $p = 30$, $n = 200$ e $\delta = 0$

		Método Proposto			Método de Mahalanobis				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	-	-	-	-	-	-	-	-
	E	-0,00199	0,00505	0,01209	0,78655	0,79185	0,79715	-167,6594	0
$\rho = 0,2$	S	-	-	-	-	-	-	-	-
	E	0,2525	0,3036	0,3547	0,78247	0,7885	0,79453	-18,4739	0
$\rho = 0,5$	S	-	-	-	-	-	-	-	-
	E	0,81314	0,82865	0,84416	0,77837	0,7848	0,79123	5,07502	0
$\rho = 0,7$	S	-	-	-	-	-	-	-	-
	E	0,87287	0,8848	0,89673	0,7799	0,7863	0,7927	13,71656	0
$\rho = 0,9$	S	-	-	-	-	-	-	-	-
	E	0,9125	0,9213	0,9301	0,77841	0,78345	0,78849	26,58586	0

Tabela 35: Resultado das simulações para $p = 30$, $n = 200$ e $\delta = 0,05$

		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	0,99147	0,99388	0,99629	0,79856	0,80474	0,81091	57,2908	0
$\rho = 0,2$	S	0,92994	0,94854	0,96713	0,60809	0,65009	0,69208	12,18888	0
	E	0,93402	0,94465	0,95527	0,79226	0,79893	0,80559	22,15884	0
$\rho = 0,5$	S	0,69185	0,73096	0,77006	0,39446	0,43128	0,4681	10,66987	0
	E	0,93235	0,94092	0,94948	0,79224	0,79769	0,80313	28,51993	0
$\rho = 0,7$	S	0,56863	0,61623	0,66383	0,34974	0,38566	0,42159	8,09204	0
	E	0,93988	0,94726	0,95463	0,79224	0,79784	0,80344	33,97794	0
$\rho = 0,9$	S	0,38989	0,43324	0,47658	0,27358	0,30815	0,34272	5,53851	0
	E	0,95138	0,95832	0,96526	0,78794	0,79308	0,79823	38,5266	0

Tabela 36: Resultado das simulações para $p = 30$, $n = 200$ e $\delta = 0,10$

		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	0,98706	0,99132	0,99557	0,83087	0,83665	0,84243	37,62889	0
$\rho = 0,2$	S	0,86831	0,89073	0,91316	0,45815	0,49406	0,52997	18,82905	0
	E	0,96047	0,96901	0,97755	0,80332	0,80867	0,81403	32,39461	0
$\rho = 0,5$	S	0,63663	0,66786	0,69909	0,31275	0,33884	0,36493	14,88463	0
	E	0,95313	0,96018	0,96723	0,78891	0,79479	0,80067	35,44143	0
$\rho = 0,7$	S	0,50901	0,54643	0,58385	0,28222	0,30753	0,33285	10,32905	0
	E	0,94605	0,95336	0,96066	0,78684	0,79228	0,79772	35,70224	0
$\rho = 0,9$	S	0,37306	0,40881	0,44456	0,26202	0,28521	0,3084	5,68583	0
	E	0,95653	0,96219	0,96785	0,78768	0,79421	0,80074	39,76028	0

Tabela 37: Resultado das simulações para $p = 30$, $n = 500$ e $\delta = 0$

		Método Proposto			**Método de Mahalanobis**				
		IC inf.	**Média**	**IC sup.**	**IC inf.**	**Média**	**IC sup.**	t	**p-valor**
$\rho = 0$	S	-	-	-	-	-	-	-	-
	E	0	0	0	0,88012	0,8827	0,88528	-669,6203	0
$\rho = 0,2$	S	-	-	-	-	-	-	-	-
	E	0,08962	0,1179	0,14618	0,87879	0,88138	0,88397	-52,56485	0
$\rho = 0,5$	S	-	-	-	-	-	-	-	-
	E	0,77823	0,79014	0,80205	0,87854	0,88092	0,8833	-14,19532	0
$\rho = 0,7$	S	-	-	-	-	-	-	-	-
	E	0,83804	0,84612	0,8542	0,8767	0,87918	0,88166	-7,79031	0
$\rho = 0,9$	S	-	-	-	-	-	-	-	-
	E	0,88443	0,8908	0,89717	0,87603	0,87828	0,88053	3,70546	0,00035

Tabela 38: Resultado das simulações para $p = 30$, $n = 500$ e $\delta = 0,05$

		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
	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
	E	0,91076	0,91845	0,92614	0,88524	0,88777	0,8903	7,19116	0
$\rho = 0,7$	S	0,65727	0,68133	0,70539	0,23784	0,2567	0,27556	27,20583	0
	E	0,91542	0,92293	0,93044	0,88306	0,88584	0,88863	9,51673	0
$\rho = 0,9$	S	0,5197	0,5438	0,5679	0,19591	0,21239	0,22888	22,42424	0
	E	0,93168	0,93721	0,94275	0,88265	0,88516	0,88766	17,62282	0

Tabela 39: Resultado das simulações para $p = 30$, $n = 500$ e $\delta = 0,10$

		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	0,96216	0,97739	0,99262	0,91073	0,91345	0,91617	8,01727	0
$\rho = 0,2$	S	0,92303	0,93367	0,94431	0,36763	0,39291	0,41819	40,63549	0
	E	0,94292	0,95225	0,96158	0,89418	0,89681	0,89944	10,96805	0
$\rho = 0,5$	S	0,71188	0,73541	0,75895	0,221	0,23535	0,24969	36,32808	0
	E	0,92859	0,93578	0,94297	0,88859	0,89104	0,89349	12,08586	0
$\rho = 0,7$	S	0,55926	0,58332	0,60739	0,19814	0,21154	0,22494	25,51769	0
	E	0,94031	0,94654	0,95277	0,88654	0,88924	0,89194	16,36282	0
$\rho = 0,9$	S	0,4425	0,46271	0,48292	0,17124	0,18151	0,19177	24,31953	0
	E	0,94533	0,95027	0,95521	0,88333	0,88576	0,88819	22,71036	0

ANEXO B – Programa utilizado nas simulações

Caso hipotético em que $p = 5$, $n = 100$, $\delta = 0,05$ e $\rho = 0.2$.

```
########## Pacotes necessarios ##########

library("mvtnorm")
library("sn")
library("mvoutlier")
library("rgl")
library("StatMatch")
library(MVA)
library(cluster)
library(mclust)
library(proxy)
library(depth)
library(robustbase)
library(fpc)
library(miscTools)

###### Inicio da geracao das populacoes contaminadas por outliers ######

########## Define matriz de covariancia populacional ##########
defpar=function(n,p,pho) {
sigar1=diag(p)
eco=diag(p)
for (i in 1:p) {
for (j in 1:p) {
```

```
########## Estrutura AR(1) ##########
if (i==j) sigar1[i,j]=1
########## Estrutura Eco ##########
if (i==j) eco[i,j]=1
if (i!=j) eco[i,j]=pho
}
}
return(list(mar1=sigar1, meco=eco))
}


########## Parametros para geracao dos dados ##########

# Parametros principais
p = 5 ; n = 100 ; delta = 0.05 ; pho2 = 0.2
# Parametros complementares
mi1 = rep(0,p) ; mi2 = rep(2,p) ; nsim = 100 ; k = max(2,ceiling(n/10))
dadosc = matrix(0,n,p) ; outver = matrix(0,n,1) ; vv = matrix(0,nsim,k)
sd2 = 1 ; ResultadoPROP = matrix(0,nsim,4) ; ResultadoM = matrix(0,nsim,4)
grupo_out = matrix(n,ncol=(p+2))
SD2 = matrix(rep(sd2,p),ncol=1)
rho2 = defpar(n,p,pho2)$meco
cross2 = SD2%*%t(SD2)
Sigma2 = rho2*cross2


########## Inicio das simulacoes ##########

for (s in 1:nsim)
{
for (r in 1:n)
{
u=runif(1)
if (u>=delta)  {
dadosc[r,] <- rmvnorm(1,mean=mi1,sigma=Sigma2)
```

```
outver[r] = 0
}
if (u<delta) {
dadosc[r,] <- rmvnorm(1,mean=mi2,sigma=Sigma2)
outver[r] = 1
}
}

####### Final da Geracao das populacoes contaminadas por outliers #######


########## Inicio Metodo Proposto ##########

pop = cbind(dadosc,outver)
## Utilizando o metodo K-medias
set.seed(1)
KM = kmeans(dadosc,k)
## Calculando os centroides
dk = KM$centers
## Identifica os individuos por grupo
cl = KM$cluster
dadosA = cbind(pop,cl)
## Mediana dos dados
medianA = colMedians(dadosc)
## Calculando a distancia entre os centroides e a mediana
d = dist(rbind(dk,medianA))
d = as.matrix(d)
di = d[(k+1),-(k+1)]
## Teste
vvv=as.numeric(di > 2.5*sd(di))
vv[s,]=as.numeric(di > 2.5*sd(di))
## Identificacao dos outliers
grupo_out = NULL
for(j in 1:length(vvv)){
if(vvv[j]==1){
```

```
grupo_out = rbind(grupo_out,dadosA[dadosA[,(p+2)]==j,])
}
}
## Calculando o numero de verdadeiros/falsos positivos/negativos
contvp=0
contfp=0
if(sum(vvv)!= 0){
for (i in 1:dim(grupo_out)[1])
{
if (grupo_out[i,(p+1)]==1){
contvp = contvp+1
}
else {
if(grupo_out[i,(p+1)]==0)
contfp=contfp+1
}
}
}else{
contvp = 0
contfp = 0
}
verd_pos=contvp
fals_pos=contfp
fals_neg=sum(outver)-verd_pos
verd_neg=(n-sum(outver)) - fals_pos
if(sum(outver)!= 0){
p1 = verd_pos/sum(outver)
p2 = fals_pos/(n-sum(outver))
p3 = fals_neg/sum(outver)
p4 = verd_neg/(n-sum(outver))
}else{
p1 = 1
p2 = fals_pos/(n-sum(outver))
p3 = 0
p4 = verd_neg/(n-sum(outver))
```

```
}
## Resultados
ResultadoPROP[s,1] = p1 ; ResultadoPROP[s,2] = p2
ResultadoPROP[s,3] = p3 ; ResultadoPROP[s,4] = p4

########## Final Metodo Proposto ##########


########## Inicio Metodo Mahalanobis ##########

medRobusta = covMcd(dadosc)$center
covRobusta = covMcd(dadosc)$cov
Fobs = (mahalanobis(dadosc,medRobusta,covRobusta))
Ftab = qchisq(0.90,p-1)
outlierM = NULL
for (i in 1:length(Fobs)){
if (Fobs[i] >= Ftab){
outlierM[i]=1
}
else (outlierM[i]=0)
}
verd_pos = 0
fals_neg = 0
fals_pos = 0
verd_neg = 0
for(i in 1:length(outver)){
if(outlierM[i]==1 & outver[i]==1){
verd_pos = verd_pos + 1
}
if(outlierM[i]==0 & outver[i]==1){
fals_neg = fals_neg + 1
}
if(outlierM[i]==1 & outver[i]==0){
fals_pos = fals_pos + 1
}
```

```
if(outlierM[i]==0 & outver[i]==0){
verd_neg = verd_neg + 1
}
}
if(sum(outver)!= 0){
p1M = verd_pos/sum(outver)
p2M = fals_pos/(n-sum(outver))
p3M = fals_neg/sum(outver)
p4M = verd_neg/(n-sum(outver))
}else{
p1M = 1
p2M = fals_pos/(n-sum(outver))
p3M = 0
p4M = verd_neg/(n-sum(outver))
}
## Resultados
ResultadoM[s,1] = p1M ; ResultadoM[s,2] = p2M
ResultadoM[s,3] = p3M ; ResultadoM[s,4] = p4M
}

########## Final Metodo Mahalanobis ##########


######## IC para as medias - Proposto ########

mediaPROP = colMeans(ResultadoPROP)
sdPROPp1 = sd(ResultadoPROP[,1])
sdPROPp2 = sd(ResultadoPROP[,2])
sdPROPp3 = sd(ResultadoPROP[,3])
sdPROPp4 = sd(ResultadoPROP[,4])
alpha = 0.025
ztab = qnorm((1 - alpha),lower.tail=T)
ICPROPp1inf = mediaPROP[1] - ztab * solve(sqrt(nsim)) * sdPROPp1
ICPROPp1sup = mediaPROP[1] + ztab * solve(sqrt(nsim)) * sdPROPp1
ICPROPp1 = cbind(ICPROPp1inf, mediaPROP[1], ICPROPp1sup)
```

```
ICPROPp2inf = mediaPROP[2] - ztab * solve(sqrt(nsim)) * sdPROPp2
ICPROPp2sup = mediaPROP[2] + ztab * solve(sqrt(nsim)) * sdPROPp2
ICPROPp2 = cbind(ICPROPp2inf, mediaPROP[2], ICPROPp2sup)
ICPROPp3inf = mediaPROP[3] - ztab * solve(sqrt(nsim)) * sdPROPp3
ICPROPp3sup = mediaPROP[3] + ztab * solve(sqrt(nsim)) * sdPROPp3
ICPROPp3 = cbind(ICPROPp3inf, mediaPROP[3], ICPROPp3sup)
ICPROPp4inf = mediaPROP[4] - ztab * solve(sqrt(nsim)) * sdPROPp4
ICPROPp4sup = mediaPROP[4] + ztab * solve(sqrt(nsim)) * sdPROPp4
ICPROPp4 = cbind(ICPROPp4inf, mediaPROP[4], ICPROPp4sup)
ICPROPp = rbind(ICPROPp1,ICPROPp2,ICPROPp3,ICPROPp4)


####### IC para as medias - Mahalanobis #######

mediaM = colMeans(ResultadoM)
sdMp1 = sd(ResultadoM[,1])
sdMp2 = sd(ResultadoM[,2])
sdMp3 = sd(ResultadoM[,3])
sdMp4 = sd(ResultadoM[,4])
alpha = 0.025
ztab = qnorm((1 - alpha),lower.tail=T)
ICMp1inf = mediaM[1] - ztab * solve(sqrt(nsim)) * sdMp1
ICMp1sup = mediaM[1] + ztab * solve(sqrt(nsim)) * sdMp1
ICMp1 = cbind(ICMp1inf, mediaM[1], ICMp1sup)
ICMp2inf = mediaM[2] - ztab * solve(sqrt(nsim)) * sdMp2
ICMp2sup = mediaM[2] + ztab * solve(sqrt(nsim)) * sdMp2
ICMp2 = cbind(ICMp2inf, mediaM[2], ICMp2sup)
ICMp3inf = mediaM[3] - ztab * solve(sqrt(nsim)) * sdMp3
ICMp3sup = mediaM[3] + ztab * solve(sqrt(nsim)) * sdMp3
ICMp3 = cbind(ICMp3inf, mediaM[3], ICMp3sup)
ICMp4inf = mediaM[4] - ztab * solve(sqrt(nsim)) * sdMp4
ICMp4sup = mediaM[4] + ztab * solve(sqrt(nsim)) * sdMp4
ICMp4 = cbind(ICMp4inf, mediaM[4], ICMp4sup)
ICMp = rbind(ICMp1,ICMp2,ICMp3,ICMp4)
```

```
########## IC para as medias e teste t ##########

rotulo_colunas=c("IC_Inf_P","Media_P","IC_Sup_P","IC_Inf_M",
"Media_M","IC_Sup_M","T","p_valor")
rotulo_linhas=c("p1","p2","p3","p4")
ST=rep(0,4) ; PV=rep(0,4)
for(i in 1:4){
ST[i]=t.test(x=ResultadoPROP[,i], y=ResultadoM[,i], paired = T)$statistic
PV[i]=t.test(x=ResultadoPROP[,i], y=ResultadoM[,i], paired = T)$p.value
}
ICp = round(cbind(ICPROPp,ICMp,ST,PV),5)
colnames(ICp)=rotulo_colunas
rownames(ICp)=rotulo_linhas
ICp
```