MARIA VITÓRIA NEVES

PROPOSIÇÃO DE TESTES ROBUSTOS COMEDIAN PARA DETECÇÃO DE OUTLIERS MULTIVARIADOS BASEADOS EM RESÍDUOS DA ANÁLISE DE COMPONENTES PRINCIPAIS

LAVRAS

2024

MARIA VITÓRIA NEVES

PROPOSIÇÃO DE TESTES ROBUSTOS COMEDIAN PARA DETECÇÃO DE OUTLIERS MULTIVARIADOS BASEADOS EM RESÍDUOS DA ANÁLISE DE COMPONENTES PRINCIPAIS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Prof. Dr. Daniel Furtado Ferreira
Orientador

LAVRAS
2024

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Neves, Maria Vitória

Proposição de testes robustos Comedian para detecção de outliers multivariados baseados em resíduos da análise de componentes principais / Maria Vitória Neves. – 2024.

91 p. : il.

Orientador: Daniel Furtado Ferreira.

Dissertação (mestrado acadêmico) – Universidade Federal de Lavras, 2024.

Bibliografia.

1.Simulação Monte Carlo. 2.robustez. 3.verdadeiros positivos. I. Ferreira, Daniel Furtado. II. Título

MARIA VITÓRIA NEVES

PROPOSIÇÃO DE TESTES ROBUSTOS COMEDIAN PARA DETECÇÃO DE OUTLIERS MULTIVARIADOS BASEADOS EM RESÍDUOS DA ANÁLISE DE COMPONENTES PRINCIPAIS

PROPOSAL OF ROBUST COMEDIAN TESTS FOR DETECTING MULTIVARIATE OUTLIERS BASED ON PRINCIPAL COMPONENT ANALYSIS RESIDUALS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 16 de outubro de 2024.

Prof. Dr. Paulo Henrique Sales Guimarães	UFLA
Prof. Dr. Ben Dêivide de Oliveira Batista	UFSJ
Prof. Dr. José Airton Rodrigues Nunes	UFLA

Prof. Dr. Daniel Furtado Ferreira
Orientador

LAVRAS
2024

AGRADECIMENTOS

À Deus, pela força, por me abençoar e iluminar meu caminho em todos os momentos.

Aos meus pais, Israel e Francisca, e aos meus irmãos, Ladson e Jonas, agradeço por todo amor, carinho, apoio e incentivo que sempre me deram para realizar meus estudos.

Ao meu noivo, Iago, por todo amor, ajuda, compreensão e por sempre estar ao meu lado durante toda esta trajetória.

Agradeço as minhas amigas que fiz no mestrado, Louziane e Victória, pela amizade, ajuda e apoio que me deram, vocês foram essenciais nesta jornada.

Ao professor Daniel Ferreira Furtado, por toda ajuda, paciência e ensinamentos compartilhados nas disciplinas ministradas e como orientador, foi muito importante no processo de desenvolvimento deste trabalho.

Aos meus familiares e minhas amigas que, de alguma forma, colaboraram para que se pudesse desenvolver e concluir essa dissertação.

À Universidade Federal de Lavras (UFLA), pela oportunidade. Aos professores do Departamento de Estatística pelas excelentes aulas e pela contribuição para minha formação.

À CAPES, pelas bolsas de estudo. O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

Em diversas pesquisas, encontram-se amostras aleatórias de variáveis univariadas ou multivariadas. Em muitos casos, a existência de observações denominadas de *outliers* são responsáveis por sérios comprometimentos da inferência estatística. A detecção dos *outliers* se torna de extrema importância nestes casos, uma vez que as inferências realizadas podem levar a conclusões equivocadas. Em um conjunto de dados multivariados, a detecção de *outliers* é mais complexa que nos casos univariados em razão da dimensão ser definida além da reta real. Existem testes para a detecção de *outliers* multivariados. Todos são dependentes de distribuições assintóticas e normais e são extremamente influenciados pela presença dos próprios *outliers* que se pretende identificar e excluir da amostra aleatória. Nesse sentido, o objetivo deste trabalho foi propor testes assintóticos robustos baseados nos estimadores *comedian* e em resíduos da análise de componentes principais para detectar *outliers* multivariados. Objetivou-se também comparar o desempenho dos testes propostos e dos testes existentes avaliados no presente trabalho por meio de simulação Monte Carlo, mensurando a capacidade dos testes de identificar os *outliers* e os não *outliers* na amostra aleatória. Para a geração dos dados simulados, utilizou-se uma amostra proveniente de uma população normal multivariada. A partir dos dados amostrais, foram obtidos os componentes principais para a realização dos testes por meio do programa R. Os testes existentes são os de Jackson e Mudholkar e de Rao. Já os testes propostos consistem no teste de Jackson e Mudholkar e no teste de Rao, ambos utilizando o estimador robusto comedian. Conclui-se que os testes propostos, utilizando o estimador robusto *comedian*, obtiveram os melhores resultados na detecção dos *outliers*. Além disso, o teste de Rao, ao utilizar o *comedian*, destacou-se como o melhor, pois também detectou corretamente os não *outliers*.

Palavras-chave: simulação Monte Carlo; robustez; verdadeiros positivos; verdadeiros negativos.

ABSTRACT

In various studies, random samples of univariate or multivariate variables are encountered. In many cases, the presence of observations referred to as outliers is responsible for significant compromises in statistical inference. Detecting outliers becomes extremely important in these cases, as the inferences made can lead to incorrect conclusions. In a multivariate dataset, the detection of outliers is more complex than in univariate cases because the dimension is defined beyond the real line. There are tests for detecting multivariate outliers. All of them depend on asymptotic and normal distributions and are highly influenced by the presence of the very outliers they aim to identify and exclude from the random sample. In this context, the objective of this study was to propose robust asymptotic tests based on the comedian estimator and residuals from principal component analysis to detect multivariate outliers. It also aimed to compare the performance of the proposed tests with existing tests evaluated in this study through Monte Carlo simulation, measuring the tests' ability to identify outliers and non-outliers in the random sample. For the generation of simulated data, a sample from a multivariate normal population was used. From the sample data, principal components were obtained for conducting the tests using the R software. The existing tests are those of Jackson and Mudholkar and Rao. The proposed tests consist of the Jackson and Mudholkar test and the Rao test, both using the robust comedian estimator. It was concluded that the proposed tests, utilizing the robust comedian estimator, achieved the best results in detecting outliers. Furthermore, the Rao test, when using the comedian estimator, stood out as the best, as it also correctly identified non-outliers.

Keywords: Monte Carlo simulation; robustness; true positives; true negatives.

INDICADORES DE IMPACTO

Este trabalho contribui em pesquisas que envolvem dados multivariados, especialmente na identificação e exclusão de observações discrepantes, conhecidas como *outliers*, que afetam a inferência estatística. A introdução de testes assintóticos robustos, baseados no estimador *comedian* e em resíduos da análise de componentes principais, oferece uma alternativa mais precisa e menos sensível aos próprios *outliers*, reduzindo o viés em pesquisas científicas e aplicações práticas. Esse trabalho beneficia diretamente áreas como biologia, economia, medicina e ciências sociais, onde dados multivariados são amplamente utilizados, aprimorando a confiabilidade dos resultados. Em termos sociais, o aprimoramento da detecção de *outliers* torna-se essencial para pesquisas que impactam diretamente a formulação de políticas públicas e decisões de saúde, por exemplo, uma vez que elimina a possibilidade de distorções causadas por dados extremos. Culturalmente, este trabalho estimula o avanço de métodos estatísticos na sociedade acadêmica, incentivando a adoção de técnicas mais robustas. A extensão deste trabalho permite, também, um impacto econômico, pois a aplicação de métodos mais eficientes reduz custos e tempo na análise de grandes volumes de dados, promovendo eficiência em setores de pesquisa e desenvolvimento.

IMPACT INDICATORS

This work contributes to research involving multivariate data, particularly in the identification and exclusion of outlying observations, known as outliers, which affect statistical inference. The introduction of robust asymptotic tests, based on the comedian estimator and residuals from principal component analysis, offers a more accurate and less sensitive alternative to the outliers themselves, reducing bias in scientific research and practical applications. This work directly benefits fields such as biology, economics, medicine, and social sciences, where multivariate data is widely used, enhancing the reliability of results. Socially, improving outlier detection becomes essential for research that directly impacts public policy formulation and health decisions, as it eliminates the possibility of distortions caused by extreme data points. Culturally, this work fosters the advancement of statistical methods within the academic community, encouraging the adoption of more robust techniques. Furthermore, the extension of this work enables an economic impact, as the application of more efficient methods reduces costs and time in analyzing large datasets, promoting efficiency in research and development sectors.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Em uma análise estatística, o objetivo é realizar inferências sobre parâmetros, buscando estimar, testar hipóteses ou prever características de uma população. Estas inferências podem ser realizadas tanto no caso univariado, quanto no caso multivariado. Porém, quando existem *outliers*, as inferências realizadas podem levar a conclusões equivocadas.

Outliers, ou observações atípicas, são pontos ou observações que se encontram distante da maioria do agrupamento amostral de dados ou que apresentam alguma inconsistência com esse agrupamento. Essas observações são também designadas por discrepantes, extremas ou aberrantes (BARNETT et al., 1994).

No espaço multivariado, uma observação é considerada *outlier* quando 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 dos dados (JOLLIFE, 2002).

De acordo com Sathe e Aggarwal (2018), os algoritmos para detecção de *outliers* são computacionalmente intensivos. Os métodos clássicos existentes são bons para identificar outliers em dados com apenas uma variável (HADI, 1992). Sathe e Aggarwal (2018) e Barbosa et al. (2018) afirmaram que a detecção dos *outliers* em alta dimensionalidade é complexa por eles estarem em diferentes subespaços, visto que as variáveis são definidas em um espaço p-dimensional. Além disso, a detecção de *outliers* em conjuntos de dados com um grande número de variáveis é mais complexa, devido a questões como o mascaramento de *outliers*, no qual um *outlier* pode ocultar outras observações que também são *outliers*, e à questão da inundação, em que uma observação que não é um *outlier* pode ser equivocadamente identificada como tal.

Na literatura, há vários métodos robustos para a detecção de *outliers* em dados multivariados, dentre os quais podemos citar os métodos Elipsoide de Volume Mínimo (*Minimum Volume Ellipsoide* - MVE), Ortogonalizado de Gnanadesikan e Kettenring (OGK), Covariância de Determinante Mínimo (*Minimum Covariance Determinant* - MCD), componentes principais para detecção de *outliers* (PCOut) e uso do estimador *comedian* (SAJESH; SRINIVASAN, 2012).

O estimador robusto *Comedian* foi escolhido para este trabalho, pois, de acordo com Casella e Berger (2014), para que o modelo estatístico assumido obtenha eficiência ótima ou próxima do ideal na presença de *outliers*, é recomendável o uso de estatísticas robustas por meio

de estimadores robustos. Além disso, a escolha foi fundamentada nas conclusões do estudo de Martins (2022), que constatou que os testes utilizando o estimador *Comedian* obtiveram bons resultados, e nos resultados obtidos por Sajesh e Srinivasan (2012), em que, ao ser comparado com outros métodos robustos, o método *Comedian* demonstrou ser mais eficiente em termos de tempo computacional.

O objetivo deste trabalho é propor testes assintóticos robustos baseados nos estimadores *comedian* e em resíduos da análise de componentes principais para detectar *outliers* multivariados. Também objetivou-se avaliar os desempenhos dos testes propostos no presente trabalho e já existentes, por meio de simulação Monte Carlo, mensurando a capacidade dos testes em identificar os *outliers* e os não *outliers* na amostra aleatória.

Este trabalho está dividido em cinco seções, em que a seção 1 é a Introdução. Na seção 2 estão as principais bases teóricas utilizadas, como o conceito de formas quadráticas, distâncias, distribuição normal e normal contaminada multivariada, *outliers*, análise de componentes principais, estimador robusto *comedian*, o método Monte Carlo, teste de hipótese e os testes de Jackson e Mudholkar e de Rao, além dos testes propostos utilizando o estimador robusto *comedian*. Na seção 3 será abordado a metodologia utilizada para identificar os *outliers* por meio dos testes existentes e propostos, utilizando simulações Monte Carlo. Na seção 4 estão os resultados e discussões da detecção dos *outliers* e não *outliers* por meio das simulações realizadas em cada teste. Por fim, na seção 5 tem-se a conclusão.

2 REFERENCIAL TEÓRICO

Nesta seção, são apresentadas as principais bases teóricas utilizadas para o desenvolvimento deste trabalho.

2.1 Formas Quadráticas

Segundo Ferreira (2018), uma forma quadrática é definida por

$$Q(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n-1}\sum_{k=i+1}^{n} a_{ik}x_i x_k = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_i x_k, \tag{2.1}$$

em que $\boldsymbol{A}$ é uma matriz simétrica $(n \times n)$ e $\boldsymbol{x} \neq \boldsymbol{0}$ é um vetor definido em $\mathbb{R}^n$. Como os elementos de $\boldsymbol{A}$ são conhecidos, $Q(\boldsymbol{x})$ representa uma função do vetor $\boldsymbol{x}$.

A forma quadrática pode ser classificada conforme o valor de Q($\boldsymbol{x}$). Assim, se $Q(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ é uma forma quadrática, então:

a) $Q(\boldsymbol{x}) > 0$ positiva definida;
b) $Q(\boldsymbol{x}) \geq 0$ semipositiva definida;
c) $Q(\boldsymbol{x}) < 0$ negativa definida;
d) $Q(\boldsymbol{x}) \leq 0$ seminegativa definida.

A matriz $\boldsymbol{A}$, correspondente a cada caso acima, tem a seguinte classificação: positiva definida se $Q(\boldsymbol{x}) > 0$, semipositiva definida se $Q(\boldsymbol{x}) \geq 0$, negativa definida se $Q(\boldsymbol{x}) < 0$ e seminegativa definida se $Q(\boldsymbol{x}) \leq 0$. Se a forma quadrática apresentar resultados negativos e positivos, então a matriz é não definida (FERREIRA, 2018).

2.2 Distâncias

Nas técnicas de análise multivariada, muitos métodos de estimação e de inferência tem como base o conceito de distância. A distância entre dois objetos é usada para estimar o quanto são parecidos.

A expressão geral para a distância quadrática é:

$$d^2(\boldsymbol{x},\boldsymbol{y}) = \| \boldsymbol{x}-\boldsymbol{y} \|^2_{\boldsymbol{\psi}} = (\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{\psi}(\boldsymbol{x}-\boldsymbol{y}), \tag{2.2}$$

em que $\boldsymbol{x}$ e $\boldsymbol{y} \in \mathbb{R}^p$, a matriz $\boldsymbol{\psi}$ positiva definida é denominada de métrica (FERREIRA, 2018).

Dado um terceiro vetor $\boldsymbol{z}$, a distância $d^2(\boldsymbol{x},\boldsymbol{y})$ tem as seguintes propriedades:

a) $d^2(\boldsymbol{x},\boldsymbol{y}) > 0, \qquad \forall \boldsymbol{x} \neq \boldsymbol{y}$;

b) $d^2(\boldsymbol{x},\boldsymbol{y}) = 0$, se, e somente se, $\boldsymbol{x} = \boldsymbol{y}$;

c) $d^2(\boldsymbol{x},\boldsymbol{y}) \leq d(\boldsymbol{x},\boldsymbol{z}) + d(\boldsymbol{y},\boldsymbol{z})$, $\forall \boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$.

No caso em que a métrica dada em (2.2) é definida por $\boldsymbol{\psi} = \boldsymbol{S}^{-1}$, em que $\boldsymbol{S}$ é a matriz de variância e covariância amostral, a equação dada por:

$$d^2(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y}) \tag{2.3}$$

é conhecida como a distância generalizada de *Mahalanobis*.

2.3 Distribuição normal multivariada

Considere p variáveis normais independentes, X_1, $X_2, \cdots$, X_p , com média μ_i e variância σ_{ii}, dispostas no vetor aleatório $\boldsymbol{X} = [X_1, X_2, \cdots, X_p]^\top$. A distribuição conjunta desses componentes gera a distribuição normal multivariada. A distribuição da variável aleatória $\boldsymbol{X}$ é denotada por $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ em que $\boldsymbol{\mu}$ é um vetor de médias p-dimensional, sendo $\boldsymbol{\mu} \in \mathbb{R}^p$ e $\boldsymbol{\Sigma}$ é a matriz de covariâncias, positiva definida e simétrica $p \times p$, sendo $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. Segundo Johnson e Wichern (2007), a função densidade de probabilidade normal multivariada é dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\}. \tag{2.4}$$

Dizemos que o vetor aleatório p - dimensional $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

2.4 Distribuição normal multivariada contaminada

De acordo com Ferreira (2018), um vetor aleatório $\boldsymbol{X} = [X_1, X_2, \cdots, X_p]^\top \in \mathbb{R}^p$ com distribuição normal multivariada contaminada possui a função densidade de probabilidade dada por:

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) =& \delta(2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_1|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) \right\} + \\ &+ (1-\delta)(2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_2|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \right\}, \end{aligned}$$

em que $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$ são as matrizes de covariâncias, positivas definidas, $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ são os vetores de médias e δ está contido entre $[0,1]$, representado a parcela de dados não contaminantes e, de forma complementar, $(1-\delta)$ os dados contaminantes.

2.5 *Outliers*

Outliers podem ser definidos como valores atípicos de um conjunto de dados, ou seja, são valores que se distanciam muito das demais observações de um conjunto de dados ao ponto de parecerem inconsistentes (HAWKINS, 1980). Para Palma e Gallo (2016), *outliers* são observações que desviam rigorosamente da maior parte dos dados assumidos em um modelo. Um único *outlier* pode causar mudanças nas estimativas dos parâmetros e interferir também nos testes de normalidade, de homocedasticidade e de correlação entre as variáveis, além de alterar os resultados de qualquer outro procedimento de inferência (SAJESH; SRINIVASAN, 2012).

Os *outliers* podem ser gerados ao escolher um modelo paramétrico errado ou quando parte das observações seguem outro modelo (HAMPEL, 1971). Os *outliers* também podem ser provenientes de uma distribuição com caudas pesadas como a *t* de *Student*, dado que algumas famílias de distribuições podem produzir *outliers* mais comumente (HAWKINS, 1980).

Porém, de acordo com Hampel et al. (2011), a principal causa da geração de *outliers* é pela ocorrência de erros brutos, devido a uma fonte de desvios que agem apenas ocasionalmente. Algumas fontes de erros brutos são: erro de digitação, erro de cópia, falha de equipamentos e outros efeitos transitórios.

Em conjunto de dados multivariados, a detecção de *outliers* pode ser extremamente difícil, necessitando de uma estatística robusta, pois a maioria dos estimadores clássicos falham quando a fração de contaminação é maior que $1/(p+1)$, em que p é a dimensão dos dados, ou seja, isso significa que em alta dimensão, uma pequena fração de *outliers* pode resultar em estimativas muito ruins (ROCKE; WOODRUFF, 1996).

2.6 Componentes Principais

A ideia central da análise de componentes principais (PCA) é reduzir a dimensão de um conjunto de dados, que consiste em um grande número de variáveis inter-relacionadas, mantendo o máximo possível da variação presente no conjunto de dados (JOLLIFE, 2002).

Os componentes principais são definidos como combinações lineares de p variáveis correlacionadas, sendo que cada combinação linear é não correlacionada com a outra. As combinações lineares são em número igual ao número de variáveis originais presentes no estudo (FERREIRA, 2018).

Considere $\boldsymbol{X}_{n\times p}$ uma matriz de dados obtida através de uma amostra aleatória, sendo a j-ésima linha dada pelo vetor aleatório $\boldsymbol{X}_j^\top = [X_{j1}, X_{j2}, \cdots, X_{jp}]^\top$, para $j = 1, 2, \cdots, n$. Devemos

inicialmente obter a matriz de covariâncias por

$$\boldsymbol{S} = \frac{1}{n-1}\left[\sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{X}_j^{\top} - \frac{\left(\sum_{j=1}^{n}\boldsymbol{X}_j\right)\left(\sum_{j=1}^{n}\boldsymbol{X}_j\right)^{\top}}{n}\right] \tag{2.5}$$

ou a matriz de correlação por

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}, \tag{2.6}$$

em que $\boldsymbol{D}^{-1/2} = \text{diag}(1/\sqrt{S_{ii}})$ (FERREIRA, 2018).

O preditor do i-ésimo componente principal $\hat{Y}_i$ é definido por

$$\hat{Y}_i = \hat{\boldsymbol{e}}_i^{\top}\boldsymbol{X}_j = \hat{\boldsymbol{e}}_{i1}X_{j1} + \hat{\boldsymbol{e}}_{i2}X_{j2} + \cdots + \hat{\boldsymbol{e}}_{ip}X_{jp}, \tag{2.7}$$

em que o vetor $\widehat{\boldsymbol{e}}_i$ estabelece a i-ésima combinação linear estimada para $i = 1, 2, \cdots, p$ (FERREIRA, 2018).

O estimador da variância amostral é

$$\begin{aligned}\widehat{\text{Var}}(\hat{Y}_i) =& \widehat{\text{Var}}\left(\hat{\boldsymbol{e}}_i^{\top}\boldsymbol{X}_j\right) = \hat{\boldsymbol{e}}_i^{\top}\widehat{\text{Var}}(\boldsymbol{X}_j)\hat{\boldsymbol{e}}_i \\ =& \hat{\boldsymbol{e}}_i^{\top}\boldsymbol{S}\hat{\boldsymbol{e}}_i.\end{aligned} \tag{2.8}$$

Os componentes principais são combinações lineares não correlacionadas $Y_1, Y_2, \cdots, Y_p$ cujas variâncias estimadas em (2.8) são tão grandes quanto possível (JOHNSON; WICHERN, 2007). A definição de componentes principais é fundamentada na maximização de sua variância, porém não existe um máximo para a variância, pois à medida que $\hat{\boldsymbol{e}}_i$ cresce, a variância tende ao infinito. Uma maneira de contornar este problema é aplicar a restrição $\hat{\boldsymbol{e}}_i^{\top}\hat{\boldsymbol{e}}_i = 1$.

Os sistemas de equações resultantes, após tomar a derivada de primeira ordem e igualar a zero, é dado por:

$$\left(\boldsymbol{S} - \hat{\lambda}_i\boldsymbol{I}\right)\hat{\boldsymbol{e}}_i = \boldsymbol{0}. \tag{2.9}$$

Os componentes principais amostrais são resultantes da determinação dos autovalores e autovetores da matriz $\boldsymbol{S}$ (FERREIRA, 2018). Da equação (2.9), verificamos a seguinte relação

$$\boldsymbol{S}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i \hat{\boldsymbol{e}}_i,$$

em que

$$\widehat{\mathrm{Var}}(Y_i) = \hat{\lambda}_i$$

e

$$\widehat{\mathrm{Cov}}(\hat{Y}_i, \hat{Y}_l) = 0, \qquad i \neq l = 1, 2, \cdots, p.$$

Se tomarmos os autovalores estimados na seguinte ordem $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$, teremos os estimadores dos componentes principais, dados por $\hat{Y}_1 = \hat{\boldsymbol{e}}_i^\top \boldsymbol{X}$, $\hat{Y}_2 = \hat{\boldsymbol{e}}_2^\top X$, $\cdots$, $\hat{Y}_p = \hat{\boldsymbol{e}}_p^\top X$ (FERREIRA, 2018).

A variância amostral total é dada por

$$\begin{aligned} tr(\boldsymbol{S}) =& tr\left(\hat{\boldsymbol{P}}\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{P}}^\top\right) \\ =& tr(\hat{\boldsymbol{\Lambda}}) = \sum_{i=1}^{p} \hat{\lambda}_i. \end{aligned}$$

Pela definição, $tr(\boldsymbol{S})$ é a soma dos elementos da diagonal, ou seja,

$$tr(\boldsymbol{S}) = \sum_{i=1}^{p} S_{ii} = \sum_{i=1}^{p} \hat{\lambda}_i.$$

Portanto, a variabilidade total contida nos componentes principais é igual à variabilidade total contida nas variáveis originais (JOHNSON; WICHERN, 1998).

Os escores dos componentes principais, que são suas realizações para a j-ésima observação amostral, são dados por

$$\hat{\boldsymbol{Y}}_j = \hat{\boldsymbol{P}}^\top \boldsymbol{X}_j.$$

Como a matriz $\hat{\boldsymbol{P}}$ é ortogonal, $\hat{\boldsymbol{P}}^{-1} = \hat{\boldsymbol{P}}^{\top}$ o que permite o vetor $\boldsymbol{X}_j$ ser recuperado pela transformação

$$\hat{\boldsymbol{X}}_j = \hat{\boldsymbol{P}}\hat{\boldsymbol{Y}}_j.$$

De acordo com Ferreira (2018), se reduzirmos o número de variáveis transformadas, componentes principais, para $k < p$, teremos o vetor $\hat{\boldsymbol{Y}} = [\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}_k]^{\top}$ $(k \times 1)$ e assumindo que os k primeiros autovetores da matriz $\hat{\boldsymbol{P}}$ sejam utilizados para compor a matriz $\hat{\boldsymbol{P}}_k$ $(p \times k)$, teremos

$$\hat{\boldsymbol{Y}}_j = \hat{\boldsymbol{P}}_k^{\top}\boldsymbol{X}_j.$$

As observações das variáveis originais são preditas pelo modelo considerando

$$\tilde{\boldsymbol{X}}_j = \hat{\boldsymbol{P}}_k\hat{\boldsymbol{Y}}_j.$$

Para o modelo reduzido com $k < p$, a covariância é dada por

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{Y}}) = \hat{\boldsymbol{\Lambda}}_k,$$

em que

$$\hat{\boldsymbol{\Lambda}}_k = \begin{bmatrix} \hat{\lambda}_1 & 0 & \cdots & 0 \\ 0 & \hat{\lambda}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\lambda}_k \end{bmatrix}.$$

De acordo com Ferreira (2018), a explicação do modelo reduzido em relação ao modelo completo é obtida pela relação das variâncias dos dois modelos, ou seja, quanto da variação total das variáveis originais é explicada pelo modelo de k componentes principais, que é estimada

por

$$R_k^2 = \frac{\sum_{i=1}^{k} \hat{\lambda}_i}{\sum_{i=1}^{p} \hat{\lambda}_i} \times 100. \tag{2.10}$$

A explicação do i-ésimo componente principal é dada por

$$\hat{P}_i^2 = \frac{\hat{\lambda}_i}{\sum_{l=1}^{p} \hat{\lambda}_l} \times 100.$$

Devemos reter um número $k < p$ de componentes principais amostrais que contemple pelo menos 70% ou 80% da variação total amostral (JOHNSON; WICHERN, 1998).

2.7 Testes de Jackson e Mudholkar e de Rao

Considere um vetor aleatório $\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_n^*$ de tamanho n, proveniente supostamente de uma população normal multivariada com média $\boldsymbol{\mu} \in \mathbb{R}^p$ e covariância $\boldsymbol{\Sigma}$ positiva definida $p \times p$. Os componentes principais populacionais são obtidos a partir da decomposição espectral de $\boldsymbol{\Sigma}$, dada por

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top, \tag{2.11}$$

em que $\boldsymbol{U}$ é uma matriz $p \times p$, com os p autovetores de $\boldsymbol{\Sigma}$ formando suas colunas e $\boldsymbol{\Lambda}$ é uma matriz diagonal $p \times p$, com os autovalores λ_i, $i = 1, 2, \cdots, p$ de $\boldsymbol{\Sigma}$. Deve-se, inicialmente, centrar as observações por $\boldsymbol{X}_j = \boldsymbol{X}_j^* - \boldsymbol{\mu}$, para $j = 1, 2, \cdots, n$, sendo, portanto $\boldsymbol{X}_j \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Assim, os componentes principais, escalados para terem variâncias unitárias e médias 0, são os vetores aleatórios dados por

$$\boldsymbol{Y}_j = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top \boldsymbol{X}_j, \qquad j = 1, 2, \cdots, n. \tag{2.12}$$

Observe que $E(\boldsymbol{Y}_j)$ e $\text{Cov}(\boldsymbol{Y}_j)$ são, respectivamente, dados por

$$\begin{aligned} E(\boldsymbol{Y}_j) &= E\left(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top \boldsymbol{X}_j\right) = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top E(\boldsymbol{X}_j) \\ &= \boldsymbol{0} \end{aligned}$$

e

$$\begin{aligned}\text{Cov}(\boldsymbol{Y}_j) =& \text{Cov}\left(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top \boldsymbol{X}_j\right) = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top \text{Cov}(\boldsymbol{X}_j)\boldsymbol{U}\boldsymbol{\Lambda}^{-1/2} \\ =& \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2} \\ =& \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} \\ =& \boldsymbol{I},\end{aligned}$$

pois $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$.

A solução inversa do problema existe e é dada pelo vetor

$$\boldsymbol{X}_j = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{Y}_j, \qquad j = 1, 2, \cdots, n. \tag{2.13}$$

Se forem fixados apenas os k primeiros componentes principais, ou seja, se for utilizado apenas uma parte do conjunto de autovalores de $\boldsymbol{\Sigma}$, diga-se $k < p$ e denotar esta matriz $p \times k$ por $\boldsymbol{U}_k$, tem-se

$$\boldsymbol{U}_k^\top \boldsymbol{\Sigma}\boldsymbol{U}_k = \boldsymbol{\Lambda}_k, \tag{2.14}$$

em que $\boldsymbol{\Lambda}_k$ é uma matriz diagonal $k \times k$, dos k primeiros autovalores de $\boldsymbol{\Sigma}$. Isso decorre como consequência da ortogonalidade dos k primeiros autovetores em relação aos $p-k$ últimos, sendo que $\boldsymbol{U} = [\boldsymbol{U}_k | \boldsymbol{U}_{p-k}]$. Assim,

$$\begin{aligned}\boldsymbol{U}_k^\top \boldsymbol{\Sigma}\boldsymbol{U}_k =& \boldsymbol{U}_k^\top \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top \boldsymbol{U}_k \\ =& [\boldsymbol{I}_k | \boldsymbol{0}] \left[\begin{array}{c|c} \boldsymbol{\Lambda}_k & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{\Lambda}_{p-k} \end{array}\right] \left[\begin{array}{c} \boldsymbol{I}_k \\ \hline \boldsymbol{0} \end{array}\right] \\ =& \boldsymbol{\Lambda}_k.\end{aligned}$$

Dessa forma, os vetores dos k componentes principais correspondentes a redução do conjunto completo (2.12), são dados por

$$\boldsymbol{Y}_{kj} = \boldsymbol{\Lambda}_k^{-1/2}\boldsymbol{U}_k^\top \boldsymbol{X}_j, \qquad j = 1, 2, \cdots, n, \tag{2.15}$$

e o vetor da predição do modelo com k componentes principais para as variáveis originais é

$$\tilde{\boldsymbol{X}}_j = \boldsymbol{U}_k \boldsymbol{\Lambda}_k^{1/2} \boldsymbol{Y}_{kj}, \qquad j = 1, 2, \cdots, n. \tag{2.16}$$

A matriz dos resíduos do modelo em que foram especificados exatos k componentes principais é dada por

$$\begin{aligned}
\boldsymbol{\Sigma}_R =& \boldsymbol{\Sigma} - \text{Var}(\tilde{\boldsymbol{X}}_j) = \boldsymbol{\Sigma} - \boldsymbol{U}_k \boldsymbol{\Lambda}_k^{1/2} \text{Var}(\boldsymbol{Y}_{kj}) \boldsymbol{\Lambda}_k^{1/2} \boldsymbol{U}_k^\top \\
=& \boldsymbol{\Sigma} - \boldsymbol{U}_k \boldsymbol{\Lambda}_k^{1/2} \boldsymbol{I}_k \boldsymbol{\Lambda}_k^{1/2} \boldsymbol{U}_k^\top \\
=& \boldsymbol{\Sigma} - \boldsymbol{U}_k \boldsymbol{\Lambda}_k \boldsymbol{U}_k^\top,
\end{aligned}$$

uma vez que

$$\begin{aligned}
\text{Var}(\boldsymbol{Y}_{kj}) =& \boldsymbol{\Lambda}_k^{-1/2} \boldsymbol{U}_k^\top \boldsymbol{\Sigma} \boldsymbol{U}_k \boldsymbol{\Lambda}_k^{-1/2} \quad \text{(usando (2.14))} \\
=& \boldsymbol{\Lambda}_k^{-1/2} \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^{-1/2} \\
=& \boldsymbol{I}_k.
\end{aligned}$$

Uma medida do ajuste do modelo é dada pela soma de quadrados, que é a distância euclidiana quadrática entre a variável aleatória original $\boldsymbol{X}_j$ e o seu valor predito $\tilde{\boldsymbol{X}}_j$ pelo modelo de k componentes principais. Essa medida é

$$D_j^2 = (\boldsymbol{X}_j - \tilde{\boldsymbol{X}}_j)^\top (\boldsymbol{X}_j - \tilde{\boldsymbol{X}}_j). \tag{2.17}$$

Jackson e Mudholkar (1979) mostraram que

$$(D_j^2/\theta_1)^{h_0} \dot{\sim} N\left(1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2}; \frac{2\theta_2 h_0^2}{\theta_1^2}\right), \tag{2.18}$$

possui aproximadamente distribuição normal univariada, em que

$$\theta_i = \sum_{j=q+1}^{p} \hat{\lambda}_j^i, \quad i = 1, 2, 3$$

e

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}.$$

Um valor observado de D_j^2/θ_1 nos extremos, definidos pelos quantis inferior e superior $100\alpha/2\%$ de uma normal com média $1+(\theta_2 h_0(h_0-1))/\theta_1^2$ e variância $2\theta_2 h_0^2/\theta_1^2$, indica que a j-ésima observação observada deve ser considerada um *outlier* para um dado valor de α escolhido *a priori*.

Por outro lado, Rao (1964) sugere que seja utilizado um procedimento mais simples. Para uma melhor convergência à distribuição qui-quadrado, recomenda-se obter D_j^2 por

$$D_j^2 = \|\boldsymbol{Y}_{(p-k)j}\|^2,$$

em que $\boldsymbol{Y}_{(p-k)j}$ é o vetor aleatório $(p-k)$-dimensional dos escores dos últimos $p-k$ componentes principais para a j-ésima unidade amostral, considerando que $D_j^2 \dot{\sim} \chi_{p-k}^2$. Espera-se que estes procedimentos tenham baixa eficiência se forem aplicados a uma amostra aleatória, uma vez que os estimadores da média e da covariância são muito influenciados pelos *outliers*. No entanto, pode-se obter grande sucesso na identificação de *outliers* multivariados se estes estimadores clássicos forem substituídos pelos respectivos estimadores robustos *comedian* (FALK, 1997).

2.8 Estatística robusta

De acordo com Hampel et al. (2011), a estatística robusta busca melhorar os modelos paramétricos já existentes, deixando-os robustos quando os dados apresentam algum tipo de desvio que poderia comprometer as pressuposições do modelo. A estatística robusta é uma extensão da estatística paramétrica clássica, reconhecendo que os modelos paramétricos são apenas aproximações da realidade, já que suposições como normalidade multivariada e independência dos dados amostrados nem sempre se verificam.

2.8.1 Estimador robusto *comedian*

O método *Comedian* (COM) foi criado por Falk (1997) ao propor uma medida de dependência robusta entre duas variáveis aleatórias. Sejam X e Y duas variáveis aleatórias, então

o *comedian* entre X e Y é definido como

$$COM = med\left((X - med(X))(Y - med(Y))\right), \tag{2.19}$$

em que *med* representa a mediana. Sendo assim, o *Comedian* é na realidade uma mediana, e como a mediana, ele também possui o maior ponto de ruptura possível (FALK, 1997).

A expressão (2.19) é a generalização do Desvio Absoluto Mediano (MAD), pois é igual ao σ^2 =MAD^2 quando X=Y, sendo que MAD é uma alternativa robusta do estimador clássico do desvio padrão (σ), sendo descrita da seguinte forma:

$$MAD(X) = med\ (|X - med(X)|). \tag{2.20}$$

Segundo Falk (1997), as propriedades e características do *Comedian* são:

a) *comedian* pode ser utilizado como uma medida alternativa de covariância robusta, portanto, COM(X,Y) é correspondente a COV(X,Y). Entretanto, COV(X,Y) requer a existência dos dois primeiros momentos das variáveis X e Y, enquanto COM(X,Y) sempre existirá, dado que seu cálculo é realizado por meio da mediana, que é uma medida de posição;
b) se X e Y são independentes, COM(X,Y) = 0;
c) *comediam* é simétrico, invariante em relação a translações e possui escala equivariante, ou seja, COM(X, aY +b) = a COM(X,Y) = aCOM(Y,X);
d) se Y = aX +b, COM(X,aX +b) = aMAD^2 (X).

De acordo com Sajesh e Srinivasan (2012), na estatística multivariada utiliza-se o *Comedian* para obter a matriz de covariâncias robusta COM(X). Seja $\boldsymbol{X}_{nxp}$, uma matriz com linhas, $\boldsymbol{X}_i^\top$, $i = 1,2,\cdots,n$ e colunas $\boldsymbol{X}_j^\top$, $j = 1,2,\cdots,p$. A matriz COM(X) será calculada por:

$$COM(X) = \begin{bmatrix} MAD^2(\boldsymbol{X}_1) & COM(\boldsymbol{X}_1,\boldsymbol{X}_2) & \cdots & COM(\boldsymbol{X}_1,\boldsymbol{X}_p) \\ COM(\boldsymbol{X}_1,\boldsymbol{X}_2) & MAD^2(\boldsymbol{X}_2) & \cdots & COM(\boldsymbol{X}_2,\boldsymbol{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ COM(\boldsymbol{X}_p,\boldsymbol{X}_1) & \cdots & \cdots & MAD^2(\boldsymbol{X}_p) \end{bmatrix}. \tag{2.21}$$

Porém, a matriz COM (X), como alternativa robusta para a matriz de covariâncias, é em geral, não positiva (semi-definida) (FALK, 1997). Com o objetivo de solucionar este problema,

Maronna e Zamar (2002) propuseram os seguintes passos para obter estimativas robustas para o vetor de médias e matriz de covariâncias:

a) estabeleça $\boldsymbol{\delta}(\boldsymbol{X}) = \boldsymbol{D}COM(\boldsymbol{X})\boldsymbol{D}^{\top}$ como a matriz de correlação mediana multivariada em que $\boldsymbol{D}$ é uma matriz diagonal com elementos 1/MAD$(\boldsymbol{X}_i)$, $i = 1, 2, \cdots, p$;
b) realiza-se a decomposição espectral $\boldsymbol{\delta}(\boldsymbol{X}) = \boldsymbol{E}\Lambda\boldsymbol{E}^{\top}$, em que Λ_{pxp} é uma matriz diagonal, cujos elementos são os autovalores λ_j e $\boldsymbol{E}_{pxp}$é uma matriz, cujas colunas são os autovetores e_j, para $j = 1, 2, \cdots, p$;
c) seja $\boldsymbol{Q} = \boldsymbol{DE}$, em que $\boldsymbol{D}$ é definida no passo 1 e $\boldsymbol{Z}_i = \boldsymbol{Q}^{-1}\boldsymbol{x}_i, i = 1, 2, \cdots, n$;
d) as estimativas robustas para o vetor de médias $\boldsymbol{m}(\boldsymbol{X})$ e matriz de covariâncias $\boldsymbol{S}(\boldsymbol{X})$ são respectivamente:

$$\boldsymbol{S}(\boldsymbol{X}) = \boldsymbol{Q}\Gamma\boldsymbol{Q}^{\top} \qquad \text{e} \qquad \boldsymbol{m}(\boldsymbol{X}) = \boldsymbol{Ql}, \tag{2.22}$$

em que $\boldsymbol{\Gamma}_{pxp} = diag(MAD(Z_1)^2, \cdots, MAD(Z_p)^2)$ e $\boldsymbol{l}_{px1} = (med(Z_1), \cdots, med(Z_p))^{\top}$.

Um processo iterativo é aplicado para melhorar as estimativas, substituindo $\boldsymbol{\delta}(\boldsymbol{X})$ por $\boldsymbol{S}(\boldsymbol{X})$ e repetindo os passos acima, até que as estimativas se estabilizem em um ponto qualquer do processo iterativo (MARONNA; ZAMAR, 2002).

Estas estimativas são positivas definida e aproximadamente afim-equivariantes. Além disso, as estimativas obtidas pelo método *comedian* têm um alto ponto de ruptura. A eficiência do método aumenta com o aumento da dimensão dos dados (SAJESH; SRINIVASAN, 2012).

2.9 Método Monte Carlo

O método Monte Carlo é um mecanismo computacionalmente intensivo que pode ser utilizado para avaliação das propriedades de um novo teste ou de um novo procedimento de estimação. Sendo útil para avaliar propriedades dos estimadores, determinar tamanhos amostrais, além de obter soluções de outros inúmeros problemas de inferência estatística, nas mais variadas áreas do conhecimento (FERREIRA, 2013).

Segundo Ferreira (2013), o método Monte Carlo realiza simulações de experimentos em que pelo menos um componente aleatório esteja presente. Desta forma, deve-se assumir uma distribuição de probabilidade para o componente aleatório do modelo, para então gerar dados aleatórios, que são na verdade pseudoaleatórios.

De acordo com Landau e Binder (2014), um dos usos mais simples e eficaz para os métodos Monte Carlo é na avaliação de integrais que são intratáveis por técnicas analíticas. É

importante destacar, que o uso das simulações só é justificado se o método for adequado para substituir um sistema real, dado que, segundo Morettin e Bussab (2010), estudos de simulação tentam reproduzir um problema real por via de um ambiente controlado. O método Monte Carlo pode ser determinado como a representação da solução deste problema, sendo utilizada uma sequência de números aleatórios para se fazer esta representação, a partir de uma sequência pseudoaleatória baseada na distribuição uniforme (0,1) (SILVA, 2009).

2.10 Teste de hipóteses

Os testes de hipóteses são procedimentos estatísticos que permitem tomar decisões sobre um parâmetro ou característica de interesse de uma população. Nos testes existem duas hipóteses, a hipótese nula H_0 que será testada e a hipótese alternativa, que será considerada válida caso a hipótese nula seja rejeitada. A hipótese nula H_0, representa o que pretende-se testar, ou seja, a afirmação realizada sobre o vetor de parâmetros $\boldsymbol{\theta}$, correspondendo ao espaço paramétrico restrito $\boldsymbol{\theta} \in \Omega_0$. Já na hipótese alternativa H_1, não é feita nenhuma restrição sobre o vetor de parâmetros $\boldsymbol{\theta}$, o que corresponde ao espaço paramétrico irrestrito $\boldsymbol{\theta} \in \Omega$ (FERREIRA, 2018). Assim, as hipóteses nula e alternativa são:

$$\begin{cases} H_0: & \boldsymbol{\theta} \in \Omega_0 \\ H_1: & \boldsymbol{\theta} \in \Omega. \end{cases}$$

O espaço restrito, Ω_0 corresponde às restrições que são impostas no espaço paramétrico. Já o espaço irrestrito, Ω , representa todo o espaço paramétrico sem restrição sobre θ.

Logo, o objetivo do teste de hipótese é afirmar se $\boldsymbol{\theta} \in \Omega_0$ ou se $\boldsymbol{\theta} \in \Omega$. Rejeita-se H_0 se o resultado da estatística do teste pertencer a região de rejeição R, que é determinada por:

$$\sup_{\boldsymbol{\theta} \in \Omega_0} P(\boldsymbol{Y} \in R; \boldsymbol{\theta}) = \alpha,$$

em que $\boldsymbol{Y}$ é a matriz $n \times p$ de dados e $0 < \alpha < 1$, sendo α conhecido como nível de significância do teste, dado que a região de rejeição é baseada na premissa do erro tipo I é controlado em um nível α, ou seja, a probabilidade de rejeitar a hipótese nula, quando ela for verdadeira, deve ser um valor pré-fixado de $100\alpha\%$, P(Rejeitar $H_0 | H_0$ verdadeira) =α (FERREIRA, 2018).

Para estabelecer a região de Rejeição de H_0, é necessário descobrir a distribuição da estatística do teste Λ, que é dada por

$$\Lambda = \frac{L_{\Omega_0}(\boldsymbol{Y};\hat{\boldsymbol{\theta}})}{L_{\Omega}(\boldsymbol{Y};\hat{\boldsymbol{\theta}})},$$

em que $L_{\Omega_0}(\boldsymbol{Y};\hat{\boldsymbol{\theta}})$ é o máximo da função de verossimilhança para o espaço restrito e $L_{\Omega}(\boldsymbol{Y};\hat{\boldsymbol{\theta}})$ é o máximo da função de verossimilhança para o espaço irrestrito (FERREIRA, 2018).

Porém, devido à dificuldade de estabelecer a distribuição nula de Λ, Ferreira (2018) comenta que se $\Omega_0 \subset \Omega$, com $\Omega_0 \subset \mathbb{R}^s$ e $\Omega \subset \mathbb{R}^r$, sob determinadas condições de regularidade, $-2\ln(\lambda)$ tem distribuição assintoticamente qui-quadrado com $r-s$ graus de liberdade. A região de rejeição é obtida por

$$R = \{\boldsymbol{Y} | \lambda = -2\ln[\Lambda](Y)] > \chi^2_{\alpha,r-s}\},$$

em que $\chi^2_{\alpha,r-s}$ é o quantil superior da distribuição qui-quadrado com $r-s$ graus de liberdade. Como a distribuição é assintoticamente qui-quadrado, a região de rejeição terá tamanho assintoticamente igual a α.

2.11 Avaliação dos testes de hipóteses

Ao realizar um teste de hipóteses, podem ocorrer dois tipos de erros. Tais erros são conhecidos como erros tipo I e tipo II, ou respectivamente, falso positivo e falso negativo (VIEIRA, 2021). O erro tipo I ocorre quando rejeita-se a hipótese nula, dado que a hipótese é verdadeira, e o erro tipo II, quando aceita a hipótese nula, dado que ela é falsa. Geralmente os testes são avaliados e comparados por meio de suas probabilidades de incorrer nestes erros (CASELLA; BERGER, 2014).

O tamanho do erro tipo I, é a probabilidade de rejeitar a hipótese nula, quando ela for verdadeira, ou seja, P(Rejeitar $H_0|H_0$ verdadeira) $= \alpha$. Já o tamanho do erro tipo II é definido pela probabilidade de não rejeitar H_0 dado que H_0 é falsa, ou seja, P(Não rejeitar $H_0|H_0$ é falsa) $= \beta$ (FERREIRA, 2013).

Silva (2009) destaca a impossibilidade de se controlar a probabilidade de ocorrência simultaneamente dos erros, pois a tentativa de diminuir um erro aumenta o outro, para um tamanho de amostra fixo. Isso reforça a importância do controle do erro tipo I e que os erros (tipo I e tipo II) são inversamente proporcionais (BORGES, 2002).

De acordo com Vieira (2021), o poder estatístico é uma probabilidade relacionada com a probabilidade de se cometer o erro tipo II, dada por $(1-\beta)$, ou seja, é a probabilidade de rejeitar a hipótese nula quando ela é falsa, fixados o tamanho do efeito, o nível de significância e o tamanho amostral.

Para avaliar o desempenho dos testes existentes e propostos na detecção dos *outliers* e não *outliers*, os resultados podem ser comparados com a classificação verdadeira em: a) Verdadeiro Positivo (VP); b) Verdadeiro Negativo (VN); c) Falso Positivo (FP); e d) Falso Negativo (FN).

O Quadro 2.1 apresenta o processo das classificações dos testes em relação aos *outliers* e não *outliers*, quando comparados com a classificação verdadeira.

Quadro 2.1 – Quadro de confusão das taxas associadas a detecção de *outliers* e não *outliers* nos testes existentes e propostos.

	Realidade	
Decisão	*Outliers*	Não *outliers*
Outliers	Verdadeiro Positivo - VP	Falso Positivo - FP
Não *outliers*	Falso Negativo - FN	Verdadeiro Negativo - VN

Fonte: Da autora (2024).

3 MÉTODOS

Serão apresentados nesta seção dois dos principais métodos existentes para identificação de *outliers* e os métodos propostos no presente trabalho. Posteriormente, serão apresentados os métodos de validação por simulação Monte Carlo dos métodos de identificação de *outliers* considerados. O desempenho será avaliado por diversas características destes métodos na identificação ou não dos *outliers*, considerando milhares de repetições Monte Carlo de determinados cenários simulados. Os cenários serão determinados de acordo com determinadas características da distribuição escolhida para simular as amostras aleatórias na presença e ausência de *outliers*.

3.1 Testes existentes e propostos

Esta seção apresentará todos os mecanismos estatísticos necessários para realização dos testes existentes e propostos baseados em estimadores clássicos e robustos, sendo todos os procedimentos computacionais realizados no *software* R (R Core Team, 2024).

Considerando uma amostra aleatória $\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_n^*$ de tamanho n, proveniente supostamente de uma população normal multivariada com média $\boldsymbol{\mu} \in \mathbb{R}^p$ e covariância $\boldsymbol{\Sigma}$ positiva definida $p \times p$, serão aplicados os testes definidos a seguir para a identificação de potenciais *outliers*. Em seguida deve-se obter a amostra aleatória $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, cuja média amostral é $\mathbf{0}$, fazendo $\boldsymbol{X}_j = \boldsymbol{X}_j^* - \bar{\boldsymbol{X}}$, para $j = 1, 2, \cdots, n$, em que a matriz de covariâncias e o vetor de médias amostrais devem ser obtidos, respectivamente, por

$$\boldsymbol{S} = \frac{1}{n-1}\left[\sum_{j=1}^{n} \boldsymbol{X}_j^* \boldsymbol{X}_j^{*\top} - \frac{\left(\sum_{j=1}^{n} \boldsymbol{X}_j^*\right)\left(\sum_{j=1}^{n} \boldsymbol{X}_j^*\right)^{\top}}{n}\right]$$

e

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j^*.$$

3.1.1 Testes de Jackson e Mudholkar e de Rao

Os componentes principais amostrais devem ser obtidos partir da decomposição espectral de $\boldsymbol{S}$, dada por

$$\boldsymbol{S} = \boldsymbol{P}\boldsymbol{\Gamma}\boldsymbol{P}^{\top}, \tag{3.1}$$

em que $\boldsymbol{P}$ é uma matriz $p \times p$ com os p autovetores de $\boldsymbol{S}$ formando suas colunas e $\boldsymbol{\Gamma}$ é uma matriz diagonal $p \times p$ com os autovalores amostrais $\hat{\lambda}_i$, $i = 1, 2, \cdots, p$ de $\boldsymbol{S}$.

Considerando os primeiros k componentes principais, os vetores dos escores dos componentes principais serão computados por

$$\boldsymbol{Y}_{kj} = \boldsymbol{\Gamma}_k^{-1/2}\boldsymbol{P}_k^{\top}\boldsymbol{X}_j, \qquad j = 1,\ 2,\ \cdots,\ n \tag{3.2}$$

e os vetores das predições do modelo, com k componentes principais, para as variáveis originais serão computados em seguida por

$$\tilde{\boldsymbol{X}}_j = \boldsymbol{P}_k\boldsymbol{\Gamma}_k^{1/2}\boldsymbol{Y}_{kj}, \qquad j = 1, 2, \cdots, n, \tag{3.3}$$

em que $\boldsymbol{\Gamma}_k$ e $\boldsymbol{P}_k$ são as versões amostrais das matrizes reduzidas correspondentes $\boldsymbol{\Lambda}_k$ e $\boldsymbol{U}_k$, do modelo populacional, descritas anteriormente em (2.7). Se forem fixados apenas os k primeiros componentes principais, ou seja, se for utilizado apenas uma parte do conjunto de autovalores de $\boldsymbol{S}$, diga-se $k < p$ e denotar esta matriz $p \times k$ por $\boldsymbol{P}_k$, tem-se

$$\boldsymbol{P}_k^{\top}\boldsymbol{S}\boldsymbol{P}_k = \boldsymbol{\Gamma}_k, \tag{3.4}$$

em que $\boldsymbol{\Gamma}_k$ é uma matriz diagonal $k \times k$ dos k primeiros autovalores de $\boldsymbol{S}$. Deve-se atentar para o fato de que $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ é resultado da amostra original transladada para a média amostral, ou seja, cuja média amostral é $\boldsymbol{0}$, fazendo $\boldsymbol{X}_j = \boldsymbol{X}_j^* - \bar{\boldsymbol{X}}$, para $j = 1, 2, \cdots, n$.

Em seguida, computa-se as distâncias euclidianas quadráticas

$$D_j^2 = (\boldsymbol{X}_j - \tilde{\boldsymbol{X}}_j)^{\top}(\boldsymbol{X}_j - \tilde{\boldsymbol{X}}_j), \tag{3.5}$$

para $j = 1, 2, \cdots, n$ e a quantidade apresentada em Jackson e Mudholkar (1979), dada por $(D_j^2/\theta_1)^{h_0}$, que possui distribuição assintótica

$$N\left(1+\frac{\theta_2 h_0(h_0-1)}{\theta_1^2};\frac{2\theta_2 h_0^2}{\theta_1^2}\right), \tag{3.6}$$

em que

$$\theta_i = \sum_{j=k+1}^{p} \hat{\lambda}_j^i, \quad i = 1,2,3$$

e

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}.$$

O valor observado de $(D_j^2/\theta_1)^{h_0}$ deve ser confrontado com os quantis inferior e superior $100\alpha/2\%$ de uma normal com média $1+(\theta_2 h_0(h_0-1))/\theta_1^2$ e variância $2\theta_2 h_0^2/\theta_1^2$. Se $(D_j^2/\theta_1)^{h_0}$ estiver nas regiões críticas, deve-se considerar que a j-ésima observação observada é um *outlier* para um dado valor de α escolhido *a priori*.

Para aplicar-se o teste de Rao (1964), recomenda-se obter D_j^2

$$D_j^2 = \|\boldsymbol{Y}_{(p-k)j}\|^2,$$

em que $\boldsymbol{Y}_{(p-k)j}$ é o vetor aleatório padronizado $(p-k)$-dimensional dos escores dos últimos $p-k$ componentes principais para a j-ésima unidade amostral, considerando que $D_j^2 \dot{\sim} \chi_{p-k}^2$.

Como estes procedimentos possuem baixa eficiência se forem aplicados a uma amostra aleatória com *outliers* presentes, uma vez que os estimadores da média e da covariância são influenciados por eles, sugere-se que estes estimadores clássicos sejam substituídos pelos respectivos estimadores robustos *comedian* (FALK, 1997).

3.1.2 Testes robustos propostos baseados nos testes de Jackson e Mudholkar e de Rao

Inicialmente a amostra original é utilizada para se obter os estimadores robustos *comedian* da média e da matriz de covariâncias populacionais de acordo com os procedimentos descritos na seção 2.8.1, propostos por Falk (1997) e Maronna e Zamar (2002). Para isso foi utilizado o pacote do R (R Core Team, 2024) denominado de *robustbase* (MAECHLER et al., 2021). Se estes estimadores forem denotados por $\bar{\boldsymbol{X}}^*$ e $\boldsymbol{S}^*$ para a média e variância, respecti-

vamente, então os componentes principais amostrais devem ser obtidos partir da decomposição espectral de $\boldsymbol{S}^*$, dada por

$$\boldsymbol{S}^* = \boldsymbol{P}^* \boldsymbol{\Gamma}^* \boldsymbol{P}^{*\top}, \tag{3.7}$$

em que $\boldsymbol{P}^*$ é uma matriz $p \times p$ com os p autovetores de $\boldsymbol{S}^*$ formando suas colunas e $\boldsymbol{\Gamma}^*$ é uma matriz diagonal $p \times p$ com os autovalores amostrais $\hat{\lambda}_i^*$, $i = 1, 2, \cdots, p$ de $\boldsymbol{S}^*$ formando sua diagonal.

Considerando os primeiros k componentes principais robustos, os vetores robustos dos escores dos componentes principais serão computados por

$$\boldsymbol{Y}_{kj} = \boldsymbol{\Gamma}_k^{*-1/2} \boldsymbol{P}_k^{*\top} \boldsymbol{X}_j, \qquad j = 1, 2, \cdots, n \tag{3.8}$$

e os vetores das predições do modelo com k componentes principais robustos para as variáveis originais serão computados da seguinte forma

$$\tilde{\boldsymbol{X}}_j = \boldsymbol{P}_k^* \boldsymbol{\Gamma}_k^{*1/2} \boldsymbol{Y}_{kj}, \qquad j = 1, 2, \cdots, n, \tag{3.9}$$

em que $\boldsymbol{\Gamma}_k^*$ e $\boldsymbol{P}_k^*$ são as versões amostrais das matrizes reduzidas correspondentes $\boldsymbol{\Lambda}_k^*$ e $\boldsymbol{U}_k$ do modelo populacional, descritas anteriormente na seção 2.7. Se forem fixados apenas os k primeiros componentes principais, ou seja, se for utilizado apenas uma parte do conjunto de autovalores de $\boldsymbol{S}^*$, diga-se $k < p$ e denotar esta matriz $p \times k$ por $\boldsymbol{P}_k^*$, tem-se

$$\boldsymbol{P}_k^{*\top} \boldsymbol{S}^* \boldsymbol{P}_k^* = \boldsymbol{\Gamma}_k^*, \tag{3.10}$$

em que $\boldsymbol{\Gamma}_k^*$ é uma matriz diagonal $k \times k$, dos k primeiros autovalores de $\boldsymbol{S}^*$. Deve-se atentar para o fato de $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ é resultado da amostra original transladada para a média amostral *comedian* robusta, ou seja, cuja média amostral é $\boldsymbol{0}$, fazendo $\boldsymbol{X}_j = \boldsymbol{X}_j^* - \bar{\boldsymbol{X}}^*$, para $j = 1, 2, \cdots, n$.

Em seguida, as distâncias euclidianas quadráticas são calculadas por

$$D_j^2 = (\boldsymbol{X}_j - \tilde{\boldsymbol{X}}_j)^\top (\boldsymbol{X}_j - \tilde{\boldsymbol{X}}_j), \tag{3.11}$$

para $j = 1,\ 2,\ \cdots,\ n$ e a quantidade apresentada em Jackson e Mudholkar (1979), dada por $(D_j^2/\theta_1)^{h_0}$, que potencialmente possui distribuição assintótica

$$N\left(1+\frac{\theta_2 h_0(h_0-1)}{\theta_1^2};\frac{2\theta_2 h_0^2}{\theta_1^2}\right), \tag{3.12}$$

em que

$$\theta_i = \sum_{j=k+1}^{p} \hat{\lambda}_j^i, \quad i = 1,2,3$$

e

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}.$$

O valor observado de $(D_j^2/\theta_1)^{h_0}$ deve ser confrontado com os quantis inferior e superior $100\alpha/2\%$ de uma normal com média $1+(\theta_2 h_0(h_0-1))/\theta_1^2$ e variância $2\theta_2 h_0^2/\theta_1^2$. Se $(D_j^2/\theta_1)^{h_0}$ estiver nas regiões críticas deve-se considerar que a j-ésima observação observada é um *outlier* para um dado valor de α escolhido *a priori*.

Para aplicar-se o teste de Rao (1964), recomenda-se obter D_j^2

$$D_j^2 = \|\boldsymbol{Y}_{(p-k)j}\|^2,$$

em que $\boldsymbol{Y}_{(p-k)j}$ é o vetor aleatório padronizado $(p-k)$-dimensional dos escores dos últimos $p-k$ componentes principais robustos para a j-ésima unidade amostral, considerando que $D_j^2 \dot{\sim} \chi_{p-k}^2$.

3.2 Simulação Monte Carlo

O desempenho dos 4 métodos apresentados será avaliado por simulação Monte Carlo. Assim, uma amostra aleatória $\boldsymbol{X}_1^*,\ \boldsymbol{X}_2^*,\ \cdots,\ \boldsymbol{X}_n^*$ de tamanho n originada de uma distribuição de probabilidade multivariada será simulada, sendo que entre as n observações haverá uma proporção média de $1-\delta$ observações consideradas *outliers*. Para isso, será considerada a distribuição normal contaminada multivariada no espaço p-dimensional, cuja função densidade

de probabilidade é dada por

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \delta(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\} + \\ + (1-\delta)(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\}, \tag{3.13}$$

em que $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$ são as matrizes de covariâncias positivas definidas, $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ são os vetores de médias e δ está contido entre $[0,1]$, representado a parcela de dados não contaminantes e de forma complementar, $(1-\delta)$ os dados contaminantes.

Para simular uma amostra aleatória, sorteou-se um número uniforme $[0,1]$, u, e comparou-se este valor com δ. Se $u \leq \delta$, então simula-se dados de uma população normal com média $\boldsymbol{\mu}_1$ e covariância positiva definida $\boldsymbol{\Sigma}_1$, que é a população não contaminante. Se por outro lado, $u > \delta$, então simula-se dados de uma população normal com média $\boldsymbol{\mu}_2$ e covariância positiva definida $\boldsymbol{\Sigma}_2$, que é a população contaminante. Para simular vetores aleatórios normais multivariados, com média $\boldsymbol{\mu}_i$ e covariância positiva definida $\boldsymbol{\Sigma}_i$, $i = 1, 2$, primeiramente simulou-se p variáveis normais padrão independentes e as reuniu em um vetor $\boldsymbol{Z} = [Z_1, Z_2, \cdots, Z_p]^\top$. Em seguida, realizou-se a transformação

$$\boldsymbol{X}^* = \boldsymbol{\Sigma}_i^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}_i, \tag{3.14}$$

em que $\boldsymbol{\Sigma}_i^{1/2}$ é a matriz raiz quadrada de $\boldsymbol{\Sigma}_i$, $i = 1, 2$, dada por

$$\boldsymbol{\Sigma}_i^{1/2} = \boldsymbol{P}_i\boldsymbol{\Lambda}_i^{1/2}\boldsymbol{P}_i^\top,$$

sendo ainda que $\boldsymbol{P}_i$ e $\boldsymbol{\Lambda}_i$ são as matrizes de autovetores e autovalores de $\boldsymbol{\Sigma}_i$, respectivamente, e $\boldsymbol{\Lambda}_i^{1/2} = \text{diag}(\lambda_{ik}^{1/2})$, $k = 1, 2, \cdots, p$ e $i = 1, 2$. Este processo foi repetido n vezes até que a amostra de tamanho n foi formada. Para cada repetição do processo, deve-se escolher se a simulação deve gerar um realização da normal não contaminante ou da contaminante, por meio do uso de um número aleatório simulado em cada uma das n repetições, conforme descrição anterior. Os *outliers* foram devidamente identificados em cada amostra simulada.

Os parâmetros e as demais quantidades necessárias para se realizarem as simulações Monte Carlo foram determinados convenientemente para se obter uma ampla gama de situações que correspondem a situações reais. Para isso, considerou-se uma ampla gama de número de variáveis p, diferentes tamanhos amostrais n, diferentes proporções de *outlier* $1-\delta$ e dife-

rentes números de componentes principais retidos k. A contaminação pode se dar por diferenças nas estruturas de covariâncias populacionais e nas médias das duas populações, contaminante e não contaminante. Como as distâncias de Mahalanobis são invariantes à escala escolhida, a estrutura de covariação foi escolhida como esférica para a população não contaminante (covariância diagonal e mesma variância) e de simetria composta, para a população contaminante, com correlação comum ρ_1 e ρ_2.

Vários cenários foram considerados levando em conta os principais fatores determinantes que foram descritos anteriormente, tais como: α, N, n, p, ρ_1, ρ_2, δ e k. Inicialmente, a população não *outlier* será dada pela $N_p(\mathbf{0}, [(1-\rho_1)\boldsymbol{I}_p + \rho_1\boldsymbol{J}_p])$ e a população contaminante *outlier* será a $N_p(\boldsymbol{\mu}, \sigma^2[(1-\rho_2)\boldsymbol{I}_p + \rho_2\boldsymbol{J}_p])$, em que J é uma matriz $p \times p$ de uns e $\boldsymbol{\Sigma} = \sigma^2[(1-\rho_2)\boldsymbol{I}_p + \rho_2\boldsymbol{J}_p]$ é uma matriz de simetria composta. Os valores de σ^2 foram iguais a 1 e 5 e os valores de $\boldsymbol{\mu}$ foram gerados com componentes μ_i que eram números uniformes $U(\mu - 0{,}5,\ \mu + 0{,}5)$, em que $\mu = 1$ e $\mu = 10$. No caso de $\mu = 1$, os *outliers* são de difícil detecção e o caso de $\mu = 10$, a detecção deles é mais fácil. Os valores do nível de significância nominal adotados foram $0{,}10$ e $0{,}05$. O número total de simulações Monte Carlo considerado será de 2000. Os tamanhos amostrais n variaram da seguinte forma: $n = 50(50)300(200)900$ e 1000. As dimensões p foram as seguintes: $p = 2$, 5, 10, 20 e 100. Os valores de ρ_1 e ρ_2 foram 0, $0{,}10$, $0{,}50$ e $0{,}90$, os valores de δ foram $0{,}70$, $0{,}80$ e $0{,}90$ e os valores de k foram 1, 3 e 9. Esses fatores foram combinados fatorialmente para formar todos os cenários simulados.

3.3 Avaliação do desempenho dos testes propostos

Em cada cenário simulado, uma vez obtida uma amostra da normal contaminada com a devida identificação dos *outliers* e dos não *outliers*, foram aplicados os quatro testes descritos anteriormente, quais sejam:

a) jackson e mudholkar (1979) original, denotado por JMO;
b) jackson e mudholkar (1979) utilizando o estimador robusto *comedian*, denotado por JMC;
c) rao (1964) original, denotado por RO;
d) jackson e mudholkar (1979) utilizando o estimador robusto *comedian*, denotado por RC.

Uma vez aplicados os testes, as observações foram classificadas como *outliers* ou não *outliers* individualmente para cada um deles e para cada nível nominal de significância α e o processo foi repetido N vezes. Como no processo de simulação cada observação tem sua procedência corretamente identificada como sendo *outlier* ou não *outlier*, os resultados de cada

teste foram identificados quando comparados com a classificação verdadeira em: a) Verdadeiro Positivo (VP); b) Verdadeiro Negativo (VN); c) Falso Positivo (FP); e d) Falso Negativo (FN). Os VPs são as observações verdadeiras *outliers* simuladas e identificadas corretamente pelos testes como *outliers*. Os VNs são as observações não *outliers* simuladas e identificadas pelos testes como não *outliers*. Os FPs são as observações não *outliers* simuladas e identificadas pelos testes como *outliers*. Os FNs são as observações *outliers* simuladas e identificadas pelos testes como não *outliers*.

Como o processo foi repetido N vezes, para cada nível nominal de significância α, os valores médios em cada uma das classificações descritas anteriormente foram computados, individualmente em todos os testes em cada nível nominal de significância. Finalmente, os resultados para os cenários foram tabulados e os desempenhos dos testes comparados entre si.

4 RESULTADOS E DISCUSSÃO

Nas próximas subseções serão apresentados os resultados que foram obtidos a partir das simulações Monte Carlo previamente previstas neste trabalho por meio do *software* R (R Core Team, 2024) com o objetivo de avaliar o desempenho dos testes propostos e compará-lo com o dos testes pré-existentes. Os resultados obtidos nas simulações para os níveis de significância $\alpha \in \{0,05, 0,10\}$ foram avaliados e foram identificados padrões de comportamento similares ao considerar as mesmas configurações em relação aos diferentes valores de α. Por isso, os resultados obtidos para o nível de significância $\alpha = 0,05$ são apresentados a seguir, enquanto os resultados para $\alpha = 0,10$ podem ser encontrados no Apêndice A, juntamente com alguns resultados similares encontrados.

4.1 Aplicação dos testes em dados não correlacionados e em populações distantes

Na Tabela 4.1 são apresentadas as taxas utilizadas na detecção dos verdadeiros *outliers* e não *outliers* em cada um dos testes, sejam eles os testes pré-existentes ou os testes propostos. Observa-se que os testes utilizando o estimador robusto *comedian* obtiveram um melhor desempenho em relação aos testes originais na detecção dos verdadeiros positivos (VP). Isso pode ser percebido comparando as taxas obtidas via simulação Monte Carlo com o valor da proporção de *outliers* dada por $1 - \delta$. Este fato ocorre porque, segundo Hampel et al. (2011), mesmo que os dados sejam de alta qualidade, ou seja, gerados por meio da distribuição normal multivariada, os métodos robustos podem fornecer uma melhoria notável em relação ao modelo estatístico clássico. A presença de *outliers* na amostra perturba as estimativas dos parâmetros populacionais, como média e matriz de covariância, causando dificuldades dos métodos que utilizam estas estimativas nas suas estatísticas detectarem os causadores destas perturbações.

Os testes JMO e RO apresentam uma melhoria na detecção dos *outliers* à medida que p aumenta, para os valores de $\sigma^2 = 5$. No entanto, ainda para $\sigma^2 = 5$, a detecção dos verdadeiros negativos (VN) diminui para os valores de $\delta = \{0,7,\ 0,8\}$ com o aumento de p. Portanto, quando a verdadeira proporção de *outliers* é baixa, $\delta = 0,90$, as influências nas estatísticas dos testes são menores do que quando essa proporção é alta, o que provavelmente implica na melhor detecção dos VN. Isso é verificado quando se compara o valor da taxa de VN obtida nas simulações com o valor de δ. Por outro lado, quando $\sigma^2 = 1$, este teste possui uma boa detecção dos VN em relação a todos os valores p. Além disso, para os valores de $p \in \{2,5,10,20\}$ e $\sigma^2 = 1$, os testes originais JMO e RO possuem uma baixa taxa de detecção dos *outliers* e uma boa

detecção dos não *outliers*, havendo algumas exceções, como pode ser visto na Tabela 4.1, para $p = 100$.

Quando $p = 100$, configura-se um caso de alta dimensionalidade, pois $p > n$. Nesse cenário, observam-se algumas diferenças em relação aos demais resultados. Nos testes originais, percebe-se que, para $\sigma^2 = 5$, eles conseguem detectar aproximadamente 100% dos *outliers*, $(1 - \delta)$. Não se encontrou explicação plausível da razão de ocorrer este fato, maior detecção de *outliers*, sob uma situação de não correlação e com a população de *outliers* com a matriz de covariância $\sigma^2\boldsymbol{\rho}$ maior, com $\sigma^2 = 5$. O teste original de Rao, ainda com $p = 100$, detectou corretamente os *outliers* para todos os valores de δ, fazendo com que a taxa de detecção dos falsos negativos (FN) chegue a 0. Porém, a detecção dos verdadeiros negativos é a menor registrada, e, com isso, a detecção dos falsos positivos está próxima de δ. Esses resultados em alta dimensionalidade $n = 100$ estão de acordo com o encontrado por Sajesh e Srinivasan (2012). Estes autores compararam alguns métodos de detecção de *outliers* usando outros métodos diferentes dos encontrados no presente trabalho, mas comparando-se os testes pré-existentes com aqueles propostos por eles utilizando os estimadores *Comedian*. Os métodos robustos utilizados pelos autores foram particularmente mais eficazes em detectar *outliers* em conjuntos de dados de alta dimensionalidade, uma área onde outros métodos pré-existentes apresentaram desempenho inferior, como ocorreu no presente trabalho.

Os resultados apresentados na Tabela 4.1 mostram que, conforme p aumenta, os testes utilizando o estimador *comedian* apresentam taxas de detecção dos verdadeiros *outliers* que se aproximam de 100%, o que também pode ser visto na Figura 4.1, embora com tamanho amostral bem maior de $n = 700$. Na detecção dos verdadeiros não *outliers* (VN), observa-se que o teste JMC para ambos os valores de σ^2 e valores de $\delta = \{0{,}7, 0{,}8\}$ apresenta um padrão de resposta contrário, pois, quando p aumenta, a taxa de detecção dos VNs diminui. Já o teste RC apresenta uma boa detecção dos verdadeiros negativos em todas as configurações utilizadas. Em alta dimensionalidade, $p = 100$, o teste JMC teve um desempenho muito ruim em relação a taxa de VN, o que não ocorreu com o teste robusto RC proposto.

Os resultados obtidos anteriormente para $\alpha = 0{,}05$ apresentaram padrões de respostas semelhantes aos apresentados na Tabela 1 para o nível de significância $\alpha = 0{,}10$, considerando $n = 50$. Da mesma forma, os resultados das Tabelas A.2 e A.3 para $n = 200$, considerando os dois níveis de significância, são semelhantes os resultados que já foram descritos anteriormente para $n = 50$.

Tabela 4.1 – Simulação Monte Carlo com tamanho amostral $n = 50$, número de componentes principais retidos $k = 1$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

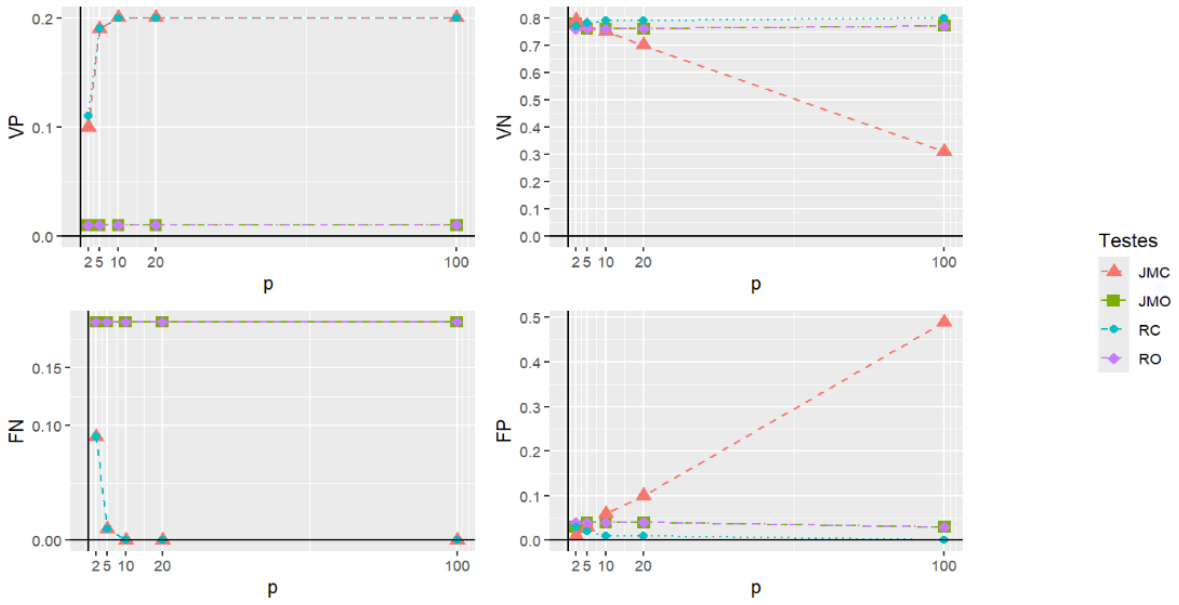
δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		2	0,0063	0,6827	0,1921	0,6902	0,0130	0,6652	0,2017	0,6819
		5	0,0104	0,6776	0,2828	0,6573	0,0101	0,6760	0,2838	0,6920
	1	10	0,0091	0,6740	0,3007	0,5642	0,0063	0,6761	0,3008	0,6887
		20	0,0061	0,6887	0,2956	0,3678	0,0012	0,6953	0,2954	0,6988
		100	0,0002	0,7010	0,2971	0,0229	0,2986	0,0055	0,2974	0,7006
0,7		2	0,0388	0,6988	0,2235	0,6951	0,0554	0,6957	0,2326	0,6896
		5	0,0720	0,6469	0,2907	0,6489	0,1033	0,69786	0,2934	0,6937
	5	10	0,1068	0,5884	0,2959	0,5594	0,1449	0,7024	0,2961	0,7000
		20	0,1580	0,4759	0,2993	0,3378	0,1549	0,6996	0,2998	0,6986
		100	0,2739	0,1871	0,3012	0,0182	0,3012	0,0048	0,3012	0,6986
		2	0,0041	0,7801	0,1687	0,7853	0,0079	0,7620	0,1710	0,7735
		5	0,0066	0,7681	0,2003	0,7547	0,0058	0,7660	0,2004	0,7777
	1	10	0,0063	0,7759	0,1983	0,7172	0,0034	0,7777	0,1983	0,7842
		20	0,0037	0,7819	0,2021	0,6047	0,0005	0,7901	0,2020	0,7861
		100	0,0004	0,7993	0,2003	0,1616	0,2003	0,0050	0,2003	0,7952
0,8		2	0,0327	0,7943	0,1719	0,7855	0,0439	0,7874	0,1749	0,7745
		5	0,0642	0,7540	0,2020	0,7514	0,0865	0,7924	0,2022	0,7783
	5	10	0,1018	0,7230	0,2025	0,7089	0,1199	0,7942	0,2025	0,7833
		20	0,1398	0,6871	0,1965	0,6141	0,1136	0,8021	0,1965	0,7928
		100	0,1951	0,5714	0,1974	0,1841	0,1974	0,0050	0,1974	0,7977
		2	0,0014	0,8782	0,0898	0,8675	0,0032	0,8572	0,0905	0,8462
		5	0,0034	0,8674	0,0986	0,8569	0,0014	0,8639	0,0986	0,8503
	1	10	0,0047	0,8710	0,1002	0,8481	0,0007	0,8715	0,1002	0,8503
		20	0,0052	0,8832	0,0996	0,8222	0,0001	0,8911	0,0996	0,8528
		100	0,0052	0,8977	0,1020	0,5695	0,1020	0,0074	0,1020	0,8577
0,9		2	0,0185	0,8868	0,0926	0,8683	0,0246	0,8731	0,0936	0,8484
		5	0,0409	0,8649	0,1003	0,8544	0,0478	0,8812	0,1003	0,8480
	5	10	0,0594	0,8656	0,0964	0,8534	0,0585	0,8905	0,0964	0,8549
		20	0,0964	0,8549	0,1003	0,8213	0,0400	0,8949	0,1003	0,8548
		100	0,1005	0,8715	0,1006	0,5765	0,1006	0,0070	0,1006	0,8601

Fonte: Da autora (2024).

Na Figura 4.1, considerando $n = 700$, pode-se observar como os testes utilizando o estimador robusto *comedian* são melhores para identificar os *outliers*. Em relação aos verdadeiros negativos, para $p = 2$, o teste JMC é o que mais se aproxima de δ, seguido pelo JMO. Nestes casos, o desempenho de todos os testes na detecção de não *outliers* foram bastante semelhantes. O

teste RC possui o melhor desempenho em detectar os não *outliers* para $p = 5$ e se mantém como o melhor para todos os outros valores de p estudados. O teste JMO se iguala ao RO com $p = 5$, mantendo-se semelhantes para os demais valores de p na detecção dos VN. Nota-se também que o teste JMC diminui acentuadamente a detecção de VN à medida que o p aumenta e com isso a detecção dos falsos positivos aumenta à medida que p aumenta. Além disso, observa-se uma relação complementar entre as taxas de verdadeiros positivos (VP) e falsos negativos (FN), assim como as taxas de verdadeiros negativos (VN) e falsos positivos (FP).

Figura 4.1 – Simulação Monte Carlo com $n = 700$, $k = 1$, $\mu = 10$, $\delta = 0,8$, $\rho_1 = 0$, $\rho_2 = 0$, $\sigma^2 = 1$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP), verdadeiro negativo (VN), falso negativo (FN) e falso positivo (FP) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes número de variáveis p.



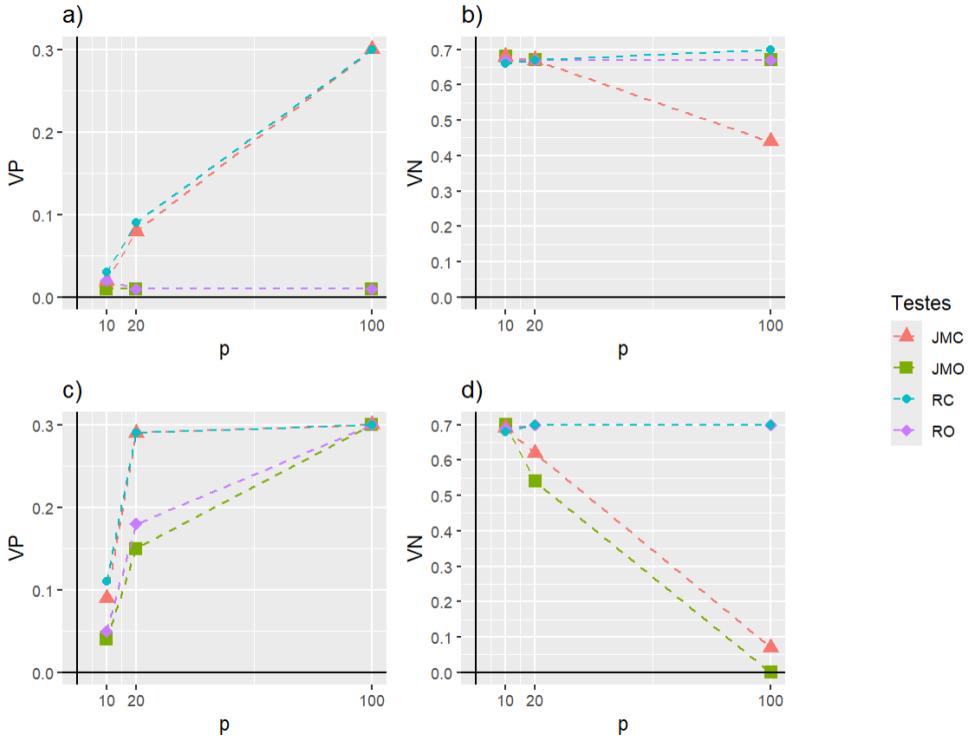
Fonte: Da autora (2024).

Nas Tabelas 4.2 e 4 são apresentados os resultados para $k = 3$, que se mostram semelhantes com os resultados obtidos para $k = 1$, apresentados na Tabela 4.1, em que é importante destacar como o teste de Rao, utilizando o estimador robusto *comedian* (RC), se sobressai na detecção dos *outliers* (VP) e dos não *outliers* (VN).

Nas Figuras 4.2, 4.3 e 4.4 e na Tabela 5, os resultados para $k = 9$ demonstram que permanecem idênticos aos encontrados para $k = 1$. Isso evidencia novamente a superioridade do teste de Rao com o estimador robusto *comedian* (RC), na detecção de *outliers* (VP) e de não *outliers* (VN). Uma possível explicação para o fato de não haver ganhos ou até mesmo

alterações nas taxas de VP e VN quando se aumenta o número de componentes retidos (k) é a ausência de correlações entre as variáveis em todas as simulações apresentadas até o presente momento.

Figura 4.2 – Simulação Monte Carlo com tamanho amostral $n = 700$, número de componentes principais retidos $k = 9$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes números de variáveis p, valores de $\sigma^2 = 1$ (a,b) e $\sigma^2 = 5$ (c,d) e $\delta = 0,7$.

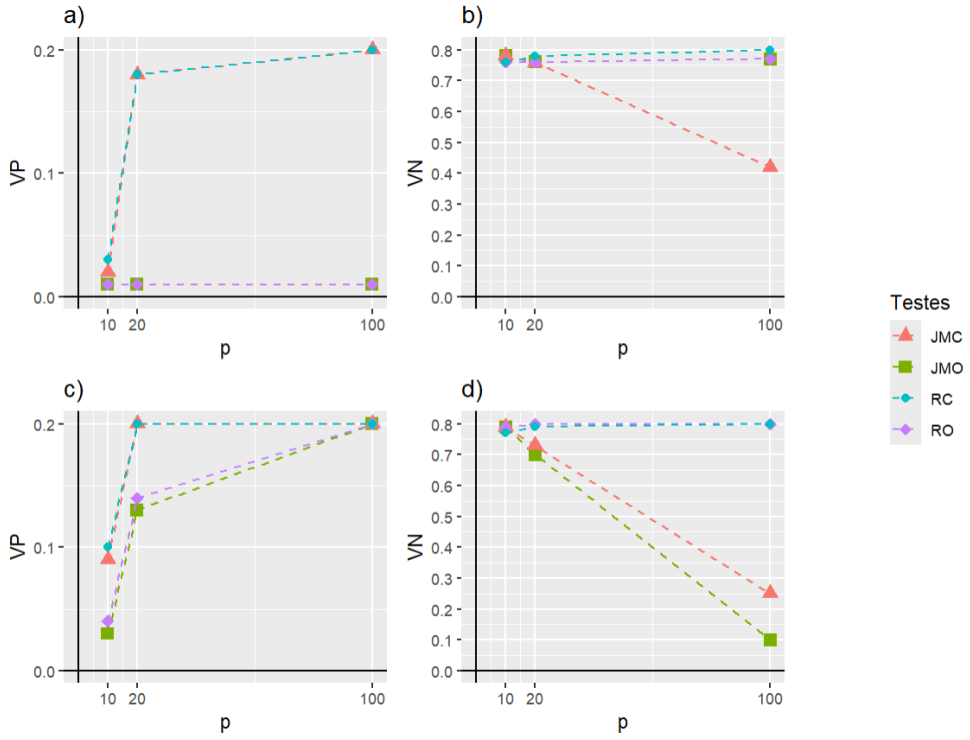


Fonte: Da autora (2024).

4.2 Detecção dos *outliers* em dados correlacionados e populações distantes

Nas Tabelas 4.3 e 10 são apresentados os resultados para $n = 50$, $k = 1$ e correlação $\rho_1 = 0,5$ e $\rho_2 = 0,9$. Estes resultados mostram que os testes utilizando o estimador robusto *comedian* são melhores para detectar as taxas dos verdadeiros positivos, em relação aos testes pré-existentes. Para $p = 100$ e $\sigma^2 = 1$, todos os testes detectam corretamente os VP, porém somente os testes JMO e RC possuem melhor detecção dos verdadeiros negativos, para a configuração apresentada. À medida que p aumenta, a detecção dos verdadeiros negativos decresce

Figura 4.3 – Simulação Monte Carlo com tamanho amostral $n = 700$, número de componentes principais retidos $k = 9$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes números de variáveis p, valores de $\sigma^2 = 1$ (a,b) e $\sigma^2 = 5$ (c,d) e $\delta = 0,8$.



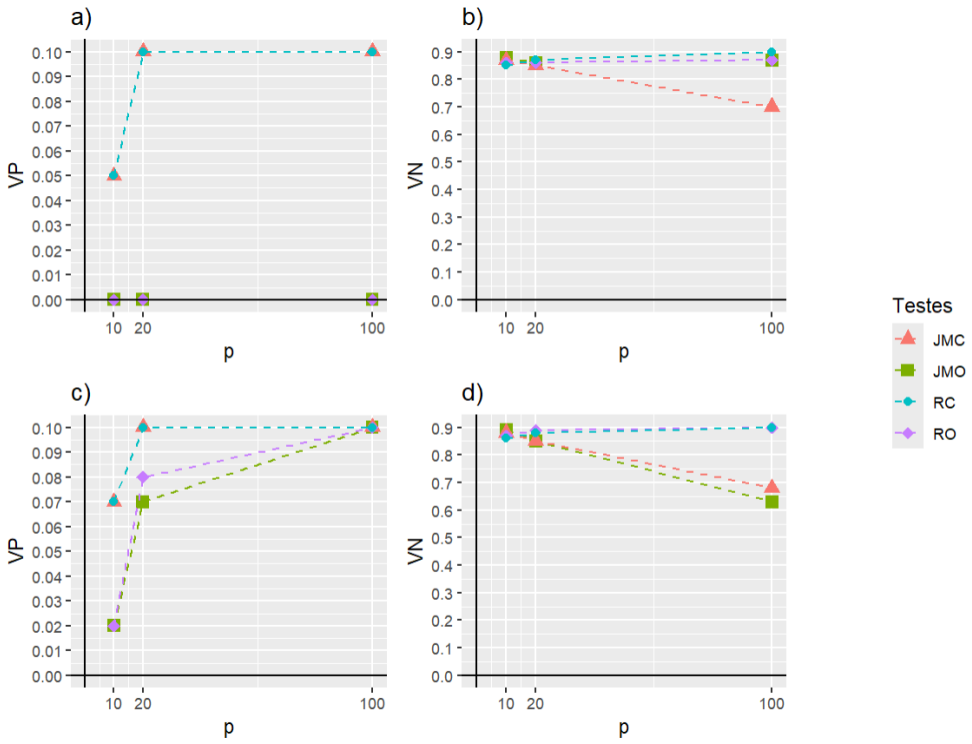
Fonte: Da autora (2024).

no teste JMC. No teste RO a taxa de detecção dos verdadeiros negativos é baixa apenas para $p = 100$.

Nas Tabelas 4.4 e 6 são apresentados os resultados para $n = 1000$, $k = 3$ e correlação da população contaminante $\rho_2 = 0,9$. Observa-se que os testes utilizando o estimador *comedian* obtiveram uma melhor detecção dos *outliers* quando comparados com os testes originais para os valores de $\delta = \{0,7,\ 0,8\}$.

Para $\sigma^2 = 1$ e $p > 5$, o teste JMO apresenta taxas de quase 100 % de detecção dos verdadeiros *outliers*, enquanto para $\sigma^2 = 5$ possui baixa detecção, exceto para $p = 100$. Esse resultado difere do obtido na seção anterior, em que verificou-se o contrário. Novamente não se encontrou explicações plausíveis para as razões da ocorrência deste fato, nas situações onde a população contaminante possui uma estrutura de correlação em que as variáveis são muito cor-

Figura 4.4 – Simulação Monte Carlo com tamanho amostral $n = 700$, número de componentes principais retidos $k = 9$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes números de variáveis p, valores de $\sigma^2 = 1$ (a,b) e $\sigma^2 = 5$ (c,d) e $\delta = 0,9$.



Fonte: Da autora (2024).

relacionadas e a população não contaminante é não correlacionada, e os melhores desempenhos ocorrem onde as variâncias σ^2 são maiores.

À medida que δ aumenta, a detecção dos *outliers* melhora com os testes propostos, uma vez que, ao aumentar δ, a proporção da população de *outliers* diminui, facilitando sua detecção. Porém o teste RC se destaca com $\delta = 0,9$, pois além de apresentar 100% de detecção dos VP para $p \geq 5$ ainda supera o JMC na detecção VN.

Nas tabelas 8 e 9 são apresentados os resultados para $n = 300$, $k = 3$ e correlação $\rho_1 = 0,1$ e $\rho_2 = 0,5$. Estes resultados mostram que os testes propostos utilizando o estimador robusto *comedian* obtiveram as melhores taxas de detecção dos verdadeiros positivos em relação aos testes pré-existentes. Para $p = 100$, todos os testes detectam corretamente os *outliers* e, em relação aos verdadeiros negativos, os testes RO e RC também os detectam corretamente.

Tabela 4.2 – Simulação Monte Carlo com tamanho amostral $n = 150$, número de componentes principais retidos $k = 3$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,0110	0,6732	0,1434	0,6767	0,0140	0,6656	0,1559	0,6760
		10	0,0126	0,6709	0,2788	0,6500	0,0123	0,6703	0,2804	0,6908
	1	20	0,0110	0,6759	0,2967	0,5411	0,0101	0,6756	0,2965	0,6981
		100	0,0034	0,6908	0,3017	0,0146	0,0000	0,6980	0,3017	0,6983
		5	0,0510	0,6803	0,2113	0,6791	0,0696	0,6956	0,2248	0,6886
0,7		10	0,1056	0,6086	0,2957	0,6252	0,1367	0,6981	0,2975	0,6964
	5	20	0,1811	0,4588	0,2988	0,4732	0,2158	0,7011	0,2988	0,7008
		100	0,2956	0,0260	0,2981	0,0081	0,1690	0,7019	0,2981	0,7019
		5	0,0069	0,7720	0,1636	0,7751	0,0087	0,7632	0,1676	0,7741
		10	0,0082	0,7675	0,1970	0,7475	0,0076	0,7659	0,1971	0,7886
	1	20	0,0077	0,7685	0,2019	0,6689	0,0063	0,7675	0,2019	0,7928
		100	0,0025	0,7916	0,1999	0,1527	0,0000	0,7998	0,1999	0,7999
		5	0,0423	0,7823	0,1810	0,7757	0,0552	0,7925	0,1839	0,7796
0,8		10	0,0906	0,7377	0,1996	0,7389	0,1112	0,7972	0,1996	0,7911
	5	20	0,1497	0,6613	0,1997	0,6586	0,1647	0,7988	0,1997	0,7966
		100	0,2009	0,2522	0,2012	0,1376	0,0994	0,7988	0,2012	0,7987
		5	0,0033	0,8675	0,0958	0,8626	0,0042	0,8576	0,0963	0,8542
		10	0,0044	0,8579	0,1027	0,8480	0,0034	0,8569	0,1027	0,8646
	1	20	0,0039	0,8667	0,1002	0,8335	0,0023	0,8653	0,1002	0,8771
		100	0,0020	0,8894	0,1004	0,6245	0,0000	0,8991	0,1004	0,8920
		5	0,0243	0,8728	0,0994	0,8607	0,0316	0,8745	0,0999	0,8537
0,9		10	0,0532	0,8574	0,1009	0,8495	0,0614	0,8846	0,1009	0,8670
	5	20	0,0818	0,8428	0,0999	0,8332	0,0845	0,8896	0,0999	0,8778
		100	0,1001	0,7681	0,1012	0,6207	0,0207	0,8985	0,1012	0,8925

Fonte: Da autora (2024).

Os resultados da Figura 4.5, para $n = 100$ e $k = 3$, mostram que os testes utilizando o estimador robusto *comedian* obtiveram as melhores taxas de detecção dos *outliers* em relação aos testes originais. Com $p = 10$ os resultados obtidos para os testes utilizando o estimador robusto *comedian* foram semelhantes ao obtido por Martins (2022), em que o *comedian* superou os demais métodos, atingindo quase 100% de detecção de *outliers*.

Além disso, percebe-se uma semelhança entre os testes JMC e RC nas configurações apresentadas, pois somente com $p = 5$ o teste RC supera o JMC e ambos apresentam desempe-

Tabela 4.3 – Simulação Monte Carlo com tamanho amostral $n = 50$, número de componentes principais retidos $k = 1$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0,5$ e $\rho_2 = 0,9$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		2	0,0000	0,6668	0,1308	0,6825	0,0000	0,6420	0,1426	0,6714
		5	0,0618	0,6506	0,2261	0,6555	0,0000	0,6223	0,2305	0,6691
	1	10	0,1784	0,6508	0,2659	0,5921	0,0000	0,6138	0,2658	0,6688
		20	0,2796	0,6489	0,2815	0,4360	0,0000	0,6314	0,2770	0,6673
		100	0,3006	0,6567	0,2960	0,0447	0,3006	0,0031	0,2923	0,6693
0,7		2	0,0065	0,6843	0,1450	0,6926	0,0130	0,6682	0,1605	0,6842
		5	0,0111	0,6748	0,2312	0,6560	0,0109	0,6734	0,2429	0,6884
	5	10	0,0086	0,6787	0,2597	0,5814	0,0064	0,6805	0,2681	0,6887
		20	0,0060	0,6881	0,2750	0,4131	0,0013	0,6953	0,2799	0,6935
		100	0,0004	0,6978	0,2911	0,0494	0,3017	0,0033	0,2941	0,6841
		2	0,0000	0,7723	0,1230	0,7820	0,0000	0,7483	0,1292	0,7671
		5	0,0496	0,7612	0,1847	0,7565	0,0000	0,7403	0,1862	0,7702
	1	10	0,1407	0,7621	0,1973	0,7176	0,0000	0,7393	0,1980	0,7673
		20	0,1920	0,7733	0,1968	0,6335	0,0000	0,7645	0,1968	0,7721
		100	0,2003	0,7891	0,2002	0,2154	0,2003	0,0035	0,2001	0,7395
0,8		2	0,0041	0,7814	0,1244	0,7826	0,0084	0,7617	0,1323	0,7690
		5	0,0076	0,7703	0,1846	0,7561	0,0055	0,7677	0,1881	0,7741
	5	10	0,0061	0,7784	0,1923	0,7213	0,0034	0,7790	0,1935	0,7784
		20	0,0046	0,7852	0,1974	0,6235	0,0007	0,7933	0,1983	0,7800
		100	0,0007	0,8024	0,1969	0,2318	0,1972	0,0037	0,1971	0,7392
		2	0,0000	0,8743	0,0668	0,8665	0,0000	0,8505	0,0703	0,8448
		5	0,0321	0,8602	0,0994	0,8513	0,0000	0,8496	0,0998	0,8430
	1	10	0,0814	0,8697	0,0988	0,8438	0,0000	0,8594	0,0988	0,8437
		20	0,0993	0,8798	0,0997	0,8112	0,0000	0,8813	0,0997	0,8419
		100	0,0987	0,8998	0,0987	0,6061	0,0987	0,0051	0,0987	0,7401
0,9		2	0,0014	0,8759	0,0711	0,8662	0,0032	0,8540	0,0746	0,8450
		5	0,0046	0,8657	0,0970	0,8525	0,0015	0,8621	0,0982	0,8433
	5	10	0,0046	0,8723	0,0989	0,8434	0,0009	0,8737	0,0990	0,8444
		20	0,0051	0,8818	0,0995	0,8109	0,0001	0,8910	0,0996	0,8410
		100	0,0062	0,8997	0,0998	0,6181	0,0999	0,0048	0,0999	0,7355

Fonte: Da autora (2024).

nhos semelhantes para valores de $p > 5$. Em relação aos testes originais, o teste RO é melhor do que o JMC na detecção dos VP até o $p = 20$. Para $p = 100$, o teste JMO é melhor.

Dos resultados das Tabelas 4.5 e 7, para $n = 500$ e $k = 9$, depreende-se que os testes propostos utilizando o *comedian* são melhores para detectar os *outliers* em relação aos testes

Tabela 4.4 – Simulação Monte Carlo com tamanho amostral $n = 1000$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0,9$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto comedian (JMC), e Rao com o estimador robusto comedian (RC), considerando diferentes proporções de não outliers δ , variância σ^2 e número de variáveis p.

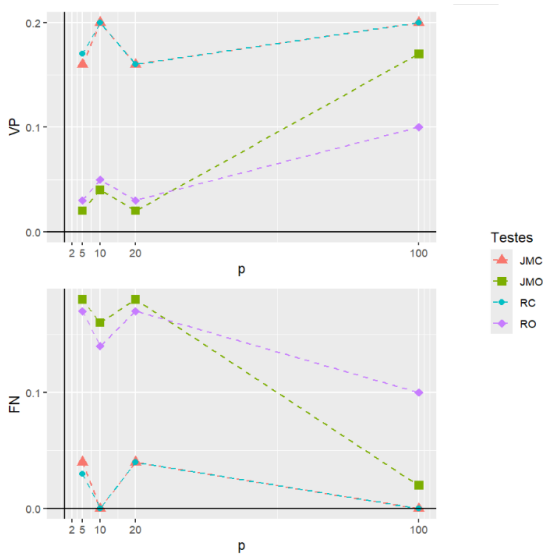
δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,0258	0,6463	0,0119	0,5858	0,0000	0,6222	0,0011	0,5420
		10	0,2646	0,6176	0,0575	0,5514	0.0000	0,5802	0,0101	0,4456
	1	20	0,3001	0,5740	0,1354	0,5724	0,0000	0,5160	0,0747	0,3438
		100	0,3000	0,2992	0,2983	0,6182	0,0000	0,1942	0,2888	0,0525
0,7		5	0,0071	0,6615	0,0133	0,6597	0,0022	0,6456	0,0143	0,6427
		10	0,0281	0,6506	0,0517	0,6609	0,0004	0,6311	0,0467	0,6414
	5	20	0,0700	0,6403	0,1587	0,6685	0,0001	0,6102	0,1396	0,6448
		100	0,2796	0,5674	0,2980	0,6158	0,0000	0,5052	0,2946	0,6593
		5	0,0193	0,7520	0,0328	0,7354	0,0000	0,7312	0,0298	0,7040
	1	10	0,1853	0,7360	0,1409	0,7515	0,0000	0,7080	0,1333	0,6957
		20	0,1998	0,7170	0,1958	0,7641	0,0000	0,6732	0,1941	0,6906
		100	0,1997	0,5856	0,1997	0,7293	0,0000	0,4763	0,1997	0,6437
0,8		5	0,0125	0,7539	0,1734	0,7666	0,0004	0,7413	0,1745	0,7726
	5	10	0,0298	0,7501	0,1970	0,7577	0,0001	0,7331	0,1969	0,7781
		20	0,0669	0,7460	0,1998	0,7361	0,0000	0,7218	0,1998	0,7842
		100	0,1942	0,7189	0,1995	0,5358	0,0000	0,6705	0,1995	0,7964
		5	0,0106	0,8578	0,0768	0,8658	0,0000	0,8416	0,0774	0,8557
	1	10	0,0957	0,8470	0,0999	0,8571	0,0000	0,8315	0,0999	0,8663
		20	0,1002	0,8431	0,1002	0,8472	0,0000	0,8194	0,1002	0,8758
		100	0,1001	0,8179	0,1001	0,7365	0,0000	0,7599	0,1001	0,8923
0,9		5	0,0070	0,8541	0,0995	0,8626	0,0001	0,8461	0,0995	0,8749
	5	10	0,0173	0,8516	0,0999	0,8535	0,0000	0,8421	0,0999	0,8808
		20	0,0400	0,8509	0,1006	0,8374	0,0000	0,8372	0,1006	0,8863
		100	0,0997	0,8498	0,1006	0,7038	0,0000	0,8220	0,1006	0,8964

Fonte: Da autora (2024).

originais para todos os valores de p, porém quando $\sigma^2 = 5$ eles melhoram sua detecção, embora para $p = 10$ eles tenham desempenho inferior aos dos demais valores de p. Para $p = 100$ todos os testes apresentam 100% de detecção dos *outliers*. À medida que o valor de p aumenta, os desempenhos para detectar os VP melhoram.

Com relação aos verdadeiros negativos, o teste JMO detecta bem apenas para $p = 10$ e sua detecção vai diminuindo até chegar a 0 quando p se aproxima de 100. O teste JMC também apresenta uma diminuição na detecção à medida que p aumenta, chegando a 0 apenas quando

Figura 4.5 – Simulação Monte Carlo com $n = 100$, $k = 3$, $\mu = 10$, $\delta = 0,8$, $\rho_1 = 0,10$, $\rho_2 = 0,5$, $\sigma^2 = 5$, $\delta = 0,8$ e $\alpha = 0,05$ para avaliar a taxa de verdadeiro positivo (VP) e falso negativo (FN) na detecção dos *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes número de variáveis p.



Fonte: Da autora (2024).

$\delta = 0,07$ e $\sigma^2 = 5$. Já os demais testes apresentam uma boa taxa de detecção dos VN para todas as configurações apresentadas.

4.3 Detecção de *outliers* em dados com populações próximas

Nesta seção são apresentadas as taxas utilizadas na detecção dos *outliers* e não *outliers* nos testes pré-existentes e propostos, considerando a situação em que a população não contaminante e contaminante estão próximas. Espera-se que a detecção dos *outliers* seja mais baixa devido a média da população contaminante estar próxima da não contaminante, quando comparados com os resultados anteriormente apresentados.

Pelos resultados da Tabela 4.6, para $n = 100$, $k = 1$ e populações sem correlações, pode-se observar que, para esta configuração, os testes apresentam baixa detecção de *outliers* para $p < 20$. Comparativamente os testes utilizando o estimador robusto *comedian* detectam melhor os *outliers* em relação aos testes originais. Para $p = 100$ todos os testes apresentam um bom desempenho na detecção dos verdadeiros positivos. Em relação aos verdadeiros negativos, a detecção decai para $p = 100$ no teste RO com $\delta = \{0,7,\ 0,8\}$ e nos testes JMO e JMC com $\delta = \{0,7,\ 0,8\}$ e $\sigma^2 = 5$. O teste RC se destaca para $p \geq 20$, pois detecta corretamente as taxas de VP e VN. Estes resultados obtidos para $\alpha = 0,05$ apresentaram padrões de respostas seme-

Tabela 4.5 – Simulação Monte Carlo com tamanho amostral $n = 500$, número de componentes retidos $k = 9$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0,9$ e $\rho_2 = 0,1$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
0,7	1	10	0,0456	0,7001	0,1157	0,6951	0,0615	0,6991	0,1328	0,6880
		20	0,1879	0,3551	0,2943	0,5985	0,2160	0,7001	0,2963	0,6993
		100	0,3000	0,0000	0,3001	0,0510	0,3001	0,6999	0,3001	0,6999
	5	10	0,0621	0,6994	0,1990	0,6970	0,0798	0,6994	0,2098	0,6928
		20	0,2349	0,0001	0,3002	0,5306	0,2566	0,6998	0,3002	0,6997
		100	0,3000	0,0000	0,3000	0,0054	0,3000	0,7000	0,3000	0,7000
0,8	1	10	0,0381	0,7986	0,0902	0,7884	0,0496	0,7946	0,1010	0,7753
		20	0,1548	0,5938	0,1974	0,7369	0,1677	0,8013	0,1979	0,7954
		100	0,2000	0,0014	0,2001	0,3175	0,2001	0,8000	0,2001	0,7998
	5	10	0,0571	0,7994	0,1415	0,7907	0,0690	0,7994	0,1479	0,7802
		20	0,1857	0,0092	0,2004	0,7126	0,1919	0,7996	0,2004	0,7969
		100	0,2004	0,0000	0,2004	0,1644	0,2004	0,7996	0,2004	0,7996
0,9	1	10	0,0218	0,8906	0,0510	0,8765	0,0277	0,8787	0,0559	0,8562
		20	0,0867	0,8212	0,0995	0,8526	0,0907	0,8961	0,0996	0,8745
		100	0,0998	0,3238	0,0998	0,7313	0,0998	0,8995	0,0998	0,8942
	5	10	0,0390	0,9002	0,0736	0,8792	0,0447	0,8990	0,0765	0,8600
		20	0,0995	0,3343	0,1008	0,8479	0,1002	0,8992	0,1008	0,8790
		100	0,1007	0,0000	0,1007	0,6667	0,1007	0,8992	0,1007	0,8960

Fonte: Da autora (2024).

lhantes aos apresentados na Tabela A.12 para o nível de significância $\alpha = 0,10$ considerando $n = 100$.

Nas Tabelas 12 e 13 são apresentados os resultados para $n = 300$ e $k = 3$, para cada nível nominal de significância α. Depreende-se que os resultados são semelhantes ao descrito para $n = 100$ e $k = 1$.

Dos resultados da Figura 4.6 para $n = 1000$, $k = 3$ e $\sigma^2 = 5$, verifica-se que os testes utilizando o estimador robusto *comedian* são melhores para detectar as taxas de VP e VN. Também, a taxa de detecção dos verdadeiros negativos decai nos testes JMO e JMC a medida que p aumenta. O teste RC se destaca na detecção dos VP e VN.

Nas tabelas 4.7 e 14 são apresentados os resultados para $n = 700$ e $k = 9$, para cada nível nominal de significância α. Observa-se novamente que os resultados são semelhantes aos obtidos anteriormente para $k = 1$ e $k = 3$. Uma possível razão para não haver alterações

Tabela 4.6 – Simulação Monte Carlo com tamanho amostral $n = 100$, número de componentes retidos $k = 1$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

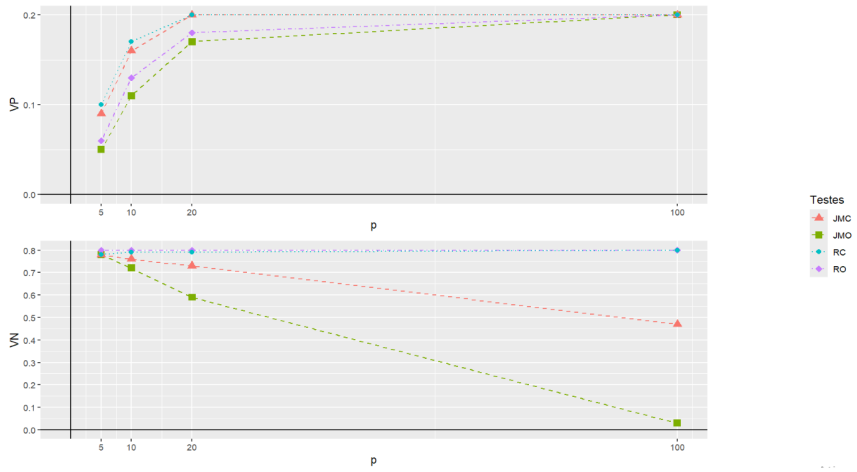
δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
0,7	1	2	0,0079	0,6822	0,0164	0,6740	0,0152	0,6656	0,0265	0,6555
		5	0,0119	0,6711	0,0199	0,6653	0,0139	0,6693	0,0318	0,6527
		10	0,0115	0,6719	0,0171	0,6672	0,0116	0,6718	0,0353	0,6464
		20	0,0098	0,6786	0,0194	0,6707	0,0073	0,6804	0,0440	0,6447
		100	0,0018	0,6978	0,0809	0,6472	0,1450	0,3594	0,1350	0,6441
	5	2	0,0382	0,6981	0,0823	0,6899	0,0537	0,6950	0,0991	0,6791
		5	0,0783	0,6452	0,1656	0,6620	0,1066	0,6987	0,1940	0,6880
		10	0,1248	0,5625	0,2350	0,6341	0,1628	0,6978	0,2605	0,6911
		20	0,1904	0,4056	0,2812	0,5714	0,2265	0,7017	0,2912	0,6984
		100	0,2953	0,0328	0,2994	0,1397	0,1481	0,3548	0,2994	0,7003
0,8	1	2	0,0058	0,7788	0,0155	0,7681	0,0117	0,7613	0,0235	0,7467
		5	0,0094	0,7669	0,0203	0,7603	0,0096	0,7657	0,0319	0,7449
		10	0,0078	0,7662	0,0218	0,7612	0,0073	0,7651	0,0402	0,7390
		20	0,0067	0,7741	0,0340	0,7640	0,0048	0,7756	0,0585	0,7394
		100	0,0013	0,7959	0,1074	0,7278	0,1020	0,3923	0,1477	0,7423
	5	2	0,0323	0,7955	0,0638	0,7824	0,0435	0,7891	0,0756	0,7677
		5	0,0696	0,7537	0,1269	0,7610	0,0893	0,7948	0,1442	0,7741
		10	0,1123	0,7073	0,1740	0,7484	0,1349	0,7968	0,1849	0,7799
		20	0,1595	0,6217	0,1951	0,7176	0,1730	0,7997	0,1975	0,7880
		100	0,1980	0,2745	0,1982	0,4191	0,1002	0,3960	0,1982	0,7981
0,9	1	2	0,0043	0,8795	0,0110	0,8654	0,0077	0,8590	0,0158	0,8407
		5	0,0051	0,8631	0,0164	0,8527	0,0061	0,8612	0,0250	0,8351
		10	0,0044	0,8655	0,0239	0,8574	0,0045	0,8658	0,0369	0,8334
		20	0,0032	0,8731	0,0371	0,8604	0,0021	0,8749	0,0530	0,8316
		100	0,0016	0,8943	0,0862	0,8367	0,0509	0,4489	0,0956	0,8207
	5	2	0,0199	0,8887	0,0351	0,8710	0,0257	0,8759	0,0401	0,8489
		5	0,0460	0,8634	0,0718	0,8556	0,0544	0,8851	0,0788	0,8504
		10	0,0692	0,8523	0,0907	0,8577	0,0756	0,8910	0,0940	0,8570
		20	0,0897	0,8315	0,0989	0,8488	0,0905	0,8931	0,0993	0,8611
		100	0,0994	0,7700	0,0994	0,7441	0,0498	0,4515	0,0994	0,8719

Fonte: Da autora (2024).

significativas nas taxas de VP e VN ao aumentar o número de componentes retidos k é a falta de correlações entre as variáveis das simulações realizadas.

Nas Tabelas 4.8 e 15 são apresentados os resultados com populações correlacionadas e próximas, para $\alpha \in \{0,05,\ 0,10\}$, respectivamente. Para $\sigma^2 = 1$, o teste JMO apresenta

Figura 4.6 – Simulação Monte Carlo com $n = 1000$, $k = 3$, $\mu = 1$, $\delta = 0,8$, $\rho_1 = 0$, $\rho_2 = 0$, $\sigma^2 = 5$, $\delta = 0,8$ e $\alpha = 0,05$ para avaliar a taxa de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes número de variáveis p.



Fonte: Da autora (2024).

uma taxa melhor de VP do que o teste JMC. Para $\sigma^2 = 5$, os testes propostos detectam melhor os *outliers* em relação aos testes originais. Considerando-se $p = 100$, o teste JMO é o que mais se destaca na detecção dos *outliers*, diferentemente das outras simulações apresentadas anteriormente, em que para $p = 100$ todos os testes identificavam corretamente os *outliers*.

Em relação aos verdadeiros negativos nas Tabelas 4.8 e 15 , para $\sigma^2 = 1$ a medida que p aumenta, a taxa de detecção decresce. Para $\sigma^2 = 5$, as taxas permanecem constantes e apresentam uma boa detecção nos testes, exceto para o teste RC, que apresenta uma pequena queda, mas menor do que em $\sigma^2 = 1$.

Nas Tabelas 16 e 17 são apresentados os resultados para $n = 900$, $k = 3$ e correlações $\rho_1 = 0$ e $\rho_2 = 0,5$. O teste JMO supera o teste JMC na detecção dos verdadeiros positivos para $p > 2$ em $\sigma^2 = 1$. Para $\sigma^2 = 5$ os testes utilizando o estimador robusto *comedian* são melhores para identificar os *outliers* do que os testes originais. Em relação aos verdadeiros negativos, todos os testes decrescem a taxa de detecção a medida que p aumenta. Para $p = 100$ e $\sigma^2 = 1$ apenas os testes JMO e JMC detectam corretamente os VP.

Na Tabela 4.9 são apresentados os resultados para $n = 300$, $k = 9$ e correlações $\rho_1 = 0,9$ e $\rho_2 = 0,1$. Os testes utilizando o estimador robusto *comedian* detectam melhor os *outliers* do que os testes originais. Para $p = 100$, todos os testes detectam corretamente os VP. O teste JMO para $\alpha \in \{0,7,\ 0,8\}$ detecta corretamente os *outliers* apenas em $p = 10$ e em seguida

decresce a sua detecção, chegando a 0 em alguns casos. O teste JMC também apresenta uma queda na detecção dos verdadeiros negativos, mas não chega a 0 como o teste JMC. Os testes RO e RC apresentam um bom desempenho na detecção dos verdadeiros positivos e verdadeiros negativos. Entretanto, o teste de RC supera o RO.

Estes resultados obtidos para dados correlacionados, são semelhantes aos obtidos por Sajesh e Srinivasan (2012), pois, segundo os autores o método Comedian se mostrou eficaz na detecção dos *outliers* em dados correlacionados, com resultados semelhantes aos métodos que preservam a propriedade afim-equivariância.

Tabela 4.7 – Simulação Monte Carlo com tamanho amostral $n = 700$, número de componentes retidos $k = 9$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0$, e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
$0,7$	1	10	0,0078	0,6818	0,0133	0,6705	0,0150	0,6654	0,0225	0,6500
		20	0,0144	0,6669	0,0203	0,6566	0,0143	0,6664	0,0284	0,6428
		100	0,0110	0,6741	0,0180	0,6655	0,0109	0,6733	0,0392	0,6517
	5	10	0,0362	0,6980	0,0791	0,6909	0,0512	0,6941	0,0963	0,6809
		20	0,1488	0,5427	0,2584	0,6292	0,1787	0,6989	0,2711	0,6956
		100	0,2992	0,0033	0,2997	0,1965	0,2995	0,7003	0,2997	0,7003
$0,8$	1	10	0,0052	0,7797	0,0093	0,7682	0,0099	0,7605	0,01563	0,7450
		20	0,0095	0,7622	0,0156	0,7514	0,0096	0,7617	0,0225	0,7393
		100	0,0073	0,7704	0,0253	0,7591	0,0071	0,7695	0,0486	0,7514
	5	10	0,0290	0,7950	0,0609	0,7844	0,0396	0,7871	0,0725	0,7694
		20	0,1255	0,7007	0,1836	0,7462	0,1424	0,7992	0,1889	0,7879
		100	0,1996	0,1026	0,1996	0,5028	0,1996	0,8004	0,1996	0,8000
$0,9$	1	10	0,0026	0,8765	0,0061	0,8640	0,0049	0,8550	0,0096	0,8383
		20	0,0048	0,8573	0,0134	0,8452	0,0048	0,8568	0,0194	0,8352
		100	0,0037	0,8665	0,0501	0,8526	0,0031	0,8658	0,06017	0,8510
	5	10	0,0169	0,8867	0,0351	0,8734	0,0226	0,8722	0,0409	0,8521
		20	0,0732	0,8456	0,0962	0,8502	0,0800	0,8901	0,0976	0,8664
		100	0,1005	0,6222	0,1005	0,7864	0,1005	0,8984	0,1005	0,8913

Fonte: Da autora (2024).

4.4 Considerações gerais

Os resultados das simulações apresentadas, mostram que para os dados sem correlação e com populações distantes, $\mu = 10$, independentemente do número de componentes principais

Tabela 4.8 – Simulação Monte Carlo com tamanho amostral $n = 250$, número de componentes retidos $k = 1$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0,1$ e $\rho_2 = 0,9$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
0,7	1	2	0,0000	0,6617	0,0009	0,6194	0,0000	0,6367	0,0017	0,5864
		5	0,1006	0,6372	0,0530	0,5620	0,0000	0,6092	0,0006	0,4970
		10	0,2566	0,6195	0,1592	0,4916	0,0000	0,5752	0,0004	0,3888
		20	0,2978	0,5901	0,2492	0,3738	0,0000	0,5210	0,0004	0,2408
		100	0,3001	0,4219	0,2906	0,0270	0,0000	0,2361	0,0001	0,0051
	5	2	0,0021	0,6734	0,0150	0,6720	0,0054	0,6534	0,0221	0,6526
		5	0,0141	0,6580	0,0265	0,6601	0,0020	0,6443	0,0248	0,6437
		10	0,0255	0,6568	0,0348	0,6613	0,0008	0,6383	0,0251	0,6360
		20	0,0487	0,6555	0,0496	0,6610	0,0002	0,6290	0,0239	0,6220
		100	0,1904	0,6427	0,1604	0,6522	0,0000	0,6301	0,0342	0,5436
0,8	1	2	0,0000	0,7674	0,0018	0,7431	0,0000	0,7430	0,0031	0,7131
		5	0,0718	0,7458	0,0400	0,7073	0,0000	0,7244	0,0010	0,6550
		10	0,1763	0,7393	0,1153	0,6799	0,0000	0,7079	0,0005	0,5934
		20	0,1995	0,7293	0,1733	0,6214	0,0000	0,6823	0,0002	0,4891
		100	0,2002	0,6734	0,1976	0,2824	0,0000	0,5555	0,0001	0,1116
	5	2	0,0013	0,7749	0,0187	0,7742	0,0031	0,7536	0,0250	0,7532
		5	0,0096	0,7574	0,0286	0,7585	0,0013	0,7477	0,0297	0,7455
		10	0,0194	0,7565	0,0318	0,7591	0,0003	0,7433	0,0279	0,7400
		20	0,0376	0,7576	0,0396	0,7606	0,0001	0,7376	0,0251	0,7310
		100	0,1420	0,7650	0,1120	0,7669	0,0000	0,7558	0,0301	0,6833
0,9	1	2	0,0000	0,8706	0,0031	0,8587	0,0000	0,84644	0,0049	0,8314
		5	0,0334	0,8527	0,0180	0,8401	0,0000	0,8406	0,0028	0,8101
		10	0,0873	0,8518	0,0510	0,8379	0,0000	0,8356	0,0010	0,7895
		20	0,0996	0,8520	0,0812	0,8291	0,0000	0,8285	0,0003	0,7507
		100	0,0998	0,8592	0,0982	0,7789	0,0000	0,8278	0,0001	0,5538
	5	2	0,0007	0,8744	0,0173	0,8709	0,0016	0,8518	0,0212	0,8478
		5	0,0052	0,8556	0,0215	0,8538	0,0006	0,8485	0,0233	0,8428
		10	0,0104	0,8572	0,0214	0,8563	0,0002	0,8492	0,0209	0,8419
		20	0,0215	0,8600	0,0243	0,8584	0,0001	0,8499	0,0195	0,8367
		100	0,0766	0,8774	0,0567	0,8675	0,0000	0,8755	0,0228	0,8123

Fonte: Da autora (2024).

retidos k, os testes utilizando o estimador robusto *comedian* são melhores para identificar os *outliers*. E o teste RC se destaca também na identificação dos não *outliers*. Em um caso de alta dimensionalidade, os testes originais conseguem detectar corretamente os *outliers* para $\sigma^2 = 5$. Mas, a detecção dos não *outliers* é uma das menores registradas.

Tabela 4.9 – Simulação Monte Carlo com tamanho amostral $n = 300$, número de componentes retidos $k = 9$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0,9$ e $\rho_2 = 0,1$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
0,7	1	10	0,0435	0,6993	0,1095	0,6914	0,0594	0,6975	0,1263	0,6821
		20	0,1794	0,4001	0,2882	0,6143	0,2103	0,7008	0,2922	0,6965
		100	0,3008	0,0000	0,3010	0,1135	0,3010	0,6990	0,3010	0,6975
	5	10	0,0608	0,6997	0,1955	0,6961	0,0780	0,6997	0,2067	0,6908
		20	0,2296	0,0011	0,2991	0,5415	0,2551	0,7009	0,2991	0,7005
		100	0,2992	0,0000	0,2992	0,0199	0,2992	0,7008	0,2992	0,7008
0,8	1	10	0,0354	0,7963	0,0830	0,7830	0,0465	0,7911	0,0941	0,7680
		20	0,1489	0,6263	0,1966	0,7406	0,1641	0,8001	0,1978	0,7874
		100	0,2011	0,0087	0,2011	0,3644	0,2011	0,7988	0,2011	0,7931
	5	10	0,0557	0,7993	0,1395	0,7887	0,0678	0,7992	0,1459	0,7770
		20	0,1818	0,0372	0,1989	0,7179	0,1900	0,8011	0,1989	0,7960
		100	0,2009	0,0000	0,2009	0,1846	0,2009	0,7991	0,2009	0,7982
0,9	1	10	0,0186	0,8888	0,0460	0,8724	0,0243	0,8754	0,0512	0,8498
		20	0,0806	0,8359	0,0996	0,8495	0,0857	0,8923	0,0998	0,8643
		100	0,0988	0,5149	0,0989	0,7225	0,0956	0,8976	0,0989	0,8771
	5	10	0,0360	0,8978	0,0734	0,8747	0,0419	0,8941	0,0763	0,8540
		20	0,0982	0,4946	0,1003	0,8478	0,0995	0,8993	0,1003	0,8740
		100	0,0998	0,0043	0,0998	0,6550	0,0987	0,8986	0,0998	0,8892

Fonte: Da autora (2024).

Para a população contaminante correlacionada e populações distantes, os testes utilizando o estimador robusto *comedian* apresentam um bom desempenho na detecção dos *outliers*. Para $\sigma^2 = 1$, o teste JMO apresenta uma melhor detecção dos *outliers* do que para $\sigma^2 = 5$ e em relação ao teste RO.

Quando a população não contaminante está correlacionada e a população contaminante está distante, observa-se que os testes utilizando o estimador robusto *comedian* apresentam um melhor desempenho na detecção dos *outliers*. O teste de RC se destaca na detecção dos *outliers* e não *outliers* em relação aos demais.

Em relação as situações em que as as populações estão próximas, $\mu = 1$, e não correlacionadas,os testes apresentam baixa detecção de *outliers* para $p < 20$. O teste de RC se destaca na detecção dos verdadeiros positivos e verdadeiros negativos. Nos casos em que as populações

estão próximas e correlacionadas, a taxa de detecção dos *outliers* é relativamente baixa, quando comparada com os casos anteriores.

5 CONCLUSÃO

Os testes propostos utilizando o estimador robusto *comedian* obtiveram um melhor desempenho em todas as simulações realizadas na detecção dos *outliers*.

O teste JMC embora se destaque na detecção dos *outliers*, na medida que p aumenta a detecção dos não *outliers* diminui.

Quando a média da população contaminante e da não contaminante estão próximas, a taxa de detecção dos *outliers* diminui nos testes existentes e propostos.

Em relação aos falsos negativos, FN, os testes utilizando o estimador robusto *comedian* apresentaram menores taxas em relação a dos testes originais.

O teste RC se destaca como sendo o melhor para detectar os *outliers* e não *outliers* com base nas configurações simuladas.

REFERÊNCIAS

BARBOSA, J. J.; PEREIRA, T. M.; OLIVEIRA, F. L. P. de. Uma proposta para identificação de outliers multivariados. **Ciência e Natura**, v. 40, 2018.

BARNETT, V.; LEWIS, T. et al. **Outliers in statistical data**. New York: Wiley, 1994. v. 3.

BORGES, L. C. **Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student Newman sob distribuições normal e não normais dos resíduos**. 94 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras, 2002.

CASELLA, G.; BERGER, R. L. **Inferência estatística-tradução da 2a edição norteamericana.** São Paulo: Centage Learning, 2014.

FALK, M. On mad and comedians. **Annals of the Institute of Statistical Mathematics**, Springer, v. 49, p. 615–644, 1997.

FERREIRA, D. F. **Estatística computacional em Java**. 1º. ed. Lavras: Editora UFLA, 2013.

FERREIRA, D. F. **Estatística Multivariada**. 3º. ed. Lavras: UFLA, 2018.

HADI, A. S. Identifying multiple outliers in multivariate data. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, Oxford University Press, v. 54, n. 3, p. 761–771, 1992.

HAMPEL, F. R. A general qualitative definition of robustness. **The annals of mathematical statistics**, Institute of Mathematical Statistics, v. 42, n. 6, p. 1887–1896, 1971. Acesso em: abr. 2024. Disponível em: <:<https://www.jstor.org/stable/2240114>>.

HAMPEL, F. R.; ROUSSEEUW, P. J.; RONCHETTI, E. M.; STAHEL, W. A. **Robust statistics: the approach based on influence functions**. New York: Wiley, 2011. v. 196.

HAWKINS, D. M. **Identification of outliers**. London: Springer, 1980. v. 11.

JACKSON, J. E.; MUDHOLKAR, G. S. Control procedures for residuals associated with principal component analysis. **Technometrics**, Taylor & Francis, v. 21, n. 3, p. 341–349, 1979.

JOHNSON, R.; WICHERN, D. **Applied Multivariate Statistical Analysis.** New Jersey: Prentice Hall International, 1998. 816 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Correspondence Analysis**. USA: Prentice-Hall,Upper Saddle River, NJ, 2007.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6º. ed. New Jersey: Pearson, 2007.

JOLLIFE, I. Principal component analysis. **Wiley Online Library**, 2002.

LANDAU, D.; BINDER, K. **A Guide to Monte Carlo Simulations in Statistical Physics**. 4º. ed. New York: Cambridge University Press, 2014. ISBN 1107074029,9781107074026.

MAECHLER, M. et al. **robustbase: Basic Robust Statistics**. 2021. Disponível em: <http://robustbase.r-forge.r-project.org/>.

MARONNA, R. A.; ZAMAR, R. H. Robust estimates of location and dispersion for high-dimensional datasets. **Technometrics**, Taylor & Francis, v. 44, n. 4, p. 307–317, 2002. Acesso em: fev. 2024. Disponível em: <https://doi.org/10.1198/004017002188618509>.

MARTINS, H. M. **Métodos para detecção de outliers multivariados: via uso dos estimadores robustos**. 93 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) — Universidade Federal de Lavras, Lavras, 2022.

MORETTIN, P. A.; BUSSAB, W. de O. **Estatística Básica**. 6º. ed. São Paulo: Saraiva, 2010. ISBN 978-85-02-08177-2.

PALMA, M. A. D.; GALLO, M. A co-median approach to detect compositional outliers. **Journal of Applied Statistics**, Taylor & Francis, v. 43, n. 13, p. 2348–2362, 2016.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>.

RAO, C. R. The use and interpretation of principal component analysis in applied research. **Sankhyā: The Indian Journal of Statistics, Series A**, JSTOR, v. 26, p. 329–358, 1964.

ROCKE, D. M.; WOODRUFF, D. L. Identification of outliers in multivariate data. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, n. 435, p. 1047–1061, 1996. Acesso em: abr. 2024. Disponível em: <http://dx.doi.org/10.1080/01621459.1996.10476975>.

SAJESH, T.; SRINIVASAN, M. Outlier detection for high dimensional data using the comedian approach. **Journal of statistical computation and simulation**, Taylor & Francis, v. 82, n. 5, p. 745–757, 2012. Acesso em: fev. 2024. Disponível em: <https://doi.org/10.1080/00949655.2011.552504>.

SATHE, S.; AGGARWAL, C. C. Subspace histograms for outlier detection in linear time. **Knowledge and Information Systems**, Springer, v. 56, p. 691–715, 2018.

SILVA, R. B. V. **Extensão do teste de normalidade de shapiro-francia para o caso multivariado**. Tese (Doutorado) — Universidade Federal de Lavras, 2009.

VIEIRA, L. V. **Modelos paramétricos de matrizes de covariância para medidas repetidas: um estudo de simulação sobre o ajuste, o erro e o poder estatístico em modelos lineares mistos**. 147 p. Tese (Doutorado) — Universidade Estadual Paulista, Botucatu, 2021.

APÊNDICE A – Resultados das simulações

Tabela 1 – Simulação Monte Carlo com tamanho amostral $n = 50$, número de componenets principais retidos $k = 1$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não outliers δ , variância σ^2 e número de variáveis p .

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		2	0,0139	0,6633	0,2026	0,6808	0,0271	0,6296	0,2147	0,6627
		5	0,0233	0,6452	0,2854	0,60954	0,0244	0,647	0,2866	0,6813
	1	10	0,0208	0,6465	0,3013	0,4799	0,0187	0,6435	0,3012	0,6814
		20	0,0152	0,6670	0,2958	0,2651	0,0072	0,6725	0,2956	0,6955
		100	0,0013	0,6990	0,2972	0,0129	0,2986	0,0055	0,2975	0,7000
$0,7$		2	0,0570	0,6951	0,2335	0,6887	0,0803	0,6850	0,2424	0,6745
		5	0,1031	0,5960	0,2929	0,5972	0,1393	0,6937	0,2955	0.6865
	5	10	0,1428	0,5071	0,2962	0,4734	0,1883	0,6991	0,2963	0,6959
		20	0,1977	0,3661	0,2996	0,2427	0,2228	0,6974	0,2998	0,6970
		100	0,2915	0,0877	0,3012	0,0083	0,3012	0,0048	0,3012	0,6985
		2	0,0083	0,7593	0,1712	0,7719	0,0167	0,7199	0,1744	0,7469
		5	0,0156	0,7316	0,2004	0,7081	0,0142	0,7253	0,2005	0,7579
	1	10	0,0140	0,7452	0,1983	0,6545	0,0100	0,7398	0,1983	0,7691
		20	0,0101	0,7585	0,2022	0,5106	0,0032	0,7635	0,2021	0,7762
		100	0,0020	0,7970	0,2003	0,0995	0,2003	0,0050	0,2003	0,7920
$0,8$		2	0,0450	0,7862	0,1754	0,7730	0,0611	0,7701	0,1787	0,7473
		5	0,0872	0,7091	0,2022	0,7049	0,1117	0,7819	0,2023	0,7596
	5	10	0,1254	0,6583	0,2025	0,6437	0,1474	0,7866	0,2025	0,7697
		20	0,1612	0,6019	0,1965	0,5201	0,1583	0,7959	0,1965	0,7832
		100	0,1967	0,4199	0,1974	0,1134	0,1974	0.0050	0,1974	0,7940
		2	0,0033	0,8545	0,0906	0,8438	0,0069	0,8110	0,0914	0,8066
		5	0,0074	0,8248	0,0986	0,8091	0,0038	0,8177	0,0986	0,8118
	1	10	0,0088	0,8350	0,1002	0,7996	0,0022	0,8293	0,1002	0,8149
		20	0,0097	0,8555	0,0996	0,7612	0,0005	0,8602	0,0996	0,8202
		100	0,0094	0,8949	0,1020	0,4460	0,1020	0,0074	0,1020	0,8343
$0,9$		2	0,0253	0,8710	0,0937	0,8457	0,0341	0,8402	0,0948	0,8068
		5	0,0516	0,8270	0,1003	0,8069	0,0596	0,8555	0,1003	0,8122
	5	10	0,0699	0,8256	0,0964	0,8052	0,0717	0,8659	0,0964	0,8208
		20	0,0905	0,8155	0,1003	0,7604	0,0674	0,8767	0,1003	0,8241
		100	0,1006	0,8199	0,1006	0,4504	0,1006	0,0070	0,1006	0,8373

Fonte: Da autora (2024).

Tabela 2 – Simulação Monte Carlo com tamanho amostral $n = 200$, número de componentes principais retidos $k = 1$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		2	0,0078	0,6818	0,1301	0,6892	0,0147	0,6648	0,1437	0,6793
		5	0,0128	0,6683	0,2704	0,6651	0,0137	0,6667	0,2730	0,6888
	1	10	0,0131	0,6718	0,2956	0,6224	0,0125	0,6706	0,2956	0,6965
		20	0,0114	0,6715	0,3003	0,4997	0,0112	0,6715	0,3002	0,6972
		100	0,0048	0,6870	0,3016	0,0154	0,0012	0,6941	0,3016	0,6983
0,7		2	0,0416	0,7000	0,1753	0,6972	0,0574	0,6979	0,1881	0,6916
		5	0,0870	0,6397	0,2859	0,6559	0,1133	0,7001	0,2892	0,6978
	5	10	0,1383	0,5389	0,2998	0,5846	0,1723	0,6996	0,3000	0,6988
		20	0,2091	0,3414	0,2999	0,4120	0,2406	0,7001	0,2999	0,7000
		100	0,3002	0,0037	0,3009	0,0050	0,3001	0,6991	0,3009	0,6991
		2	0,0048	0,7798	0,1376	0,7890	0,0094	0,7607	0,1424	0,7771
		5	0,0087	0,7636	0,1973	0,7609	0,0089	0,7615	0,1977	0,7872
	1	10	0,0087	0,7651	0,2002	0,7268	0,0084	0,7638	0,2002	0,7928
		20	0,0078	0,7687	0,2001	0,6502	0,0069	0,7678	0,2001	0,7968
		100	0,0032	0,7856	0,2016	0,1411	0,0005	0,7930	0,2016	0,7982
0,8		2	0,0360	0,7971	0,1590	0,7918	0,0472	0,7921	0,1633	0,7820
		5	0,0771	0,7514	0,1972	0,7599	0,0950	0,7990	0,1976	0,7928
	5	10	0,1226	0,6920	0,1995	0,7201	0,1422	0,7995	0,1995	0,7963
		20	0,1695	0,5772	0,1982	0,6330	0,1810	0,8014	0,1982	0,8002
		100	0,1990	0,1048	0,1991	0,1221	0,1990	0,8009	0,1991	0,8009
		2	0,0022	0,8762	0,0866	0,8800	0,0045	0,8549	0,0878	0,8627
		5	0,0041	0,8592	0,1004	0,8601	0,0039	0,8561	0,1004	0,8718
	1	10	0,0041	0,8612	0,0988	0,8511	0,0034	0,8604	0,0988	0,8798
		20	0,0040	0,8644	0,0997	0,8279	0,0027	0,8631	0,0997	0,8848
		100	0,0022	0,8858	0,0997	0,6176	0,0001	0,8942	0,0997	0,8955
0,9		2	0,0232	0,8920	0,0888	0,8818	0,0292	0,8808	0,0900	0,8640
		5	0,0503	0,8625	0,0998	0,8606	0,0590	0,8884	0,0999	0,8729
	5	10	0,0758	0,8429	0,0993	0,8506	0,0824	0,8929	0,0993	0,8807
		20	0,0957	0,8039	0,1008	0,8245	0,0977	0,8943	0,1008	0,8849
		100	0,1010	0,6108	0,1010	0,6058	0,1010	0,8979	0,1010	0,8953

Fonte: Da autora (2024).

Tabela 3 – Simulação Monte Carlo com tamanho amostral $n = 200$, número de componentes principais retidos $k = 1$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		2	0,0155	0,6628	0,1450	0,6780	0,0296	0,6306	0,1638	0,6563
		5	0,0278	0,6338	0,2763	0,6270	0,0288	0,6315	0,2794	0,6755
	1	10	0,0272	0,6383	0,2962	0,5588	0,0270	0,6360	0,2963	0,6884
		20	0,0244	0,6395	0,3003	0,3990	0,0242	0,6372	0,3002	0,6934
		100	0,0131	0,6678	0,3016	0,0058	0,0070	0,6781	0,3016	0,6983
0,7		2	0,0592	0,6974	0,1895	0,6909	0,0822	0,6898	0,2046	0,6757
		5	0,1151	0,5805	0,2892	0,6104	0,1452	0,6985	0,2923	0,6921
	5	10	0,1697	0,4439	0,3000	0,5058	0,2057	0,6993	0,3002	0,6970
		20	0,2353	0,2323	0,2999	0,3042	0,2631	0,7000	0,2999	0,6995
		100	0,3006	0,0009	0,3009	0,0018	0,3008	0,6991	0,3009	0,6991
		2	0,0100	0,7583	0,1429	0,7753	0,0193	0,7206	0,1492	0,7492
		5	0,0187	0,7234	0,1978	0,7176	0,0185	0,7215	0,1983	0,7708
	1	10	0,0183	0,7266	0,2002	0,6682	0,0181	0,7244	0,2002	0,7822
		20	0,0170	0,7331	0,2001	0,5603	0,0157	0,7294	0,2001	0,7918
		100	0,0086	0,7639	0,2016	0,0778	0,0034	0,7740	0,2016	0,7980
0,8		2	0,0483	0,7912	0,1638	0,7807	0,0643	0,7759	0,1687	0,7577
		5	0,0957	0,7000	0,1976	0,7144	0,1162	0,7927	0,1979	0,7791
	5	10	0,1404	0,6129	0,1995	0,6575	0,1593	0,7969	0,1995	0,7890
		20	0,1793	0,4666	0,1982	0,5384	0,1884	0,8004	0,1982	0,7971
		100	0,1991	0,0484	0,1991	0,0666	0,1991	0,8002	0,1991	0,8008
		2	0,0048	0,8522	0,0879	0,8604	0,0093	0,8101	0,0893	0,8250
		5	0,0092	0,8141	0,1004	0,8137	0,0084	0,8115	0,1004	0,8412
	1	10	0,0084	0,8182	0,0988	0,8033	0,0075	0,8151	0,0988	0,8548
		20	0,0086	0,8242	0,0997	0,7690	0,0064	0,8193	0,0997	0,8665
		100	0,0055	0,8616	0,0997	0,5006	0,0008	0,8731	0,0997	0,8895
0,9		2	0,0298	0,8791	0,0901	0,8618	0,0373	0,8516	0,0913	0,8267
		5	0,0595	0,8177	0,0999	0,8147	0,0686	0,8693	0,0999	0,8424
	5	10	0,0823	0,7891	0,0993	0,8017	0,0882	0,8796	0,0993	0,8565
		20	0,0979	0,7321	0,1008	0,7650	0,0993	0,8850	0,1008	0,8678
		100	0,1010	0,4712	0,1010	0,4871	0,1010	0,8927	0,1010	0,8902

Fonte: Da autora (2024).

Tabela 4 – Simulação Monte Carlo com tamanho amostral $n = 150$, número de componentes principais retidos $k = 3$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,0258	0,6378	0,1618	0,6444	0,02835	0,6305	0,1762	0,6506
		10	0,0269	0,6373	0,2831	0,6041	0,0265	0,6353	0,2853	0,6781
	1	20	0,0240	0,6452	0,2970	0,4527	0,0234	0,6421	0,2971	0,6928
		100	0,0102	0,6755	0,3017	0,0057	0,0013	0,6931	0,3017	0,6983
		5	0,0758	0,6467	0,2262	0,6449	0,0989	0,6877	0,2395	0,6738
0,7		10	0,1365	0,5391	0,2974	0,5652	0,1725	0,6966	0,2989	0,6921
	5	20	0,2124	0,3507	0,2988	0,3716	0,2444	0,7006	0,2988	0,7000
		100	0,2971	0,0088	0,2981	0,0031	0,2484	0,7016	0,2981	0,7019
		5	0,0167	0,7310	0,1686	0,7344	0,0188	0,7232	0,1730	0,7451
		10	0,0175	0,7292	0,1972	0,6981	0,0165	0,7259	0,1973	0,7717
	1	20	0,0162	0,7333	0,2019	0,5877	0,0148	0,7291	0,2019	0,7854
		100	0,0071	0,7736	0,1999	0,0891	0,0004	0,7938	0,1999	0,7995
		5	0,0592	0,7461	0,1843	0,7351	0,0751	0,7762	0,1871	0,7547
0,8		10	0,1115	0,6820	0,1996	0,6867	0,1328	0,7907	0,1997	0,7788
	5	20	0,1652	0,5727	0,1997	0,5750	0,1789	0,7956	0,1997	0,7913
		100	0,2011	0,1464	0,2012	0,07921	0,1500	0,7977	0,2012	0,7984
		5	0,0082	0,8220	0,0964	0,8141	0,0087	0,8110	0,0969	0,8126
		10	0,0094	0,8154	0,1027	0,7996	0,0078	0,8115	0,1027	0,8318
	1	20	0,0085	0,8276	0,1002	0,7789	0,0057	0,8219	0,1002	0,8533
		100	0,0049	0,8693	0,1004	0,5132	0,0000	0,8925	0,1004	0,8832
		5	0,0336	0,8315	0,0999	0,8118	0,0418	0,8436	0,1003	0,8142
0,9		10	0,0628	0,8134	0,1009	0,8007	0,0715	0,8616	0,1009	0,8344
	5	20	0,0876	0,7916	0,0999	0,7780	0,0905	0,8723	0,0999	0,8538
		100	0,1004	0,6673	0,1012	0,5064	0,0427	0,8945	0,1012	0,8846

Fonte: Da autora (2024).

Tabela 5 – Simulação Monte Carlo com tamanho amostral $n = 700$, número de componentes principais retidos $k = 9$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
$0,7$	1	10	0,0161	0,6636	0,0279	0,6579	0,0301	0,6308	0,0455	0,6240
		20	0,0290	0,6319	0,1024	0,6304	0,0292	0,6312	0,1282	0,6414
		100	0,0238	0,6442	0,2997	0,3476	0,0242	0,6414	0,2996	0,6955
	5	10	0,0530	0,6931	0,1114	0,6822	0,0754	0,6810	0,1338	0,6612
		20	0,1787	0,4519	0,2909	0,5631	0,2101	0,7000	0,2944	0,6947
		100	0,3007	0,0009	0,3009	0,0327	0,3009	0,6991	0,3009	0,6991
$0,8$	1	10	0,0104	0,7576	0,0310	0,7529	0,0198	0,7195	0,0449	0,7151
		20	0,0198	0,7227	0,1834	0,7136	0,0193	0,7214	0,1872	0,7526
		100	0,0160	0,7364	0,1995	0,3066	0,0153	0,7329	0,1995	0,7987
	5	10	0,0414	0,7850	0,0979	0,7714	0,0569	0,7642	0,1113	0,7428
		20	0,1430	0,6264	0,2000	0,6783	0,1600	0,7960	0,2001	0,7836
		100	0,2004	0,0503	0,2005	0,1569	0,2005	0,7995	0,2005	0,7995
$0,9$	1	10	0,0051	0,8524	0,0516	0,8500	0,0099	0,8100	0,0575	0,8086
		20	0,0097	0,8128	0,1001	0,8007	0,0094	0,8112	0,1001	0,8448
		100	0,008	0,8277	0,1001	0,6013	0,0069	0,8237	0,1001	0,8913
	5	10	0,0230	0,8710	0,0724	0,8570	0,0308	0,8378	0,0763	0,8178
		20	0,0803	0,7945	0,1004	0,7969	0,0868	0,8740	0,1004	0,8528
		100	0,0998	0,5048	0,0998	0,5748	0,0998	0,8962	0,0998	0,8942

Fonte: Da autora (2024).

Tabela 6 – Simulação Monte Carlo com tamanho amostral $n = 1000$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0$ e $\rho_2 = 0,9$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ , variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,0711	0,6008	0,0322	0,5303	0,0000	0,5707	0,0019	0,4784
	1	10	0,2888	0,5659	0,0887	0,4885	0,0000	0,5134	0,0128	0,3688
		20	0,3001	0,5116	0,1570	0,5089	0,0000	0,4338	0,0813	0,2670
		100	0,3000	0,2189	0,2986	0,5582	0,0000	0,1217	0,2898	0,0306
0,7		5	0,0211	0,6210	0,0285	0,6189	0,0068	0,6017	0,0274	0,5987
	5	10	0,0521	0,6087	0,0716	0,6239	0,0017	0,5791	0,0658	0,5961
		20	0,1084	0,5946	0,1763	0,6357	0,0003	0,5482	0,1590	0,6014
		100	0,2907	0,4976	0,2983	0,5570	0,0000	0,4105	0,2954	0,6239
		5	0,0526	0,7042	0,0438	0,6844	0,0000	0,6787	0,0345	0,6470
	1	10	0,1960	0,6854	0,1478	0,7079	0,0000	0,6438	0,1383	0,6352
		20	0,1998	0,6604	0.,1964	0,7260	0,0000	0,5960	0,1948	0,6282
		100	0,1997	0,4940	0,1997	0,6742	0,0000	0,3622	0,1997	0,5708
0,8		5	0,0259	0,7091	0,1777	0,7286	0,0015	0,6913	0,1794	0,7438
	5	10	0,0507	0,7052	0,1976	0,7158	0,0004	0,6788	0,1976	0,7539
		20	0,0964	0,6995	0,1998	0,6831	0,0000	0,6604	0,1998	0,7652
		100	0,1974	0,6604	0,1995	0,4294	0,0000	0,5829	0,1995	0,7900
		5	0,0286	0,8087	0,0789	0,8193	0,0000	0,7900	0,0792	0,8136
	1	10	0,0991	0,7970	0,0999	0,8118	0,0000	0,7734	0,0999	0,8310
		20	0,1002	0,7924	0,1002	0,7978	0,0000	0,7544	0,1002	0,8486
		100	0,1001	0,7556	0,1001	0,6441	0,0000	0,6646	0,1001	0,8817
0,9		5	0,0143	0,8064	0,0996	0,8183	0,0005	0,7963	0,0997	0,8447
	5	10	0,0289	0,8044	0,0999	0,8069	0,0001	0,7899	0,0999	0,8559
		20	0,0555	0,8042	0,1006	0,7831	0,0000	0,7812	0,1006	0,8686
		100	0,1003	0,8016	0,1006	0,6034	0,0000	0,7527	0,1006	0,8909

Fonte: Da autora (2024).

Tabela 7 – Simulação Monte Carlo com tamanho amostral $n = 500$, número de componentes retidos $k = 9$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0,9$ e $\rho_2 = 0,1$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		10	0,0632	0,6988	0,1345	0,6869	0,0863	0,6939	0,1549	0,6686
	1	20	0,2142	0,2379	0,2963	0,5292	0,2410	0,7001	0,2979	0,6971
		100	0.3001	0,0000	0,3001	0,0273	0,3001	0,6999	0,3001	0,6999
		10	0,0815	0,6992	0,2109	0,6921	0,1052	0,6994	0,2237	0,6795
0,7	5	20	0,2533	0,0000	0,3002	0,4378	0,2718	0,6998	0,3002	0,6993
		100	0,3000	0,0000	0,3000	0,0024	0,3000	0,7000	0,3000	0,7000
		10	0,0508	0,7938	0,1021	0,7736	0,0666	0,7808	0,1145	0,7451
	1	20	0,1668	0,4816	0,1979	0,6827	0,1782	0,8010	0,1982	0,7852
		100	0,2001	0,0004	0,2001	0,2227	0,2001	0,7999	0,2001	0,7994
		10	0,0702	0,7992	0,1485	0,7787	0,0857	0,7993	0,1558	0,7544
0,8	5	20	0,1906	0,0019	0,2004	0,6460	0,1953	0,7996	0,2004	0,7918
		100	0,2004	0,0000	0,2004	0,1013	0,2004	0,7996	0,2004	0,7995
		10	0,0283	0,8770	0,0564	0,8534	0,0362	0,8484	0,0622	0,8138
	1	20	0,0906	0,7563	0,0996	0,8051	0,0940	0,8875	0,0997	0,8440
		100	0,0998	0,2087	0,0998	0,6418	0,0998	0,8975	0,0998	0,8865
		10	0,0453	0,8987	0,0768	0,8574	0,0525	0,8931	0,0800	0,8187
0,9	5	20	0,1000	0,2003	0,1008	0,7988	0,1005	0,8992	0,1008	0,8542
		100	0,1008	0,0000	0,1007	0,5642	0,1007	0,8990	0,1007	0,8915

Fonte: Da autora (2024).

Tabela 8 – Simulação Monte Carlo com tamanho amostral $n = 300$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0,1$ e $\rho_2 = 0,5$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto comedian (JMC), e Rao com o estimador robusto comedian (RC), considerando diferentes proporções de não outliers δ , variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,0070	0,6646	0,0317	0,6668	0,0031	0,6502	0,0376	0,6533
		10	0,0220	0,6545	0,0805	0,6679	0,0010	0,6379	0,0802	0,6591
	1	20	0,0486	0,6536	0,1242	0,6628	0,0002	0,6283	0,1088	0,6699
		100	0,2217	0,6268	0,1840	0,4942	0,0000	0,6095	0,0128	0,6832
0,7		5	0,0368	0,6843	0,1116	0,6818	0,0511	0,6912	0,1323	0,6859
		10	0,0692	0,6515	0,2122	0,6525	0,0943	0,6951	0,2321	0,6925
	5	20	0,1225	0,6020	0,2727	0,6010	0,1538	0,6986	0,2828	0,6975
		100	0,2838	0,2562	0,2997	0,1964	0,2808	0,7002	0,2998	0,7002
		5	0,0049	0,7638	0,0747	0,7682	0,0017	0,7496	0,0816	0,7584
	1	10	0,0164	0,7574	0,1420	0,7626	0,0005	0,7449	0,1452	0,768
		20	0,0403	0,7540	0,1719	0,7491	0,0001	0,7344	0,1711	0,7739
		100	0,1713	0,7544	0,1965	0,6184	0,0000	0,7389	0,1938	0,7892
0,8		5	0,0289	0,7786	0,1209	0,7750	0,0398	0,7824	0,1313	0,7752
	5	10	0,0585	0,7573	0,1824	0,7537	0,0764	0,78930	0,1877	0,7849
		20	0,1046	0,7330	0,1968	0,7259	0,1238	0,7957	0,1981	0,7927
		100	0,1980	0,5488	0,2004	0,4749	0,1952	0,7991	0,2004	0,7989
		5	0,0024	0,8629	0,0691	0,8630	0,0006	0,8500	0,0721	0,8528
	1	10	0,0101	0,8563	0,0945	0,8562	0,0001	0,8483	0,0955	0,8631
		20	0,0249	0,8579	0,0985	0,8519	0,0000	0,8471	0,0986	0,8692
		100	0,0937	0,8723	0,1000	0,7882	0,0000	0,8638	0,1000	0,8842
0,9		5	0,0167	0,8722	0,0809	0,8649	0,0223	0,8692	0,0838	0,8573
		10	0,0348	0,8608	0,0981	0,8541	0,0440	0,8775	0,0988	0,8673
	5	20	0,0622	0,8547	0,1001	0,8464	0,0702	0,8842	0,1001	0,8755
		100	0,0992	0,8257	0,0996	0,7628	0,0968	0,8958	0,0996	0,8912

Fonte: Da autora (2024).

Tabela 9 – Simulação Monte Carlo com tamanho amostral $n = 300$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0,1$ e $\rho_2 = 0,5$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto comedian (JMC), e Rao com o estimador robusto comedian (RC), considerando diferentes proporções de não outliers δ , variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,020	0,6252	0,0502	0,6293	0,0090	0,6071	0,0569	0,6139
		10	0,0419	0,6147	0,1028	0,6351	0,0034	0,5888	0,1060	0,6242
	1	20	0,0798	0,6116	0,1491	0,6264	0,0008	0,5706	0,13805	0,6408
		100	0,2553	0,5687	0,2084	0,4017	0,0000	0,5226	0,1574	0,6662
0,7		5	0,0589	0,6522	0,1363	0,6485	0,0772	0,6749	0,1588	0,6656
		10	0,0975	0,6074	0,2321	0,6102	0,1288	0,6879	0,2515	0,6822
	5	20	0,1556	0,5357	0,2818	0,5351	0,1913	0,6959	0,2896	0,6937
		100	0,2917	0,1611	0,2997	0,1229	0,2922	0,6999	0,2998	0,7001
		5	0,0141	0,7197	0,0869	0,7264	0,0051	0,7035	0,0952	0,7206
	1	10	0,0313	0,7141	0,1513	0,7225	0,0015	0,6937	0,1560	0,7365
		20	0,0640	0,7105	0,1795	0,7039	0,0003	0,6776	0,1800	0,7476
		100	0,1860	0,7080	0,1980	0,5259	0,0000	0,6668	0,1965	0,7767
0,8		5	0,0445	0,7418	0,1329	0,7360	0,0581	0,7574	0,1444	0,7466
	5	10	0,0785	0,7149	0,1877	0,7101	0,1001	0,7745	0,1922	0,7662
		20	0,1257	0,6767	0,1980	0,6684	0,1458	0,7869	0,1989	0,7814
		100	0,1994	0,4319	0,2004	0,3663	0,1986	0,7970	0,2004	0,7975
		5	0,0071	0,8155	0,0730	0,8155	0,0019	0,8018	0,0762	0,8093
	1	10	0,0183	0,8095	0,0958	0,8099	0,0005	0,7973	0,0967	0,8251
		20	0,0379	0,8128	0,0987	0,8043	0,0000	0,7939	0,0988	0,8371
		100	0,0973	0,8355	0,1000	0,7130	0,0000	0,8091	0,1000	0,8653
0,9		5	0,0246	0,8296	0,0843	0,8171	0,0317	0,8332	0,0873	0,8168
	5	10	0,0453	0,8171	0,0987	0,8072	0,0559	0,8483	0,0993	0,8332
		20	0,0715	0,8086	0,1001	0,7963	0,0798	0,8607	0,1002	0,8483
		100	0,0995	0,7609	0,0996	0,6787	0,0985	0,8843	0,0996	0,8791

Fonte: Da autora (2024).

Tabela 10 – Simulação Monte Carlo com tamanho amostral $n = 50$, número de componentes principais retidos $k = 1$, média da população contaminante $\mu = 10$, correlação $\rho_1 = 0,5$ e $\rho_2 = 0,9$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção de *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		2	0,0001	0,6390	0,1442	0,6695	0,0008	0,5957	0,1604	0,6489
		5	0,1089	0,6065	0,2413	0,6120	0,0000	0,5648	0,2441	0,6500
	1	10	0,2280	0,6072	0,2749	0,5180	0,0000	0,5420	0,2731	0,6505
		20	0,2934	0,5990	0,2869	0,3394	0,0000	0,5374	0,2820	0,6496
		100	0,3006	0,5885	0,2970	0,0312	0,3006	0,0031	0,2934	0,6586
0,7		2	0,0136	0,6662	0,1619	0,6826	0,0277	0,6323	0,1806	0,6653
		5	0,0251	0,6426	0,2458	0,6101	0,0245	0,6385	0,2558	0,6758
	5	10	0,0205	0,6517	0,2686	0,5056	0,0175	0,6478	0,2757	0,6791
		20	0,0148	0,6667	0,2809	0,3184	0,0078	0,6724	0,2842	0,6868
		100	0,0021	0,6959	0,2930	0,0348	0,3017	0,0033	0,2954	0,6805
		2	0,0000	0,7454	0,1299	0,7652	0,0002	0,6994	0,1386	0,7374
		5	0,0842	0,7161	0,1886	0,7132	0,0000	0,6831	0,1903	0,7442
	1	10	0,1690	0,7211	0,1988	0,6550	0,0000	0,6741	0,1989	0,7457
		20	0,1957	0,7334	0,1969	0,5493	0,0000	0,6914	0,1968	0,7516
		100	0,2003	0,7615	0,2002	0,1559	0,2003	0,0035	0,2002	0,7250
0,8		2	0,0090	0,7590	0,1331	0,7673	0,0175	0,7206	0,1422	0,7393
		5	0,0166	0,7330	0,1888	0,7119	0,0137	0,7256	0,1916	0,7506
	5	10	0,0138	0,7465	0,1936	0,6597	0,0095	0,7420	0,1946	0,7599
		20	0,0106	0,7610	0,1981	0,5341	0,0036	0,7662	0,1986	0,7651
		100	0,0025	0,7999	0,1969	0,1691	0,1972	0,0037	0,1971	0,7266
		2	0,0000	0,8474	0,0706	0,8422	0,0001	0,8000	0,0741	0,8020
		5	0,0527	0,8152	0,0999	0,8050	0,0000	0,7942	0,1002	0,8034
	1	10	0,0919	0,8300	0,0988	0,7956	0,0000	0,8039	0,0989	0,8082
		20	0,0997	0,8492	0,0997	0,7464	0,0000	0,8316	0,0997	0,8072
		100	0,0987	0,8925	0,0987	0,5021	0,0987	0,0051	0,0987	0,7116
0,9		2	0,0034	0,8516	0,0749	0,8426	0,0076	0,80822	0.07834	0.80446
		5	0,0088	0,8237	0,0980	0,8055	0,0046	0,8164	0,0990	0,8060
	5	10	0,0092	0,8377	0,0991	0,7936	0,0027	0,8305	0,0992	0,8081
		20	0,0101	0,8550	0,0996	0,7470	0,0009	0,8607	0,0996	0,8077
		100	0,0105	0,8970	0,0998	0,5130	0,0999	0,0048	0,0999	0,7087

Fonte: Da autora (2024).

Tabela 11 – Simulação Monte Carlo com tamanho amostral $n = 100$, número de componentes retidos $k = 1$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		2	0,0160	0,6634	0,0278	0,6533	0,0306	0,6299	0,0447	0,6198
	1	5	0,0266	0,6375	0,0373	0,6292	0,0292	0,6347	0,0518	0,6145
		10	0,0254	0,6404	0,0339	0,6327	0,0258	0,6377	0,0568	0,6082
		20	0,0210	0,6512	0,0362	0,6387	0,0203	0,6481	0,0686	0,6046
0,7		100	0,0064	0,6877	0,1047	0,6007	0,1450	0,3594	0,1664	0,6072
		2	0,0552	0,6944	0,1008	0,6775	0,0785	0,6840	0,1225	0,6558
	5	5	0,1067	0,5927	0,1903	0,6223	0,1391	0,6957	0,2172	0,6730
		10	0,1577	0,4739	0,2530	0,5814	0,1991	0,6967	0,2737	0,6817
		20	0,2214	0,2935	0,2876	0,4953	0,2544	0,7012	0,2940	0,6933
		100	0,2980	0,0115	0,2994	0,0799	0,1481	0,3548	0,2994	0,7000
		2	0,0125	0,7589	0,0245	0,7442	0,0226	0,7219	0,0373	0,7072
	1	5	0,0193	0,7292	0,0340	0,7174	0,0194	0,7256	0,0488	0,7034
		10	0,0167	0,7300	0,0368	0,7221	0,0165	0,7264	0,0594	0,6965
		20	0,0148	0,7422	0,0511	0,7273	0,0126	0,7399	0,0796	0,6960
0,8		100	0,0044	0,7845	0,1300	0,6683	0,1020	0,3923	0,1637	0,7038
		2	0,0445	0,7881	0,0768	0,7657	0,0606	0,7699	0,0914	0,7347
	5	5	0,0895	0,7084	0,1418	0,7199	0,1113	0,7865	0,1571	0,7493
		10	0,1322	0,6368	0,1818	0,7032	0,1549	0,7926	0,1902	0,7601
		20	0,1734	0,5185	0,1968	0,6568	0,1850	0,7965	0,1982	0,7728
		100	0,1981	0,1543	0,1982	0,3049	0,1002	0,3960	0,1982	0,7937
		2	0,0080	0,85612	0,0164	0,8375	0,0139	0,8143	0,0232	0,7933
	1	5	0,0105	0,8200	0,0252	0,8032	0,0119	0,8182	0,0362	0,7859
		10	0,0093	0,8250	0,0336	0,8134	0,0099	0,8224	0,0482	0,7833
		20	0,0072	0,8377	0,0485	0,8196	0,0054	0,8351	0,0641	0,7823
0,9		100	0,0038	0,8819	0,0923	0,7805	0,0509	0,4489	0,0982	0,7712
		2	0,0265	0,8743	0,0406	0,8463	0,0343	0,8434	0,0477	0,8062
	5	5	0,0555	0,8209	0,0777	0,8075	0,0650	0,8625	0,0839	0,8088
		10	0,0769	0,8056	0,0933	0,8130	0,0830	0,8725	0,0956	0,8192
		20	0,0935	0,7713	0,0992	0,8003	0,0946	0,8784	0,0995	0,8265
		100	0,0994	0,6584	0,0994	0,6487	0,0498	0,4515	0,0994	0,8479

Fonte: Da autora (2024).

Tabela 12 – Simulação Monte Carlo com tamanho amostral $n = 300$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
$0,7$	1	5	0,0112	0,6733	0,0178	0,6635	0,0147	0,6658	0,0246	0,6496
		10	0,0139	0,6669	0,0189	0,6588	0,0137	0,6659	0,0255	0,6480
		20	0,0130	0,6698	0,0174	0,6635	0,0127	0,6697	0,0274	0,6507
		100	0,0070	0,6846	0,0189	0,6681	0,0051	0,6870	0,0565	0,6553
	5	5	0,0545	0,6804	0,1189	0,6791	0,0735	0,6968	0,1394	0,6840
		10	0,1141	0,5990	0,2179	0,6486	0,1440	0,7002	0,2386	0,6948
		20	0,1965	0,4131	0,2822	0,5920	0,2262	0,7002	0,2900	0,6985
		100	0,2984	0,0034	0,2990	0,1639	0,2987	0,7010	0,2990	0,7010
$0,8$	1	5	0,0076	0,7682	0,0139	0,7575	0,0101	0,7595	0,0198	0,7430
		10	0,0089	0,764	0,0153	0,7556	0,0092	0,7633	0,0221	0,7461
		20	0,0087	0,7646	0,0181	0,7574	0,0084	0,7639	0,0297	0,7463
		100	0,0047	0,7815	0,0755	0,7562	0,0030	0,7840	0,0979	0,7568
	5	5	0,0469	0,7814	0,0923	0,7740	0,0604	0,7934	0,1059	0,7730
		10	0,1000	0,7292	0,1630	0,7555	0,1187	0,7988	0,1730	0,7849
		20	0,1613	0,6222	0,1963	0,7262	0,1745	0,7989	0,1982	0,7911
		100	0,2004	0,0989	0,2004	0,4575	0,2004	0,7995	0,2004	0,7990
$0,9$	1	5	0,0038	0,8659	0,0110	0,8545	0,0053	0,8559	0,0155	0,8384
		10	0,0046	0,8588	0,0157	0,8488	0,0048	0,8575	0,0224	0,8395
		20	0,0045	0,8616	0,02705	0,8537	0,0041	0,8606	0,0365	0,8429
		100	0,0026	0,8803	0,0863	0,8523	0,0010	0,8827	0,0886	0,8532
	5	5	0,0291	0,8760	0,0534	0,8617	0,0363	0,8798	0,0598	0,8525
		10	0,0619	0,8535	0,0886	0,8528	0,0700	0,8883	0,0921	0,8620
		20	0,0894	0,8262	0,0984	0,8497	0,0924	0,8946	0,0987	0,8727
		100	0,1001	0,6233	0,1001	0,7713	0,1000	0,8977	0,1001	0,8894

Fonte: Da autora (2024).

Tabela 13 – Simulação Monte Carlo com tamanho amostral $n = 300$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
$0,7$	1	5	0,0262	0,6379	0,03597	0,6224	0,0294	0,6309	0,0430	0,6099
		10	0,0287	0,6329	0,0359	0,6204	0,0289	0,6311	0,0450	0,6062
		20	0,0272	0,6374	0,0337	0,6271	0,0271	0,6357	0,0482	0,6103
		100	0,0172	0,6608	0,0351	0,6344	0,0151	0,6612	0,0868	0,6175
	5	5	0,0796	0,6445	0,1433	0,6442	0,1017	0,6908	0,1649	0,6645
		10	0,1442	0,5227	0,2371	0,6025	0,1780	0,6994	0,2558	0,6853
		20	0,2236	0,3027	0,2888	0,5227	0,2512	0,7001	0,2941	0,6953
		100	0,2988	0,0008	0,2990	0,0974	0,2989	0,7010	0,2990	0,7010
$0,8$	1	5	0,0179	0,7273	0,0267	0,7117	0,0202	0,7195	0,0331	0,6978
		10	0,0188	0,7247	0,0275	0,7118	0,0191	0,7239	0,0364	0,7005
		20	0,0180	0,7263	0,0314	0,7165	0,0177	0,7245	0,0470	0,7010
		100	0,0115	0,7541	0,0935	0,7145	0,0092	0,7543	0,1207	0,7193
	5	5	0,0638	0,7448	0,1080	0,7331	0,0797	0,7794	0,1220	0,7428
		10	0,1190	0,6667	0,1722	0,7116	0,1389	0,7942	0,1813	0,7640
		20	0,1736	0,5217	0.1979	0,6705	0,1848	0,7977	0,1994	0,7793
		100	0,2004	0,0466	0,2004	0,3472	0,2004	0,7992	0,2004	0,7979
$0,9$	1	5	0,0091	0,8197	0,0184	0,8026	0,0107	0,8115	0,0234	0,7883
		10	0,0097	0,8149	0,0237	0,8004	0,0098	0,8134	0,0325	0,7895
		20	0,0092	0,8195	0,0367	0,8072	0,0087	0,8165	0,0484	0,7950
		100	0,0061	0,8496	0,0907	0,8057	0,0033	0,8489	0,0928	0,8118
	5	5	0,0379	0,8346	0,0607	0,8129	0,0460	0,8520	0,0598	0,8525
		10	0,0704	0,8058	0,0919	0,8059	0,0784	0,8706	0,0949	0,8241
		20	0,0927	0,7642	0,0987	0,8018	0,0953	0,8828	0,0989	0,8419
		100	0,1001	0,4949	0,1001	0,6903	0,1001	0,8917	0,10013	0,8758

Fonte: Da autora (2024).

Tabela 14 – Simulação Monte Carlo com tamanho amostral $n = 700$, número de componentes retidos $k = 9$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0$, e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
$0,7$	1	10	0,0159	0,6632	0,0236	0,6475	0,0300	0,6305	0,0401	0,6101
		20	0,0291	0,6322	0,0379	0,6173	0,0294	0,6318	0,0495	0,5976
		100	0,0239	0,6442	0,0346	0,6308	0,0244	0,6408	0,0651	0,6103
	5	10	0,0528	0,6935	0,0980	0,6795	0,0753	0,6812	0,1206	0,6569
		20	0,1789	0,4500	0,2707	0,5753	0,2104	0,6985	0,2816	0,6894
		100	0,2995	0,0009	0,2997	0,1232	0,2996	0,7003	0.2997	0,7003
$0,8$	1	10	0,0105	0,7581	0,0164	0,7423	0,0199	0,7205	0,0276	0,7001
		20	0,0194	0,7226	0,0281	0,7066	0,0196	0,7218	0,0380	0,6896
		100	0,0159	0,7362	0,0410	0,7188	0,0159	0,7324	0,0718	0,7089
	5	10	0,0407	0,7858	0,0736	0,7674	0,0560	0,7647	0,0884	0,7357
		20	0,1425	0,6270	0,1888	0,6987	0,1596	0,7963	0,1930	0,7709
		100	0,1996	0,0515	0,1996	0,3979	0,1996	0,8003	0,1996	0,7990
$0,9$	1	10	0,0052	0,8522	0,0100	0,8352	0,0098	0,8101	0,0160	0,7887
		20	0,0097	0,8128	0,0210	0,7951	0,0097	0,8123	0,0292	0,7817
		100	0,0078	0,8282	0,0610	0,8070	0,0071	0,8240	0,0710	0,8069
	5	10	0,0232	0,8702	0,0415	0,8495	0,0309	0,8370	0,0487	0,8084
		20	0,0803	0,7949	0,0976	0,8026	0,0866	0,8744	0,0987	0,8315
		100	0,1005	0,5012	0,1005	0,7132	0,1005	0,8958	0,1005	0,8793

Fonte: Da autora (2024).

Tabela 15 – Simulação Monte Carlo com tamanho amostral $n = 250$, número de componentes retidos $k = 1$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0,1$ e $\rho_2 = 0,9$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
0,7	1	2	0,0001	0,6340	0,0019	0,5829	0,0001	0,5917	0,0038	0,5333
		5	0,1613	0,5906	0,0933	0,5024	0,0000	0,5504	0,0012	0,4262
		10	0,2809	0,5677	0,2006	0,4204	0,0000	0,5047	0,0007	0,3119
		20	0,2996	0,5279	0,2639	0,2970	0,0000	0,4357	0,0006	0,1747
		100	0,3001	0,3221	0,2926	0,0141	0,0000	0,1339	0,0002	0,0025
	5	2	0,0058	0,6509	0,0230	0,6503	0,0139	0,6134	0,0366	0,6143
		5	0,0304	0,6183	0,0441	0,6220	0,0061	0,5983	0,0389	0,6008
		10	0,0477	0,6172	0,0555	0,6244	0,0027	0,5871	0,0366	0,5892
		20	0,0790	0,6146	0,0757	0,6237	0,0008	0,5703	0,0330	0,5681
		100	0,2296	0,5905	0,1887	0,6105	0,0000	0,5465	0,0410	0,4688
0,8	1	2	0,0000	0,7399	0,0033	0,7095	0,0001	0,6952	0,0060	0,6586
		5	0,1139	0,6986	0,0692	0,6493	0,0000	0,6673	0,0017	0,5835
		10	0,1898	0,6898	0,1411	0,6149	0,0000	0,6418	0,0009	0,5095
		20	0,2002	0,6760	0,1810	0,5427	0,0000	0,6021	0,0003	0,3949
		100	0,2002	0,5882	0,1983	0,2008	0,0000	0,4117	0,0001	0,0689
	5	2	0,0034	0,7509	0,0257	0,7508	0,0085	0,7091	0,0355	0,7114
		5	0,0208	0,7144	0,0410	0,7154	0,0039	0,6997	0,0399	0,7008
		10	0,0350	0,7143	0,0460	0,7181	0,0015	0,6920	0,0368	0,6923
		20	0,0603	0,7158	0,0577	0,7206	0,0004	0,6821	0,0323	0,6791
		100	0,1648	0,7243	0,1296	0,7298	0,0000	0,6899	0,0349	0,6156
0,9	1	2	0,0000	0,8434	0,0051	0,8282	0,0000	0,7975	0,0077	0,7798
		5	0,0544	0,8044	0,0311	0,7877	0,0000	0,7861	0,0041	0,7478
		10	0,0943	0,8038	0,0629	0,7838	0,0000	0,7777	0,0016	0,7199
		20	0,0998	0,8048	0,0859	0,7721	0,0000	0,7638	0,0005	0,6686
		100	0,0998	0,8125	0,0986	0,7026	0,0000	0,7358	0,0001	0,4466
	5	2	0,0018	0,8488	0,0216	0,8449	0,0043	0,8053	0,0276	0,8019
		5	0,0110	0,8084	0,0281	0,8058	0,0015	0,7996	0,0289	0,7942
		10	0,0186	0,8119	0,0286	0,8116	0,0007	0,7990	0,0258	0,7931
		20	0,0335	0,8171	0,0327	0,8142	0,0002	0,7970	0,0236	0,7845
		100	0,0860	0,8451	0,0650	0,8281	0,0000	0,8281	0,0253	0,7515

Fonte: Da autora (2024).

Tabela 16 – Simulação Monte Carlo com tamanho amostral $n = 900$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0,5$ e $\alpha = 0,05$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,0071	0,6630	0,0080	0,64972	0,0023	0,6473	0,0060	0,6278
	1	10	0,0269	0,6522	0,02123	0,6374	0,0006	0,6330	0,0037	0,6072
		20	0,0673	0,6406	0,04558	0,6228	0,0001	0,6109	0,0033	0,5762
		100	0,2802	0,5691	0,2252	0,5159	0,0000	0,5067	0,0012	0,3923
0,7		5	0,0361	0,6833	0,0664	0,6802	0,0506	0,6900	0,0851	0,6820
	5	10	0,0670	0,6544	0,1272	0,6608	0,0908	0,6955	0,1545	0,6898
		20	0,1187	0,6100	0,2014	0,6359	0,1498	0,6990	0,2269	0,6958
		100	0,2850	0,2561	0,2990	0,3898	0,2909	0,7000	0,2997	0,6999
		5	0,0048	0,7613	0,0059	0,7511	0,0013	0,7469	0,0045	0,7308
	1	10	0,0203	0,7533	0,0160	0,7429	0,0002	0,7388	0,0043	0,7194
		20	0,0523	0,7486	0,0334	0,7386	0,0000	0,7263	0,0053	0,7037
		100	0,1931	0,7218	0,1314	0,7157	0,0000	0,6735	0,0193	0,6197
0,8		5	0,0448	0,7674	0,0889	0,7668	0,0594	0,7880	0,1040	0,7804
	5	10	0,0572	0,7591	0,1122	0,7598	0,0748	0,7907	0,1279	0,7816
		20	0,1024	0,7334	0,1613	0,7457	0,1222	0,7955	0,1725	0,7892
		100	0,1979	0,5265	0,1999	0,5919	0,1987	0,7997	0,2000	0,7989
		5	0,0024	0,8621	0,0047	0,8547	0,0005	0,8491	0,0050	0,8379
	1	10	0,0110	0,8533	0,0114	0,8471	0,0001	0,8447	0,0083	0,8339
		20	0,03044	0,8518	0,0242	0,8483	0,0000	0,8399	0,0159	0,8315
		100	0,0975	0,8524	0,06861	0,8565	0,0000	0,8256	0,0367	0,8175
0,9		5	0,0281	0,8642	0,0575	0,8614	0,0360	0,8759	0,0641	0,8678
	5	10	0,0359	0,8596	0,06997	0,8557	0,0450	0,8775	0,0761	0,8658
		20	0,0638	0,8518	0,0903	0,8524	0,0721	0,8842	0,0933	0,8742
		100	0,1000	0,7967	0,1002	0,8016	0,1001	0,8951	0,1002	0,8903

Fonte: Da autora (2024).

Tabela 17 – Simulação Monte Carlo com tamanho amostral $n = 900$, número de componentes retidos $k = 3$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0$ e $\rho_2 = 0,5$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
		5	0,0210	0,6226	0,0221	0,6052	0,0071	0,6039	0,0144	0,5787
	1	10	0,0505	0,6103	0,0403	0,5911	0,0020	0,5817	0,0084	0,5482
		20	0,1051	0,5953	0,0747	0,5725	0,0003	0,5496	60,0004	0,5073
		100	0,2912	0,4992	0,2475	0,4404	0,0000	0,4113	0,0018	0,3003
0,7		5	0,0583	0,6519	0,0912	0,6463	0,0765	0,6731	0,1132	0,6583
	5	10	0,0944	0,6111	0,1551	0,6221	0,1250	0,6875	0,1856	0,6749
		20	0,1508	0,5464	0,2250	0,5855	0,1860	0,6960	0,2491	0,6882
		100	0,2920	0,1669	0,2995	0,2940	0,2960	0,6999	0,2999	0,6995
		5	0,0145	0,7173	0,0156	0,7029	0,0041	0,7006	0,0099	0,6793
	1	10	0,0375	0,7088	0,0296	0,6947	0,0009	0,6867	0,0081	0,6612
		20	0,0794	0,7026	0,0531	0,6894	0,0001	0,6676	0,0085	0,6382
		100	0,1971	0,6634	0,1446	0,6573	0,0000	0,5868	0,0223	0,5293
0,8		5	0,0625	0,7282	0,1056	0,7273	0,0806	0,7693	0,1224	0,7554
	5	10	0,0766	0,7166	0,1282	0,7181	0,0979	0,7748	0,1449	0,7581
		20	0,1230	0,6784	0,1718	0,6979	0,1439	0,7867	0,1816	0,7730
		100	0,1990	0,4169	0,1999	0,4957	0,1995	0,7988	0,2000	0,7964
		5	0,0075	0,8137	0,0837	0,8227	0,0015	0,7994	0,0854	0,8239
	1	10	0,0202	0,8060	0,0185	0,7974	0,0004	0,7934	0,0121	0,7789
		20	0,0447	0,8054	0,0331	0,8002	0,0000	0,7845	0,0197	0,7750
		100	0,0989	0,8052	0,0740	0,8122	0,0000	0,7578	0,0393	0,7551
0,9		5	0,0370	0,8198	0,1001	0,8170	0,0462	0,8429	0,1002	0,8488
	5	10	0,0459	0,8160	0,0762	0,8099	0,0565	0,8485	0,0823	0,8292
		20	0,0726	0,8043	0,0932	0,8051	0,0808	0,8617	0,0958	0,8442
		100	0,1002	0,7228	0,1002	0,7321	0,1002	0,8863	0,1002	0,8770

Fonte: Da autora (2024).

Tabela 18 – Simulação Monte Carlo com tamanho amostral $n = 300$, número de componentes retidos $k = 9$, média da população contaminante $\mu = 1$, correlação $\rho_1 = 0,9$ e $\rho_2 = 0,1$ e $\alpha = 0,10$ para avaliar as taxas de verdadeiro positivo (VP) e verdadeiro negativo (VN) na detecção dos *outliers* e não *outliers* pelos testes de Jackson e Mudholkar (JMO), Rao (RO) e os testes propostos: Jackson e Mudholkar com o estimador robusto *comedian* (JMC), e Rao com o estimador robusto *comedian* (RC), considerando diferentes proporções de não *outliers* δ, variância σ^2 e número de variáveis p.

δ	σ^2	p	JMO		JMC		RO		RC	
			VP	VN	VP	VN	VP	VN	VP	VN
$0,7$	1	10	0,0610	0,6972	0,1279	0,6808	0,0842	0,6903	0,1488	0,6593
		20	0,2067	0,2834	0,2919	0,5508	0,2367	0,7008	0,2951	0,6905
		100	0,3009	0,0000	0,3010	0,0789	0,3010	0,6990	0,3010	0,6965
	5	10	0,0799	0,6996	0,2078	0,6899	0,1036	0,6997	0,2213	0,6755
		20	0,2494	0,0002	0,2991	0,4531	0,2705	0,7009	0,2991	0,6995
		100	0,2992	0,0000	0,2992	0,0121	0,2992	0,7008	0,2992	0,7008
$0,8$	1	10	0,0477	0,7901	0,0952	0,7661	0,0632	0,7747	0,1084	0,7344
		20	0,1628	0,5232	0,1977	0,6891	0,1759	0,7993	0,1985	0,7717
		100	0,2011	0,0029	0,2011	0,2763	0,2011	0,7987	0,2011	0,7894
	5	10	0,0690	0,7991	0,1466	0,7754	0,0849	0,7988	0,1542	0,7498
		20	0,1875	0,0119	0,1989	0,6535	0,1935	0,8011	0,1989	0,7884
		100	0,2009	0,0000	0,2009	0,1243	0,2009	0,7991	0,2009	0,7974
$0,9$	1	10	0,0249	0,8735	0,0517	0,8469	0,0329	0,8418	0,0581	0,8044
		20	0,0861	0,7794	0,0998	0,8013	0,0908	0,8795	0,0999	0,8283
		100	0,0989	0,3827	0,0989	0,6328	0,0967	0,8894	0,0989	0,8608
	5	10	0,0425	0,8935	0,0766	0,8515	0,0499	0,8822	0,0802	0,8116
		20	0,0991	0,3537	0,1003	0,7987	0,0999	0,8986	0,1003	0,8457
		100	0,0998	0,0015	0,0998	0,5528	0,0990	0,8940	0,0998	0,8803

Fonte: Da autora (2024).

APÊNDICE B – Comandos usados no R

```
#Quadratic Euclidian Distance
QuadEucDist <- function(x)
{
  return(sum(x^2))
}

# Outlier detection based on Principal
# components, X (n x p) is the data matrix
# k is the retained PC
# Jackson and Muldholkar
JacksonMuldholkar <- function(Xs, k = 2)
{
  n <- nrow(Xs)
  p <- ncol(Xs)
  Xb <- apply(Xs,2,mean)
  X <- Xs-matrix(rep(Xb, each=n),n,p)
  UL <- eigen(var(X))
  Uk <- UL$vectors[,1:k]
  Uk <- matrix(Uk, p, k)
  Lk <- UL$values[1:k]
  Yk <- X%*%Uk%*%diag(1/sqrt(Lk),nrow=k)
  Xp <- Yk%*%diag(sqrt(Lk),nrow=k)%*%t(Uk)
  D2 <- apply(X-Xp,1,QuadEucDist)
  Lpk <- UL$values[(k+1):p]
  T1 <- sum(Lpk)
  T2 <- sum(Lpk^2)
  T3 <- sum(Lpk^3)
  h0 <- 1 - 2*T1*T3 / (3 * T2^2)
  mu <- 1 + T2*h0*(h0-1) / T1^2
  sig2 <- 2*T2*h0^2 / T1^2
  Z <- ((D2/T1)^h0-mu) / sqrt(sig2)
```

```
  val.p <- 2*(1-pnorm(abs(Z)))
  return(list(D2=D2,Z=Z,val.p=val.p,
              mu=mu,sig2=sig2))
}

# Outlier detection based on Principal
# components, X (n x p) is the data matrix
# k is the retained PC
# Rao 1964
Rao1964 <- function(Xs, k = 2)
{
  n <- nrow(Xs)
  p <- ncol(Xs)
  Xb <- apply(Xs,2,mean)
  X <- Xs-matrix(rep(Xb, each=n),n,p)
  S  <- var(X)
  UL <- eigen(S)
  Uk <- UL$vectors[,(k+1):p]
  Uk <- matrix(Uk, p, p-k)
  Lk <- UL$values[(k+1):p]
  Lk[Lk<=0] <- 0
  Lk <- 1 / sqrt(Lk)
  Yk <- X%*%Uk%*%diag(Lk,nrow=p-k)
  D2 <- apply(Yk,1,QuadEucDist)
  val.p <- 1-pchisq(D2,p-k)
  return(list(D2=D2,val.p=val.p))
}

# Outlier detection based on Principal
# components, X (n x p) is the data matrix
# k is the retained PC
# JacksonMuldholkar - Comedian
```

```
library(robustbase)
JacksonMuldholkarCom <- function(Xs, k = 2)
{
  n <- nrow(Xs)
  p <- ncol(Xs)
  comed <- covComed(Xs)
  Xb <- comed$raw.center
  X <- Xs-matrix(rep(Xb, each=n),n,p)
  UL <- eigen(comed$raw.cov)
  Uk <- UL$vectors[,1:k]
  Uk <- matrix(Uk, p, k)
  Lk <- UL$values[1:k]
  Yk <- X%*%Uk%*%diag(1/sqrt(Lk),nrow=k)
  Xp <- Yk%*%diag(sqrt(Lk),nrow=k)%*%t(Uk)
  D2 <- apply(X-Xp,1,QuadEucDist)
  Lpk <- UL$values[(k+1):p]
  T1 <- sum(Lpk)
  T2 <- sum(Lpk^2)
  T3 <- sum(Lpk^3)
  h0 <- 1 - 2*T1*T3 / (3 * T2^2)
  mu <- 1 + T2*h0*(h0-1) / T1^2
  sig2 <- 2*T2*h0^2 / T1^2
  Z <- ((D2/T1)^h0-mu) / sqrt(sig2)
  val.p <- 2*(1-pnorm(abs(Z)))
  return(list(D2=D2,Z=Z,val.p=val.p,
              mu=mu,sig2=sig2))
}

# Outlier detection based on Principal
# components, X (n x p) is the data matrix
# k is the retained PC
# Rao 1964 -using Comedian
```

```
Rao1964Comed <- function(Xs, k = 2)
{
  n <- nrow(Xs)
  p <- ncol(Xs)
  comed <- covComed(Xs)
  Xb <- comed$raw.center
  X <- Xs-matrix(rep(Xb, each=n),n,p)
  S  <- comed$raw.cov
  UL <- eigen(S)
  UL <- eigen(S)
  Uk <- UL$vectors[,(k+1):p]
  Uk <- matrix(Uk, p, p-k)
  Lk <- UL$values[(k+1):p]
  Yk <- X%*%Uk%*%diag(1/sqrt(Lk),nrow=p-k)
  D2 <- apply(Yk,1,QuadEucDist)
  val.p <- 1-pchisq(D2,p-k)
  return(list(D2=D2,val.p=val.p))
}

#simular amostras normal multivariada contaminada
library(MASS)
rNCMWL <- function(n, delta , mu1, mu2,  Sig1,  Sig2)
{
  u <- runif(n)
  p <- nrow(Sig1)
  n1 <- length(u[u <= delta])
  n2 <- n - n1
  X <- matrix(0, n, p)
  if (n1 > 0) X[u <= delta, ] <- mvrnorm(n1, mu1, Sig1)
  if (n2 > 0) X[u > delta, ] <- mvrnorm(n2, mu2, Sig2)
  outliers <- u > delta
  return(list(X=X, outliers=outliers))
```

```
}

# Programa para Simular e contar os erros e acertos
# da identificação dos outliers
SimulaOutlierMC<-function(N=1000,n=100,p=3,rho1=0.9,rho2=0.9,
                          m2=10,Sigma2=3,alpha1=0.10,alpha2=0.05,
                          k=2,delta=0.8)
{
  MCRes <- matrix(0,8,4)
  rownames(MCRes) <- c("JMO10", "JMO05","JMC10", "JMC05",
                       "RO10", "RO05","RC10", "RC05")
  colnames(MCRes) <- c("VP","VN","FN","FP")
  for (i in 1:N)
  {
    mu1 <- rep(c(0),times=p)
    Sig1 <- ((1-rho1)*diag(p) + matrix(rho1, p, p))
    mu2 <- runif(p,m2-0.5,m2+0.5)
    Sig2 <- Sigma2*((1-rho2)*diag(p) + matrix(rho2, p, p))
    res <- rNCMWL(n, delta, mu1, mu2, Sig1, Sig2)
    Xs <- res$X
    #res$outliers
    #plot(Xs)
    resJMO <- JacksonMuldholkar(Xs,k)
    resJMC <- JacksonMuldholkarCom(Xs,k)
    resRO <- Rao1964(Xs,k)
    resRC <- Rao1964Comed(Xs,k)
    res1 <- resJMO$val.p[res$outliers]
    MCRes[1,1] <- MCRes[1,1] + length(res1[res1<=alpha1])/n
    MCRes[2,1] <- MCRes[2,1] + length(res1[res1<=alpha2])/n
    MCRes[1,3] <- MCRes[1,3] + length(res1[res1>alpha1])/n
    MCRes[2,3] <- MCRes[2,3] + length(res1[res1>alpha2])/n
    res1 <- resJMO$val.p[!res$outliers]
```

```
MCRes[1,2] <- MCRes[1,2] + length(res1[res1>alpha1])/n
MCRes[2,2] <- MCRes[2,2] + length(res1[res1>alpha2])/n
MCRes[1,4] <- MCRes[1,4] + length(res1[res1<=alpha1])/n
MCRes[2,4] <- MCRes[2,4] + length(res1[res1<=alpha2])/n
res1 <- resJMC$val.p[res$outliers]
MCRes[3,1] <- MCRes[3,1] + length(res1[res1<=alpha1])/n
MCRes[4,1] <- MCRes[4,1] + length(res1[res1<=alpha2])/n
MCRes[3,3] <- MCRes[3,3] + length(res1[res1>alpha1])/n
MCRes[4,3] <- MCRes[4,3] + length(res1[res1>alpha2])/n
res1 <- resJMC$val.p[!res$outliers]
MCRes[3,2] <- MCRes[3,2] + length(res1[res1>alpha1])/n
MCRes[4,2] <- MCRes[4,2] + length(res1[res1>alpha2])/n
MCRes[3,4] <- MCRes[3,4] + length(res1[res1<=alpha1])/n
MCRes[4,4] <- MCRes[4,4] + length(res1[res1<=alpha2])/n
res1 <- resRO$val.p[res$outliers]
MCRes[5,1] <- MCRes[5,1] + length(res1[res1<=alpha1])/n
MCRes[6,1] <- MCRes[6,1] + length(res1[res1<=alpha2])/n
MCRes[5,3] <- MCRes[5,3] + length(res1[res1>alpha1])/n
MCRes[6,3] <- MCRes[6,3] + length(res1[res1>alpha2])/n
res1 <- resRO$val.p[!res$outliers]
MCRes[5,2] <- MCRes[5,2] + length(res1[res1>alpha1])/n
MCRes[6,2] <- MCRes[6,2] + length(res1[res1>alpha2])/n
MCRes[5,4] <- MCRes[5,4] + length(res1[res1<=alpha1])/n
MCRes[6,4] <- MCRes[6,4] + length(res1[res1<=alpha2])/n
res1 <- resRC$val.p[res$outliers]
MCRes[7,1] <- MCRes[7,1] + length(res1[res1<=alpha1])/n
MCRes[8,1] <- MCRes[8,1] + length(res1[res1<=alpha2])/n
MCRes[7,3] <- MCRes[7,3] + length(res1[res1>alpha1])/n
MCRes[8,3] <- MCRes[8,3] + length(res1[res1>alpha2])/n
res1 <- resRC$val.p[!res$outliers]
MCRes[7,2] <- MCRes[7,2] + length(res1[res1>alpha1])/n
MCRes[8,2] <- MCRes[8,2] + length(res1[res1>alpha2])/n
```

```
    MCRes[7,4] <- MCRes[7,4] + length(res1[res1<=alpha1])/n
    MCRes[8,4] <- MCRes[8,4] + length(res1[res1<=alpha2])/n
  }
  MCRes <- MCRes / N
  return(MCRes)
}
# Simulações - Configuração 1
N <- 1000
n <- 100
p <- 5
m2 <- 10
rho1 <- 0.90
rho2 <- 0.10
Sigma2 <- 1
k <- 1
delta <- 0.8
N;n;p;rho1;rho2;k;m2;Sigma2;delta
SimulaOutlierMC(N,n,p,rho1,rho2,m2,Sigma2,
                alpha1=0.10,alpha2=0.05,k,delta)
```