HUMBERTO MARCÍLIO MARTINS

MÉTODOS PARA DETECÇÃO DE *OUTLIERS* MULTIVARIADOS:

VIA USO DOS ESTIMADORES ROBUSTOS

LAVRAS – MG

2022

HUMBERTO MARCÍLIO MARTINS

MÉTODOS PARA DETECÇÃO DE *OUTLIERS* MULTIVARIADOS:
VIA USO DOS ESTIMADORES ROBUSTOS

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

Prof. Dr. Daniel Furtado Ferreira
Orientador

LAVRAS – MG
2022

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

Martins, Humberto Marcílio.
Métodos para detecção de *Outliers* multivariados : Via uso dos estimadores robustos / Humberto Marcílio Martins. – 2022.
90 p.

Orientador: Daniel Furtado Ferreira.

Dissertação (mestrado acadêmico)– Universidade Federal de Lavras, 2022.
Bibliografia.

1. *comedian*. 2. PCOut. 3. OGK. Ferreira, Daniel Furtado. II. Título.

HUMBERTO MARCÍLIO MARTINS

MÉTODOS PARA DETECÇÃO DE *OUTLIERS* MULTIVARIADOS: VIA USO DOS ESTIMADORES ROBUSTOS
METHODS FOR DETECTION OF MULTIVARIATE OUTLIERS: : VIA THE USE OF ROBUST ESTIMATORS

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

APROVADA em 14 de Março de 2022.

Prof.Dr. Daniel Furtado Ferreira	UFLA-MG
Prof.Dr. Denismar Alves Nogueira	UNIFAL-MG
Prof.Dr. Luiz Alberto Beijo	UNIFAL -MG
Prof. Dr. Ben Deivide de Oliveira Batista	UFSJ

Prof. Dr. Daniel Furtado Ferreira
Orientador

LAVRAS – MG
2022

Dedico essa simples dissertação a todos os pesquisadores, professores e alunos. Dedico também a universidade, UFLA-MG e a todos funcionários.

AGRADECIMENTOS

Agradeço aos meus pais, Oswaldo Martins e Maria Aparecida Jeremias Martins, que mesmo sem entender o que de fato eu estudo sempre se sentiram orgulhosos de mim.

Agradeço também a minha esposa e minha filha, que sempre me apoiaram mesmo nos finais de semanas a frente do computador.

Agradeço a todos professores que pude ter aulas pelo aprendizado em especial ao meu orientador que me deu um tema e me ajudou nessa caminhada, além disso, foi paciente e compreensivo, principalmente no momento em que vivemos, foi de fato um grande desafio.

Importante salientar que o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, sem a bolsa nada disse seria possível.

Agradeço também ao Programa de Pós-graduação em Estatística e Experimentação Agropecuária e à UFLA-MG pela oportunidade.

Agradeço em especial a Deus que me manteve com saúde diante da pandemia protegeu a todos que esteve por perto de mim.

Com toda certeza do mundo, nada disso seria possível sem as pessoas e entidades aqui mencionadas.

RESUMO

Na aplicação da análise multivariada é necessário seguir alguns procedimentos para não obter uma relação equivocada do fenômeno de interesse com as demais variáveis, ou seja, o modelo precisa ser bem ajustado às características do fenômeno sob estudo. A detecção de *outliers* é um importante método a ser aplicado nas análises estatísticas, pois um único *outlier* pode causar mudanças nas estimativas dos parâmetros, interferir também nos testes de normalidade e de correlação entre as variáveis, além de alterar os resultados de qualquer outro procedimento de inferência. Portanto, o objetivo desse trabalho é apresentar e comparar alguns métodos de detecção de *outliers* em dados multivariados. Foram comparados os métodos elipsóide de volume mínimo (MVE), Covariância de volume mínimo (MCD), Ortogonalizado de Gnanadesikan e Kettenring (OGK), componentes principais para detecção de *outliers* (PCOut) e o *Comedian*. Para realizar as comparações foi utilizado uma série de simulações prevendo diversas situações utilizando a distribuição normal contaminada. As comparações foram avaliadas atráves da taxa de sucesso (TS), que aponta a porcentagem de *outliers* que os métodos identificaram corretamente e da taxa de falsa detecção (TFD), que aponta a porcentagem de observações que não são *outliers*, mas foram identificadas como *outliers*. Conclui-se que o ideal é utlizar ao menos dois métodos de detecção de *outliers*, visto que apontar o único método como melhor é uma tarefa díficil. No entanto, os métodos PCOut e *Comedian* obtiveram as TS melhores na maioria dos cenários simulados. O método *comedian* obteve as melhores TFD.

Palavras-chave: *comedian*, PCOut, OGK.

ABSTRACT

In the application of the multivariate analysis, it is necessary to follow some procedures in order not to obtain an erroneous relationship between the phenomenon of interest and the other variables, that is, the model needs to be well adjusted to the characteristics of the phenomenon under study. The detection of *outliers* is an important method to be applied in statistical analyses, because a single *outlier* can cause changes in parameter estimates, also interfere with normality and correlation tests between variables, in addition to alter the results of any other inference procedure. Therefore, the objective of this work is to present and compare some methods for detecting *outliers* in multivariate data. The minimum volume ellipsoid (MVE), minimum volume covariance (MCD), orthogonalized Gnanadesikan and Kettenring (OGK) methods, principal components for detection of *outliers* (PCOut) and *Comedian* were compared. To perform the comparisons, a series of simulations was used, predicting different situations using the contaminated normal distribution. Comparisons were evaluated through the success rate (TS), which indicates the percentage of *outliers* that the methods correctly identified, and the false detection rate (TFD), which indicates the percentage of observations that are not *outliers*, but were identified as *outliers*. It is concluded that the ideal is to use at least two methods to detect *outliers*, since pointing out the only method as the best is a difficult task. However, the PCOut and *Comedian* methods obtained the best TS in most of the simulated scenarios. The *comedian* method obtained the best TFD.

Keywords: *comedian*, PCOut, OGK.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Modelos estatísticos têm sido utilizados nas mais diversas áreas do conhecimento, desde as ciências exatas, biológicas até as sociais. A aplicação de modelos estatísticos para explicação de um fenômeno é muito comum em pesquisas dentro e fora do meio acadêmico. Dentre os métodos estatísticos, um dos mais usados é a análise multivariada, que consiste em estudar um fenômeno de interesse por meio de várias variáveis mensuradas na investigação científica em questão. Entre estes estudos encontram-se aqueles casos em que se deseja avaliar a relação de mais de uma variável aleatória dependente com uma ou várias variáveis aleatórias independentes, como na regressão linear multivariada simples e múltipla e na análise de variância multivariada, além de muitos outros casos.

Existem muitas técnicas usuais de análise multivariada, por exemplo, componentes principais, análise de agrupamento, análise discriminantes e regressão multivariada, que utilizam a média, as matrizes de covariância e de correlação (HUBERT; ROUSSEEUW; AELST, 2008). Na verdade, a média e matriz de covariância são estatísticas básicas na maioria dos procedimentos em análise multivariada (MARONNA; YOHAI, 2017). Por isso, na aplicação da análise multivariada é necessário seguir alguns procedimentos para não obter uma relação equivocada da variável de interesse com as demais variáveis. Dentre os procedimentos, envolve aplicar alguns testes estatísticos para verificar se os pressupostos para aplicação destas técnicas foram atendidos, como por exemplo, teste de normalidade multivariada e de homogeneidade de covariâncias. Um dos testes que são primordiais de serem aplicados tanto em uma análise multivariada quanto univariada é o da detecção de *outliers*, embora sejam muitas vezes negligenciado. Vale salientar que em alguns casos como análise de valores extremos as observações de interesse podem ser *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 métodos clássicos existentes são bons para identificar *outliers* em dados com apenas uma variável (HADI, 1992). Na literatura existem vários métodos para detecção de *outliers* em da-

dos multivariados, dentre eles podemos citar o método Elipsóide 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). Os métodos citados são menos sensíveis as observações discrepantes, por isso são considerados métodos robustos. Alguns métodos tentam minimizar as distâncias por meio da matriz de covariância como por exemplo o MVE e MCD. Outros métodos são baseados em projeções da matriz de covariância, como por exemplo, o Kurtosis e OGK. Todos esses métodos são bons quando o conjunto de dados possuem poucas variáveis. Mas à medida que o número de variáveis fica maior, esses métodos ficam mais dispendiosos computacionalmente (MARONNA; YOHAI, 2017). Além disso, a detecção de *outliers* em dados com muitas variáveis envolvidas é mais complexa por que envolve problemas com mascaramento de *outliers*, ou seja, um *outlier* pode mascarar outras observações que também são *outliers* e inundação, isto é, uma observação que não é *outlier* ser identificada como *outlier*.

Embora os métodos MVE e MCD sejam muito utilizados para comparação de procedimentos de detecção de *outliers*, eles apresentam problemas na formação das amostras iniciais, pois elas não podem conter *outliers*. Além disso, os métodos MVE e MCD não possibilitam detecção de *outliers* em dados com alta dimensão. Na literatura os métodos mais utilizados para comparação de detecção de *outliers*, que são o MVE, MCD e OGK. Além desses métodos, foi reportado o uso dos métodos *Comedian* e PCOut para detecção de *outliers* em dados com observações (n) maiores que o número de dimensões (p) e para banco de dados com alta dimensões ($p > n$). As diversas comparações encontradas na literatura foram realizadas considerando: dados sem *outliers*, dados com *outliers* de mesma média que as observações normais porém com variância diferente (*outliers* de dispersão), dados com *outliers* de médias e variâncias diferentes das observações normais (*outliers* de locação e dispersão) e por fim dados com alta dimensão, como exemplo pode ser citado por Filzmoser, Maronna e Werner (2008), Sajesh e Srinivasan (2012), Barbosa (2021) e Cabana, Lillo e Laniado (2021). A principal crítica é grande influência da alta dimensionalidade na eficiência da detecção de *outlier* nos métodos existentes, visto que quando o número de variáveis é muito maior que o número de observações é mais difícil detectar o comportamento "padrão" dos dados (CHUNG; AHN, 2021). Para Chung e Ahn (2021) esse problema de detecção em dados com alta dimensão fica mais intensificado quando o tamanho amostral é pequeno.

O objetivo desse trabalho é apresentar e comparar os seguintes métodos de detecção de *outliers*: MVE, MCD, OGK, PCOut e *comedian*. Para realizar as comparações foi utilizado uma série de simulações prevendo diversas situações utilizando a distribuição normal contaminada. As comparações foram avaliadas por meio da taxa de sucesso (TS), que aponta a porcentagem de *outliers* que os métodos identificaram corretamente e da taxa de falsa detecção (TFD), que aponta a porcentagem de observações que não são *outliers*, mas foram identificadas como *outliers*.

Por fim, essa pesquisa está dividida em quatro seções, sendo a primeira o referencial teórico contendo explicação de *outliers*, exemplos dos problemas de mascaramento e inundação, a descrição para obter os estimadores robustos (estimador pode ser considerado robusto quando a estatística que será usada como estimador ser menos sensível à valores discrepantes) para detecção dos *outliers* e trabalhos relacionados. A segunda seção é a metodologia utilizada que contém explicações de como gerar a dados com distribuição normal contaminada, como usar as estimativas robustas para detectar os *outliers* e como foram feitas as simulações. A terceira seção contém os resultados das aplicações dos métodos de detecção nos dados simulados. Por fim, a última seção é composta pela conclusão dos resultados.

2 REFERENCIAL TEÓRICO

2.1 Exemplo de *outlier*

Para introduzir o conceito de detecção de *outliers* é interessante observar o caso univariado, em que se utilizou de uma demonstração parecida com a proposta apresentada por Rousseeuw e Hubert (2011). Considere o seguinte conjunto de dados com 5 observações:

$$X = 5,15; \quad 5,23; \quad 5,13; \quad 5,18; \quad 5,17 \tag{2.1}$$

A média pode ser calculada por meio da estatística $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Nesse caso $\bar{x} = 5,172$. Agora supondo que um desses dados está com erro de digitação, ou seja,

$$X = 5,15; \quad 5,23; \quad 5,13; \quad 51,8; \quad 5,17 \tag{2.2}$$

Para o conjunto (2.2) (com erro de digitação), a média obtida foi de 14,496. No entanto, para ambos conjuntos a mediana será 5,17. Isso demonstra facilmente o quanto a média é sensível à potenciais valores discrepantes na amostra. A mediana, por outro lado, por ser uma medida de posição, é considerada robusta (menos sensível aos valores discrepantes). O ponto de ruptura é uma medida dos estimadores robustos que funciona como um alerta contra *outliers*. Em outras palavras, o ponto de ruptura indica a menor fração (sempre entre 0 e 1, ou expresso em porcentagem) de *outliers* em um conjunto de dados que quebre o estimador, ou seja, faça com que o estimador apresente eventualmente estimativas pobres ou ruins. A interpretação de estimativas pobres ou ruins, depende de todo um contexto e fica sempre a critério do pesquisador (HUBERT; DEBRUYNE, 2009). Para a média o ponto de ruptura é $1/n$, pois como visto qualquer observação pode alterar o valor da média. No caso da mediana o ponto de ruptura seria de $\frac{1}{n}\left[\frac{n+1}{2}\right]$ (HUBERT; DEBRUYNE, 2009). O ponto de ruptura do desvio padrão usando a seguinte equação $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$ será o mesmo da média($\frac{1}{n}$). Para saber mais sobre ponto de ruptura o artigo de Hubert e Debruyne (2009) pode ser consultado.

No conjunto de dados (2.1), $s = 0,0376$, enquanto no conjunto de dados (2.2) $s = 20,8536$. Logo, a média e o desvio padrão não são estatísticas robustas. Como sugestão para esse exemplo, pode-se citar o desvio absoluto da mediana (*median absolute deviation* - MAD),

como outra estatística robusta, a qual é dada por

$$MAD(X) = 1,483\, med(|x_i - med(x_i)|), \tag{2.3}$$

em que *med* significa mediana e 1,483 é uma constante que representa o fator de correção que torna *MAD* imparcial para distribuição normal. Nesse caso para ambos os conjuntos de dados o $MAD(x) = 0,0296$. Supondo que os dados possui distribuição normal, espera-se que o centro seja μ e o parâmetro de escala seja σ. Dessa maneira, como modelo clássico para detecção de *outliers* univariados pode ser usada a distribuição normal padrão (ROUSSEEUW; HUBERT, 2011), cuja estatística é

$$z_i = (x_i - \bar{x})/s \tag{2.4}$$

em que s é o desvio padrão. A observação será considerada um *outlier*, usando um critério conservador, se $|z_i| > 2,5$ (ROUSSEEUW; HUBERT, 2011). No caso do conjunto de dados (2.2) (com erro de digitação), usando a média, o desvio padrão e este critério, para nenhum dos 5 dados foi encontrado um *outlier*. O conjunto dos valores de z é:

$$z_i = (0,4481; 0,4443; 0,4491; 1,788; 0,4472)$$

Isso porque o valor da média e do desvio padrão foram atraídos pelo *outlier*. No entanto se for utilizado a mediana e o *MAD* como parâmetros robustos por meio da seguinte modificação no estimador z.

$$(|x_i - med(x)|)/MAD, \tag{2.5}$$

é possível encontrar facilmente o *outlier* do conjunto de dados. Observe os resultados da equação (2.5) são:

$$0,6756; \quad 2,027; \quad 1,35; \quad 1.575,33; \quad 0.0$$

Esse exemplo apresentado demonstra duas características importantes. Primeira, na detecção de *outliers* é necessário uma medida de centro e uma de escala. Segundo, o uso de estimadores robustos na detecção de *outliers* aumenta as chances de sucesso de acerto. Embora no caso univariado a detecção de *outliers* possa ser feita graficamente sem muitas dificuldades, para dados multivariados a detecção fica mais complexa (ROCKE; WOODRUFF, 1996). Isso acontece porque a presença de *outliers* pode mascarar outros *outliers* e pode fazer também que uma observação que não é *outlier* seja classificada como *outlier*. Além disso, uma amostra

pode não conter *outliers* quando observada variável por variável isoladamente, mas o que muda na análise multivariada em que as variáveis conjuntamente são consideradas e os potenciais *outliers* são mais facilmente detectados.

Na literatura existem diversos métodos para detecção de *outliers*. Na maioria dos métodos propostos pelos pesquisadores, os testes realizados utilizam a distribuição multivariada contaminada para realizar simulação e alguns casos há alguma comparação com outros métodos. Dentre essas comparações do novo método os modelos robustos mais utilizados são: MVE, MCD e OGK. Adiante nessa pesquisa apresenta-se o conceito de distância de Mahalanobis e seus principais problemas. Em seguida será demonstrado como obter estimadores robustos usando os métodos, MVE, MCD, *Comedian*, OGK e PCOut.

2.2 Distância de Mahalanobis (DM)

Dentre muitos métodos propostos na literatura para detecção de *outliers*, a distância generalizada de Mahalanobis tem sido muito utilizada (SAJESH; SRINIVASAN, 2012), cujo interesse recai na distância entre as obervações $\boldsymbol{X}$ e a média. A distância de Mahalanobis entre uma observação aleatória $\boldsymbol{X}$ e o centro $\boldsymbol{M}$ é

$$DM(\boldsymbol{X},\boldsymbol{M}) = (\boldsymbol{X}-\boldsymbol{M})^{\top}\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{M}), \tag{2.6}$$

em que, $\boldsymbol{X}$ é observação multivariada com p variáveis, $\boldsymbol{M}$ o centro, geralmente representado pela média de $\boldsymbol{X}$ com dimensão p e $\boldsymbol{V}$ a matriz de covariância $p \times p$. Hadi (1992) também cita essa equação como sendo amplamente usada na detecção de *outliers*, porém obtendo sua raiz quadrada. Por se tratar da distância da observação em relação ao vetor de médias e levando em consideração a matriz de covariância, ela indica que um valor alto de *DM* pode ser um *outlier*. Porém, nem sempre uma observação com *DM* alto é um *outlier* e nem sempre uma observação com *DM* baixo não é um *outlier*. Esses dois problemas geralmente ocorrem em dados com muitos *outliers*, em que a presença de um *outlier* muito maior que um grupo de *outliers*, pode mascarar alguns *outliers*, visto que o *DM* tende a ser mais baixo. O contrário também pode ocorrer, ou seja, imagine uma quantidade de *outliers* muito próxima e longe das observações de maior densidade, de modo que isso fosse capaz de alterar a matriz de covariância. Dessa forma, a *DM* ficaria alta para as observações que não são *outliers*. Esses problemas são conhecidos como mascaramento (*masking*) e inundação (*swamping*), respectivamente. Eles acontecem por-

que o vetor de média $\boldsymbol{M}$ e $\boldsymbol{V}_{p\times p}$, matriz de covariância amostrais, não são robustas e, portanto, são mais sensíveis as alterações nas observações (HADI, 1992).

A função densidade de probabilidade da normal multivariada é:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}. \tag{2.7}$$

Observe que ela depende da distância de Mahalanobis $(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Fazendo uma análise mais profunda, pode-se verificar que essa expressão são hiperelipsóides no espaço p-dimensional (FERREIRA, 2008). Abaixo segue o teorema 3.8 da referida citação, sobre Elipsóides de concentração:

> "**Teorema 3.8** (Elipsóides de concentração). Se o vetor aleatório $\boldsymbol{X} \in R^p$ segue uma distribuição normal multivariada com função de densidade de probabilidade $f_{\boldsymbol{X}}(\boldsymbol{x})$, dada em 2.7, com média $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$, então
>
> $$(\boldsymbol{X}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$$
>
> tem distribuição qui-quadrado com p graus de liberdade (χ^2_p) e a região
>
> $$(\boldsymbol{X}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \leq \chi^2_{\alpha,p}$$
>
> define a concentração 100(1- α)% das realizações das variáveis aleatórias, em que $\chi^2_{\alpha,p}$ é o quantil superior a 100α% da distribuição de qui-quadrado com $\nu = p$ graus de liberdade, obtido de acordo com a afirmativa probabilística $P(\chi^2_p > \chi^2_{\alpha,p}) = \alpha$ (FERREIRA, 2008)."

Esse teorema é importante para compreensão do método de elipsóide de volume mínimo (MVE), que será apresentado mais adiante e também para determinar o valor de corte. O valor de corte é o quantil de distribuição $\chi_{\alpha,p}$. Quando a DM da observação multivariada é maior que o $\chi_{\alpha,p}$ trata-se de um *outliers* (em geral utiliza-se $\alpha = 0,975$).

2.3 Mascaramento e inundação

Como já mencionado, ao utilizar o vetor de médias e matriz de covariância tradicionais o uso da distância de Mahanalobis pode levar à problemas com mascaramento e inundação. A seguir segue um exemplo de cada caso para demonstrar os efeitos citados. Vale ressaltar que os exemplos foram criados apenas para demonstração e não são dados reais.

2.3.1 Exemplo 01 - Mascaramento

O problema de mascaramento ocorre quando se tem dois ou mais *outliers*, sendo que um *outliers* está muito mais distante dos demais *outliers*. Isso faz com que a distância de Mahalanobis fique baixa para os *outliers* menos distantes. Observe a Tabela (2.1), claramente as observações 12, 13, 14,e 15 são *outliers*. Porém pode-se notar que a observação 15 está mais distante dos outros *outliers*.

Tabela 2.1 – Observações por indivíduos - Exemplo 01

Indivíduo (i)	X	*DM*
1	2	0,591
2	9	0,040
3	3	0,473
4	6	0,197
5	4	0,367
6	5	0,275
7	8	0,079
8	5	0,275
9	6	0,197
10	7	0,131
11	7	0,131
12	20	0,480
13	20	0,480
14	20	0,480
15	50	9,797

Fonte: Do autor (2022).

Observe que a observação 15 atraiu os valores da média e variância. Isso fez com que o valor da *DM* das observações 12, 13 e 14 ficassem mascaradas pelo *outlier* de maior valor. O valor de corte foi $\chi_{0,975;1} = 5,02$. Vale ressaltar que, os dados utilizados nesse exemplo são apenas para ilustrar o efeito de mascaramento e não tem nenhuma intenção de verificar se há relação entre eles.

2.3.2 Exemplo 02 - Inundação

Os casos de inundação ocorrem quando um grupo de *outliers* altera a média e covariância de modo que algumas obervações que não são *outliers* sejam tratadas como *outliers*. A Tabela 2.2 apresenta 50 observações com duas variáveis. Os dados apresentados possuem 15 *outliers*, sendo eles as observações 12, 13, 14, 15, 27, 28, 29, 30, 39, 40, 41, 42, 48 ,49, e 50.

Tabela 2.2 – Observações por indivíduos - Exemplo 02

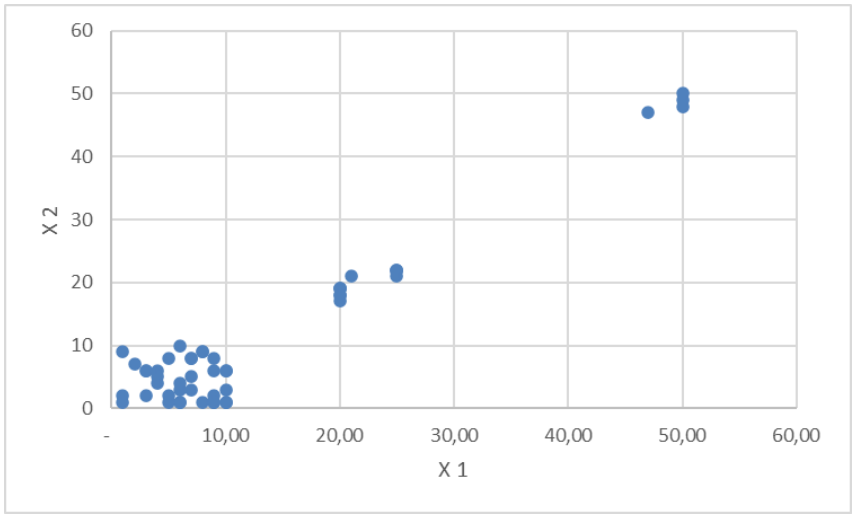
Indivíduo (i)	X1	X2	*DM*	Indivíduo (i)	X1	X2	*DM*
1	10	3	2,23	26	1	1	1,05
2	8	9	0,69	27	20	18	0,29
3	8	9	0,69	28	20	19	0,35
4	5	1	0,86	29	21	21	0,67
5	6	1	1,20	30	47	47	7,13
6	10	6	0,47	31	7	8	0,75
7	8	1	2,37	32	7	5	0,24
8	4	4	0,69	33	7	8	0,75
9	9	2	2,30	34	2	7	3,98
10	6	1	1,20	35	5	8	2,00
11	3	6	2,18	36	3	2	0,64
12	20	19	0,35	37	4	6	1,46
13	25	21	1,15	38	5	2	0,56
14	25	22	0,92	39	20	19	0,35
15	50	48	8,29	40	20	17	0,37
16	10	1	4,17	41	25	22	0,92
17	3	6	2,18	42	50	49	8,47
18	9	6	0,24	43	6	4	0,32
19	6	3	0,46	44	10	1	4,17
20	7	3	0,67	45	10	6	0,47
21	9	1	3,19	46	9	8	0,14
22	6	10	2,70	47	4	5	1,00
23	**1,00**	**9**	**7,76**	48	20	18	0,29
24	1	2	1,36	49	25	22	0,92
25	9	1	3,19	50	50	50	8,81

Fonte: Do autor (2022).

Observando a Tabela 2.2 pode-se verificar que as observações 15, 23, 42 e 50 possuem as *DM's* mais altas e poderiam ser classificadas como *outliers*. Porém a observação 23 em negrito não é um *outlier*. O gráfico de dispersão apresentado na Figura 2.1 mostra as posições dos dados. Com relação a *DM* pode-se concluir que as observações 12, 13, 14, 27, 28, 29,30, 39, 40, 41, 48 e 49 ficaram mascaradas e a observação 23 foi inundada pela presença dos *outliers*. Vale ressaltar que nesse exemplo o valor de corte foi $\chi_{0,975;2} = 7,37$.

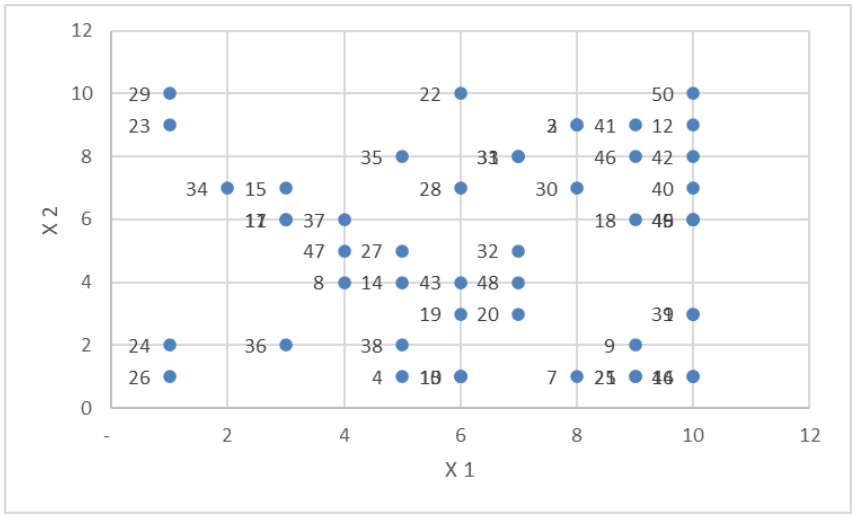
Importante dizer: mesmo que não haja *outliers* nas marginais dos conjuntos de dados X1 e X2, ainda assim pode haver *outliers* quando esses dados estiveram combinados. Observando a Figura 2.2, a observação 29 no ponto (1;10), está um pouco distante das demais e a *DM* ficou muito próxima de ser classificada como *outlier*, chegando em 6,57.

Figura 2.1 – Gráfico de dispersão de X1 e X2 com a presença de outliers.



Fonte: Do autor (2022).

Figura 2.2 – Gráfico de dispersão de X1 e X2 sem a presença de outliers.



Fonte: Do autor (2022).

A Tabela 2.3 apresenta todas as *DM's* obtidas no conjunto de dados sem a presença de *outliers* nas variáveis independentes. Como ja mencionado a observação 29 obteve o maior valor da *DM*. Embora nesse exemplo o valor de corte seria $\chi_{0,975;2} = 7,37$. Mas ja é possível perceber que pode ocorrer a presença de *outliers* em dados que não possuem *outliers* nas marginais.

Tabela 2.3 – Observações por indivíduos - Exemplo 02

Indivíduo (i)	X1	X2	*DM*	Indivíduo (i)	X1	X2	*DM*
1	10	3	1,94	26	1	1	5,47
2	8	9	1,94	27	5	5	0,30
3	8	9	1,94	28	6	7	0,46
4	5	1	2,09	29	1	10	6,57
5	6	1	1,85	30	8	7	0,41
6	10	6	1,52	31	7	8	0,97
7	8	1	2,10	32	7	5	0,02
8	4	4	0,91	33	7	8	0,97
9	9	2	1,79	34	2	7	2,98
10	6	1	1,85	35	5	8	1,27
11	3	6	1,65	36	3	2	2,52
12	10	9	3,09	37	4	6	0,90
13	6	1	1,85	38	5	2	1,31
14	5	4	0,41	39	10	3	1,94
15	3	7	1,98	40	10	7	1,82
16	10	1	3,32	41	9	9	2,40
17	3	6	1,65	42	10	8	2,35
18	9	6	0,81	43	6	4	0,16
19	6	3	0,50	44	10	1	3,32
20	7	3	0,50	45	10	6	1,52
21	9	1	2,59	46	9	8	1,65
22	6	10	2,75	47	4	5	0,79
23	1	9	5,57	48	7	4	0,15
24	1	2	4,71	49	10	6	1,52
25	9	1	2,59	50	10	10	4,06

Fonte: Do autor (2022).

No apêndice A, foi apresentado a utilização dos métodos robustos para detecção de *outliers* usando dados da Tabela 2.1. Os métodos robustos usados foram MVE, MCD, *Comedian*, OGK e PCOut. A seguir segue os procedimentos de cada método.

2.4 Estimador robusto: Elipsóide de volume mínimo (MVE)

O método MVE tem o objetivo de encontrar estimadores robustos, ou seja, menos sensíveis aos valores das observações discrepantes, e portanto, é possível evitar os problemas na detecção de *outliers* encontrados na distância de Mahalanobis usando a matriz de covariância e média. Os estimadores robustos utilizados no MVE são, a matriz de covariância $C(\boldsymbol{X}_j)$ e o centro $T(\boldsymbol{X}_j)$, definidos adiante.

Rousseeuw e Zomeren (1990) apresentaram um algoritmo para calcular o MVE por meio de re-amostragens. A ideia do algoritmo de re-amostragens é criar subamostras j com tamanho $p+1$, onde p é o número de variáveis do conjunto de dados (a subamostra j pode assumir tamanho entre $p+1 < j < n$, em que n é o número de obervações do conjunto de dados, por padrão o algoritmo fica definido em $p+1$) e obter a matriz de covariância ($\boldsymbol{C}_j$) e as médias de $\boldsymbol{x_j}$ como $T(\boldsymbol{X}_j)$.

O MVE pode ser definido como um par de $(\boldsymbol{T},\boldsymbol{C})$, em que $T(\boldsymbol{X})$ é um vetor de dimensão p e $C(\boldsymbol{X})$ é uma matriz positiva semi-definida de tamanho $p \times p$. Serão escolhidos os subconjuntos que atenda a seguinte equação (ROUSSEEUW; ZOMEREN, 1990):

$$\#\{i;(\boldsymbol{x}_i - T)^\top C^{-1}(\boldsymbol{x}_i - T) \leq a^2\} \geq h \tag{2.8}$$

em que # é o número de observações presente no conjunto, $h = [(n+p+1)/2]$, sendo que h assume somente valores inteiros, a^2 é um valor fixo obtido por meio da distribuição χ^2_p, normalmente usado com $\chi^2_{p,0.5}$, desde que $\boldsymbol{X}$ tenha distribuição normal multivariada. Essa equação também garante que dado a^2, o par de $(\boldsymbol{T},\boldsymbol{C})$ tenha pelo menos h pontos dentro da elipse.

A distância robusta do MVE é definida na equação (2.9).

$$RD_i = \sqrt{(\boldsymbol{x}_i - T(\boldsymbol{X}))^\top C(\boldsymbol{X})^{-1}(\boldsymbol{x}_i - T(\boldsymbol{X}))}, \qquad i = 1,...,n. \tag{2.9}$$

Como estimadores robustos do MVE pode usar as médias ponderadas (com os pesos) e a matriz de covariância ponderada (ROUSSEEUW; ZOMEREN, 1990) por:

$$T_1(\boldsymbol{X}) = \left(\sum_{i=1}^{n} w_i\right)^{-1} \sum_{i=1}^{\mathrm{n}} \boldsymbol{x}_i wi \tag{2.10}$$

e

$$C_1(\boldsymbol{X}) = \left(\sum_{i=1}^{\mathrm{n}} w_i - 1\right)^{-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - T(\boldsymbol{X}))^\top (\boldsymbol{x}_i - T(\boldsymbol{X})) \tag{2.11}$$

em que os pesos $w_i = w(RD_i)$ dependem das distâncias robustas. A ideia é encontrar a subamostra com volume da elipse mínimo. Os passos principais do algoritmo segue descritos a seguir. O primeiro passo consiste em obter subamostras com $p+1$ observações diferentes, indexadas

em $j = (i_1, ..., i_{p+1})$. Depois encontrar as médias e covariâncias usando as seguintes equações:

$$\boldsymbol{T}_j = \frac{1}{p+1} \sum_{i \in j} \boldsymbol{x}_i \tag{2.12}$$

e

$$\boldsymbol{C}_j = \frac{1}{p} \sum_{i \in j} (\boldsymbol{x}_i - T(\boldsymbol{X}))(\boldsymbol{x}_i - T(\boldsymbol{X}))^\top \tag{2.13}$$

Observe que as médias e covariâncias são as tradicionais, porém ao invés de serem calculadas com a matriz de dados $\boldsymbol{X}$ completa, foi realizado o procedimento com apenas $p+1$ observações.

O próximo passo é encontrar e separar os subconjuntos que atende a equação (2.8). Note que implicitamente nessa equação encontra-se as distância de Mahalanobis. Nos subconjuntos selecionados deve-se aplicar a equação

$$m_j^2 = med(\boldsymbol{x}_i - T_j)^\top \boldsymbol{C}_j^{-1} (\boldsymbol{x}_i, -T_j), \tag{2.14}$$

com o objetivo de encontrar o fator de ampliação (WOODRUFF; ROCKE, 1994), em que m_j^2 é este fator de ampliação e *med* é a mediana das distâncias encontradas. Nesse caso interessa o menor valor de $m_j^{2p} det(\boldsymbol{C}_j)$ (que representa o volume do elipsóide). O melhor subconjunto será aquele que apresentar o menor volume. Depois para encontrar o melhor subconjunto o algoritmo ajusta a matriz de covariância e vetor de médias usando a equação

$$T(\boldsymbol{X}) = \boldsymbol{T}_J \quad \text{e} \quad C(\boldsymbol{X}) = (\chi^2_{0,5;p})^{-1} m_J \boldsymbol{C}_J. \tag{2.15}$$

Com o vetor de médias e matriz de covariância ajustados, pode-se aplicar as equações (2.10) e (2.11) para fazer a ponderação. Para desenvolver a ponderação w_i segue a seguinte equação:

$$w_i = \begin{cases} 1 & se \quad (\boldsymbol{x}_i - T(\boldsymbol{X}))^\top C(\boldsymbol{X})^{-1} (\boldsymbol{x}_i - T(\boldsymbol{X})) \leq v_c \\ 0 & \text{caso contrário} \end{cases} \tag{2.16}$$

em que $v_c = \chi_{0,975;p}$ definido como valor de corte.

O vetor e a matriz de covariância encontrada na equação (2.10) e equação (2.11) são estimativas robustas para ao vetor de média e matriz de covariância que deverão ser usadas no calculo da distância de Mahanalobis.

2.5 Estimador robusto: Covariância de Determinante Mínimo (MCD)

O método MCD procura pelo determinante mínimo para a matriz de covariância de subconjuntos da matriz $\boldsymbol{X}$ (HADI, 1992). Embora parecido conceitualmente com o MVE, segundo Rousseeuw e Driessen (1999) existem algumas vantagens ao utilizar o MCD, como, por exemplo, a eficiência da estatística que é assintoticamente normal, sendo que as distâncias robustas são precisas e, portanto, mais adequadas para detectar *outliers*. Por outro lado, o MCD era difícil de ser calculado, até o surgimento do algoritmo Fast-MCD (ROUSSEEUW; DRIESSEN, 1999).

Por meio de qualquer aproximação MCD é possível computar outra aproximação que leva ao determinante mínimo (ROUSSEEUW; DRIESSEN, 1999). Considere um conjunto de dados $\boldsymbol{X}_n = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, com p variáveis observadas. Seja $H_1 \subset \{1, ..., n\}$, sendo $|H_1|$ com h observações, $\boldsymbol{T}_1 = \frac{1}{h} \sum_{i \in H_1} \boldsymbol{x}_i$ e $\boldsymbol{S}_1 = \frac{1}{h} \sum_{i \in H_1} (\boldsymbol{x}_1 - \boldsymbol{T}_1)(\boldsymbol{x}_1 - \boldsymbol{T}_1)^\top$. Caso o $det(\boldsymbol{S}_1) \neq 0$, calcula-se a distância por meio da equação:

$$d_i(i) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{T}_i)^\top \boldsymbol{S}_i^{-1} (\boldsymbol{x}_i - \boldsymbol{T}_i)}, \text{ para } i = 1, ..., n. \tag{2.17}$$

Depois, considera-se um subconjunto H_2 de modo que $\{d_i(i); i \in H_2 = \{(d_1)_{1:n}, \cdots, (d_1)_{h:n}\}$ em que $\{(d_1)_{1:n} \leq (d_1)_{2:n} \leq ... \leq (d_1)_{h:n}\}$ são as distâncias ordenadas, ou seja, H_2 será preenchido com as observações de 1 até h, em ordem. Então calcule $\boldsymbol{T}_2$ e $\boldsymbol{S}_2$ baseados em H_2, dessa forma $det(\boldsymbol{S}_2) \leq det(\boldsymbol{S}_1)$, a prova disso pode ser encontrada nos apêndices de Rousseeuw e Driessen (1999).

As equações apresentadas para esse algoritmo requerem que o $det(\boldsymbol{S}_1) \neq 0$, ou seja, $det(\boldsymbol{S}_1) = 0$ é uma restrição. Nesse caso, Rousseeuw e Driessen (1999) recomendaram acrescentar observações no sub-conjunto inicial até que o $det(\boldsymbol{S}_1) \neq 0$. Quando $det(\boldsymbol{S}_1) > 0$ basta repetir o processo anteriormente descrito até que o determinante não seja mais reduzido, desse modo pode-se dizer que houve convergência e, portanto, o processo é finalizado (ROUSSEEUW; DRIESSEN, 1999).

Para definir a quantidade de observações h, pode ser usando como padrão $h = [(n + p + 1)/2]$, ou ainda, qualquer inteiro em que h esteja entre $[(n + p + 1)/2 \leq h \leq n]$. Caso a contaminação seja menor que 25%, recomenda-se adotar $h = 0,75n$. Se $h = n$, então a quantidade $\boldsymbol{T}$ é a própria média dos dados e $\boldsymbol{S}$ é a própria covariância, portanto não faz sentido a busca por

um determinante mínimo e quando $p = 1$, tem-se o caso de dados univariados (ROUSSEEUW; DRIESSEN, 1999).

O algoritmo Fast-MCD é inicializado com um subconjunto H_0 com $p+1$ observações, então calcula-se a média e a matriz de variância do subconjunto H_0. As regras seguem às mesmas do procedimento já mencionado. Se $det(S_0) = 0$ adiciona-se mais observações até que o $det(S_0) > 0$. Então calcula-se as distâncias $d_0^2(i) = (\boldsymbol{x}_i - \boldsymbol{T}_0)^\top \boldsymbol{S}_0^{-1}(\boldsymbol{x}_i - \boldsymbol{T}_0)$ e as colocar em ordem. Em seguida é construído o conjunto H_1 com a primeira até a h-ésima observação (tamanho de h foi definido acima). Depois, repetir os processos três vezes, até encontrar o $det(S_3)$. Rousseeuw e Driessen (1999) recomendaram fazer o procedimento com vários subconjuntos H_0 e separar os 10 determinantes menores dentre todos $det(S_3)$. Basta continuar o processo até a convergência com apenas esses determinantes. Serão selecionadas o subconjunto que obtiver o menor determinante. A matriz de covariância e o vetor de médias desse subconjunto serão as estimativas robustas para serem usada no calculo da distância de Mahanalobis.

2.6 Estimador robusto: *Comedian*

O método *Comedian* foi criado por Falk (1997) como uma alternativa de dependência entre duas variáveis aleatórias conhecidas. O estimador da covariância *Comedian* entre duas variáveis se trata de uma estatística robusta e pode ser definido como:

$$COM(X,Y) = med((X - med(X))(Y - med(Y))), \tag{2.18}$$

em que *med* representa a mediana. O *Comedian* é uma generalização do *MAD*, quando $X = Y$, $COM(X,Y) = MAD^2$ (ver equação (2.3) do MAD, lembrando que o valor da constante, 1,483, de correção também fica elevado ao quadrado). A $COM(X,Y)$ faz um paralelo com a covariância $COV(X,Y)$, mas a $COM(X,Y)$ vai sempre existir dado que ele é calculado por meio de uma medida de locação (a mediana), enquanto a $COV(X,Y)$ não necessariamente existe (CABANA; LILLO; LANIADO, 2021). Além disso, *Comedian* é simétrico, invariante em relação a translações e à escala, isto é: $COM(X, aY + b) = aCOM(X,Y) = aCOM(Y,X)$ (SAJESH; SRINIVASAN, 2012). O coeficiente de correlação *Comedian* estabelecido é definido pela equação:

$$\delta(X,Y) = \frac{COM(X,Y)}{[MAD(X)MAD(Y)]}, \tag{2.19}$$

com $\delta \in [-1,1]$ para dados bivariados.

Sendo o *Comedian* um método de estatística robusto e como ele é uma técnica multivariada, pode-se utilizá-lo para detecção de *outliers*. Considere $\boldsymbol{X}$ uma matriz $n \times p$, com linhas $\boldsymbol{X}_i^T (i = 1,2,...,n)$ e colunas $\boldsymbol{X}_j (j = 1,2,...,p)$. Então a matriz $\boldsymbol{COM(X)}$ é dada por:

$$\boldsymbol{COM(X)} = (COM(\boldsymbol{X}_i, \boldsymbol{X}_j)), i, j = 1,2,...,p. \tag{2.20}$$

Da mesma forma pode ser obtido a matriz de coeficiente de correlação:

$$\delta(\boldsymbol{X}) = \boldsymbol{D}COM(\boldsymbol{X})\boldsymbol{D}^\top, \tag{2.21}$$

sendo $\boldsymbol{D}$ a matriz diagonal dos elementos $1/MAD(\boldsymbol{X}_i)(i = 1,2,...,p)$ (SAJESH; SRINIVASAN, 2012). Uma desvantagens que caso os dados das variáveis X e Y sejam simétricos, então $med(X) = med(Y) = 0, med(X^2) = med(Y^2) = 1$, portanto $MAD(X) = MAD(Y) = 1$,isso implica que $COM(X,Y) > 1$. Isso porque, a matriz de covariância *Comedian* nem sempre é positiva definida, sendo em geral semi-definida. Essa questão já foi tratada por Maronna e Zamar (2002). Eles apresentaram um método geral para encontrar matrizes robustas de covariância positiva definida e aproximadamente invariante em relação as transformações lineares. Para obter essa transformação de matrizes semidefinidas em matrizes positivas definidas, o autor utilizou os passos a seguir.

1 calcular os autovalores λ_j, autovetores $\boldsymbol{e}_j$ de $\delta(\boldsymbol{X})(j = 1,2,...,p)$, e chame $\boldsymbol{E}$ de matriz cujas colunas são os $\boldsymbol{e}_j^\top$s, de modo que $\delta(\boldsymbol{X}) = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^\top$, onde $\boldsymbol{\Lambda} = diag(\lambda_1,...,\lambda_1)$.

2 Seja $\boldsymbol{Q} = D(\boldsymbol{X})^{-1}\boldsymbol{E}$, onde $\boldsymbol{D}$ é definido como a matriz diagonal dos elementos $1/MAD(\boldsymbol{X}_i)(i = 1,2,...,p)$ e $\boldsymbol{z}_i = \boldsymbol{Q}^{-1}\boldsymbol{X}_i$.

3 As estimativas robustas resultantes para locação $(m(\boldsymbol{X}))$ e dispersão $(S(\boldsymbol{X}))$ são então definidas como:

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

em que $\boldsymbol{\Gamma} = diag(MAD(\boldsymbol{Z}_1)^2,...,MAD(\boldsymbol{Z}_p)^2)$ e $\boldsymbol{l} = (med(\boldsymbol{Z}_1),...,med(\boldsymbol{Z}_1))^\top$.

Segundo Maronna e Zamar (2002) as estimativas podem ser aperfeiçoadas por meio de um processo iterativo, substituindo $\boldsymbol{\delta}$ por $\boldsymbol{S}$ e repetir os passos citados acima.

2.7 Estimador robusto: Ortogonalizado de Gnanadesikan e Kettenring (OGK)

O estimador de Gnanadesikan e Kettenring (1972) tem como base a covariância definida da seguinte forma:

$$cov(X,Y) = \frac{1}{4}(\sigma(X+Y)^2 - \sigma(X-Y)^2), \tag{2.23}$$

em que σ é o desvio padrão, X e Y é um par de variáveis aleatórias e $cov(X,Y)$ a matriz de covariância definida por Gnanadesikan e Kettenring. Segundo Maronna e Zamar (2002) o uso de σ como escalar, foi para se obter uma matriz de covariância robusta. O resultado encontrado foi uma matriz simétrica, embora não positiva-definida e também não equivariante.

Sejam $\boldsymbol{V}$ uma matriz de covariância p-dimensional, $\boldsymbol{X}$ um vetor aleatório e σ o desvio padrão, então

$$\sigma(\boldsymbol{a}^\top \boldsymbol{X})^2 = \boldsymbol{a}^\top \boldsymbol{V} \boldsymbol{a}, \tag{2.24}$$

para todo $\boldsymbol{a} \in R^p$. O estimador robusto de σ funciona como um escalar para algumas direções determinadas por $\boldsymbol{a}$, porém com o prejuízo que essa matriz deixa de ser positiva definida. O método proposto por Maronna e Zamar (2002) é uma modificação na equação (2.24) para obter uma matriz próxima de ser positiva definida. Para obter essa matriz de covariância os autores usaram a mesma técnica demonstrada na estimativa robusta *Comedian*. Embora o método para obtenção da matriz robusta de covariância seja o mesmo usado no *Comedian* as matrizes são muito diferentes. Enquanto no método OGK utiliza-se de um estimador robusto de locação e dispersão para criar a matriz de covariância robusta, no *Comedian*, utiliza-se a matriz de covariância obtida com o $\boldsymbol{COM(X)}$. Vale ressaltar que, existem mais de um estimador robusto que podem ser usados no método OGK, como por exemplo, MAD e IQR (por padrão o algoritmo usa o "τ escala", segundo Maronna e Zamar (2002) possue maior eficiência em dados com distribuição normal).

Seja $\boldsymbol{X}$ uma matriz $n \times p$ com $\boldsymbol{X}_i^\top (i = 1,...,n)$ linhas e $\boldsymbol{X}_j (j = 1,...,p)$ colunas. Sendo $\sigma(.)$ um estimador robusto univariado de dispersão, $\mu(.)$ um estimador robusto univariado de locação e $v(.,.)$ a estimativa robusta da covariância de duas variáveis aleatórias. A definição da matriz de escala $\boldsymbol{V(X)}$ e o vetor centro $\boldsymbol{t(X)}$ segue os seguintes passos:

1. Sejam $\boldsymbol{D} = diag(\sigma(\boldsymbol{X}_1),...,\sigma(\boldsymbol{X}_p))$ e $\boldsymbol{y}_i = \boldsymbol{D}^{-1}\boldsymbol{x}_i \qquad i = 1,...,n.$

2. Calcular a matriz de correlação $\boldsymbol{U} = [U_{jk}]$, usando v como estimador robusto de $\boldsymbol{Y}$, definido como:

$$U_{jj} = 1 \qquad e \qquad U_{jk} = v(\boldsymbol{Y}_j, \boldsymbol{Y}_k), \quad j \neq k$$

.

3. calcular os autovalores $\boldsymbol{\lambda}_j$, autovetores $\boldsymbol{e}_j$ de $U(j = 1,2,...,p)$, e chame $\boldsymbol{E}$ de matriz cujas colunas são os $\boldsymbol{e}_j^\top$, de modo que $U = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^\top$, onde $\boldsymbol{\Lambda} = diag(\lambda_1, ..., \lambda_p)$.

4. Seja $\boldsymbol{A} = \boldsymbol{D}(\boldsymbol{X})\boldsymbol{E}$ e $\boldsymbol{z}_i = \boldsymbol{A}^{-1}\boldsymbol{x}_i$. De modo que $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{z}_i$.

5. As estimativas robustas resultantes para locação $(t(\boldsymbol{X}))$ e dispersão$(V(\boldsymbol{X}))$ são então definidas como:

   $V(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^\top \quad e \quad t(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{v}$

   Em que $\boldsymbol{\Gamma} = diag(\sigma(\boldsymbol{Z}_1)^2, ..., \sigma(\boldsymbol{Z}_p)^2)$ e $\boldsymbol{v} = (\mu(\boldsymbol{Z}_1), ..., \mu(\boldsymbol{Z}_p))^\top$.

Como estimativa robusta da covariância de duas variáveis aleatórias $v(.,.)$ eles usaram o estimador de Gnanadesikan-Kettenring definido na equação (2.23). Segue:

$$\boldsymbol{U}_{jk} = \frac{1}{4}[\sigma(\boldsymbol{Y}_j + \boldsymbol{Y}_k)^2 - \sigma(\boldsymbol{Y}_j - \boldsymbol{Y}_k)^2], \quad j \neq k. \tag{2.25}$$

em que U_{jk} é a estimativa "orthogonalized Gnana- desikan-Kettenring"(OGK). O procedimento pode ter várias interações, mas, por padrão, o algorítimo usa apenas duas. Em seguida, a distância de Mahalanobis usando $\boldsymbol{V}$ e $\boldsymbol{t}$ são obtidas por:

$$d_i = d(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{t})^\top \boldsymbol{V}^{-1}(\boldsymbol{x}_i - \boldsymbol{t}) \tag{2.26}$$

em que $\boldsymbol{t} = \boldsymbol{T}(\boldsymbol{X})$ e $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{X})$. Por meio do cálculo destas distâncias foram atribuídos pesos W, ao centro $\boldsymbol{t}$ e à matriz de correlação $\boldsymbol{V}$. Para cada obervação $\boldsymbol{x_i}$ obtém-se um $w_i = W(d_i)$, as equações do centro e da covariância ponderadas com os pesos são:

$$\boldsymbol{t}_w = \frac{\sum_i w_i \boldsymbol{x}_i}{\sum_i w_i} \qquad \text{e} \qquad \boldsymbol{V}_w = \frac{\sum_i w_i (\boldsymbol{x}_i - \boldsymbol{t}_w)(\boldsymbol{x}_i - \boldsymbol{t}_w)^\top}{\sum_i w_i}. \tag{2.27}$$

O valor de corte ou ponto de rejeição para o modelo OGK pode ser definido como:

$$d_0 = \frac{\chi^2_{\alpha,p} med(d_1, ..., d_n)}{\chi^2_p(0,5)} \tag{2.28}$$

em que $\chi_{\alpha,p}$ é o quantil da distribuição qui-quadrado com p graus de liberdade e *med* é a mediana. A equação (2.26) apresenta as distâncias de cada observação para o centro e, portanto, qualquer valor de $d_i > d_0$ pode ser considerado um *outlier*.

2.8 Estimador robusto: Componentes principais para detecção de *outliers* (PCOut)

O uso da técnica de componentes principais tem como objetivo reduzir o número de variáveis. Por isso ele é indicado para dados com alta dimensão ($p > n$). O algoritmo proposto por Filzmoser, Maronna e Werner (2008) consistem em duas fases. A primeira fase detecta *outliers* de locação (possuem média diferentes dos dados normais) e a segunda fase detecta *outliers* de dispersão (*outliers* que possuem variância diferente dos dados normais). O primeiro passo da primeira fase é realizar a padronização ou redimensionamento de cada observação multivariada usando a mediana e o MAD por meio da equação

$$x_{ij}^* = \frac{x_{ij} - med(x_{1j},...,x_{nj})}{MAD(x_{1j},...,x_{nj})}, \qquad j = 1,...,p, \tag{2.29}$$

em que x_{ij}^* são os dados padronizados. Na sequência deve-se calcular a matriz de covariância dos dados padronizados e sem seguida obter os autovalores e autovetores. Apenas autovalores que mais contribuem com a variância total serão mantidos até obter 99% da variância total acumulada. Com a retirada dos componentes que pouco contribuíram com a variância total, obtém-se uma nova matriz com dimensão p^*. Com essa nova matriz, $\boldsymbol{V}_{p^* \times p^*}$ de autovetores obtém-se a matriz dos componentes principais ($\mathbf{Z}$) por meio da equação abaixo:

$$\boldsymbol{Z} = \boldsymbol{X}^* \boldsymbol{V},$$

em que $\mathbf{X}^*$ é a matriz com os elementos x_{ij}^*. Os componentes principais são usados para ponderar novamente a mediana e o MAD com a equação:

$$z_{ij}^* = \frac{z_{ij} - med(z_{1j},...,z_{nj})}{MAD(z_{1j},...,z_{nj})}, \qquad j = 1,...,p^*. \tag{2.30}$$

Depois de realizado a ponderação da matriz de componentes principais, inicia-se a detecção dos *outliers* usando o valor absoluto de uma medida robusta de Kurtosis por meio da equação

$$w_j = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(z_{ij}^* - med(z_{1j}^*,...,z_{nj}^*))^4}{MAD(z_{1j}^*,...,z_{nj}^*)^4} - 3 \right|, \quad j = 1,...,p^* \tag{2.31}$$

em que **W** é vetor de peso das componentes. Se fizer $w_j/\sum_i w_i$ obtém-se uma medida relativa entre $0 \leq w_j \leq 1$, que é a probabilidade de revelar *outliers*, ou seja, quanto mais próximo de 1 maior a probabilidade da componente possuir algum *outliers* e quanto mais próximo de zero indica que os dados estão todos normalmente distribuídos, portanto não há *outliers*. Essa probabilidade é usada para reponderar o conjunto de dados Z e obter as distâncias por meio da seguinte equação

$$\boldsymbol{Zr}_i = \boldsymbol{Z} \quad diag\left(w_j/\sum_i w_i\right), \tag{2.32}$$

em que Zr são os dados reponderados com o coeficiente de Kurtosis dado por

$$RD_i = \sum_{j=1}^{p}\left(\frac{Zr_{ij} - \mu(Zr_j)}{\sigma(Zr_j)}\right)^2, \quad i = 1,...,n, \tag{2.33}$$

em que RD_i é a distância dos dados reponderados para o centro, zr_{ij} são os dados reponderados usando $w_j/\sum_i w_i$, Zr_j são os p componentes principais, μ é um estimador robusto de locação e σ é um estimador robusto de dispersão. Porém, quando aplica-se o coeficiente de curtose para ponderar os componentes principais, perde-se qualquer semelhança com a distribuição χ^2. Então é necessário fazer um ajuste na distância usando a equação

$$d_i = RD_i \frac{\sqrt{\chi^2_{0,5;p}}}{med(RD_1,...,RD_n)} \quad para \quad i = 1,...,n, \tag{2.34}$$

em que d_i é a distância aproximada com distribuição χ^2.

A segunda fase do algoritmo é semelhante a primeira, porém não faz uso do coeficiente de curtose para reponderar os componentes principais. Nesse caso, com os dados obtidos na equação (2.30) aplica-se a função da distância de Mahanalobis. Como não foi aplicado a ponderação usando o coeficiente de curtose o resultado obtido já possui distribuição χ^2. Usando os dois passos encontra-se duas distâncias, sendo a primeira encontrada por meio da locação e a segunda por meio da dispersão dos dados.

2.9 Artigos e trabalhos relacionados

Rocke e Woodruff (1996) apresentaram um método para detecção de *outliers*. Eles abordaram a fundo questões sobre os *outliers* em dados multivariados e nas razões porque eles nem sempre são detectados de maneira simples. Por meio do estudo apresentado, os autores cria-

ram o método híbrido, baseado no determinante mínimo da matriz de covariância (*Minimum Covariance Determinant* - MCD) de Rousseeuw e Zomeren (1990), para detecção de *outliers*, mas não fizeram nenhuma comparação com outros métodos. Rocke e Woodruff (1996) apenas simularam diversas situações para detecção de *outliers*, envolvendo diferentes dimensões e número de observações. Por outro lado, tiveram o cuidado de apresentar o tempo de execução do algoritmo (até então, o tempo de execução era um problema para detecção de *outliers* em análises multivariadas), além de apresentarem também às *taxas de sucesso* (TS). Nas simulações apresentadas ficou claro que o aumento da taxa de contaminação e das dimensões (p) o método proposto perde eficiência.

Filzmoser, Maronna e Werner (2008) introduziram um método para detecção de *outliers* fazendo uso de analises de componentes principais. O método foi indicado principalmente para dados com alta dimensão. Por meio da aplicação da análise de componentes principais foi possível reduzir o número de variáveis reduzindo a dimensão da matriz de covariância, o que demandaria menos recursos computacionais. O algoritmo proposto pelos autores faz uso de duas fases e seis passos. A primeira fase tem como objetivo encontrar o ponto de localização dos *outliers* e a segunda fase detecta a dispersão dos *outliers*. Por último, ele combina as duas fases gerando um valor para cada observação, o qual seria uma distância em relação a um valor de referência para determinar se a observação é um *outlier* ou não. Vale ressaltar que o autor fez uso da mediana e do desvio absoluto da mediana, para ponderar as variáveis em seguida ele usou o coeficiente de Kurtosis para identificar se há *outliers* ou não em cada variável. O método proposto foi comparado com o OGK, MCD e kurtosis. Para as simulações com dimensões (p) menor que o número de observações (n), o MCD obteve os melhores resultados. Para simulações em alta dimensão, o autor não apresentou comparações com os modelos anteriores. Esse método ficou conhecido como Principais componentes para *outliers* (PCOut).

Sajesh e Srinivasan (2012) abordaram a detecção de *outliers* para dados multivariados com distribuição normal de média e variâncias conhecidas, usando o método *comedian* (método criado por Falk,1997). Nesse artigo, os autores realizaram diversas simulações, considerando dados com 5, 10 e 20 variáveis e alterando também os valores da covariância e média. Por outro lado, o autores não alteraram o tamanho da amostra n, mantendo-a sempre igual a 100. O método *comedian*, conforme demonstrado por Sajesh e Srinivasan (2012), obteve os melhores resultados para detecção de *outliers* quando comparado aos métodos kurtosis, MCD e OGK. Para verificar os acertos, os autores realizaram as contagens das taxas de sucesso e das *taxas*

de falsa detecção (TFD). Ao final, os autores fizeram uma comparação alterando o tamanho das amostras e incluindo um método híbrido, mas não alteraram os valores das médias e covariâncias. Segundo Sajesh e Srinivasan (2012), quanto maiores as dimensões dos conjuntos de dados, mais eficiente é o método. Embora o autor tenha citado que o método pode ser usado para dados em alta dimensão, não foi apresentado nenhum resultado com esse cenário.

Ro et al. (2015) publicaram uma proposta de detecção de *outliers* para dados em altas dimensões, ou seja, $p > n$. O método proposto é baseado no MCD, porém em vez de procurar por um determinante mínimo, o método utiliza o produto mínimo da diagonal da matriz de covariância estimada (MDP - *minimum diagonal product*). Para os autores o novo método proposto, obteve resultados satisfatórios. Mas nas simulações apresentadas, em poucos casos simulados o método proposto foi superior aos métodos comparados. Quando o método era superior na taxa de falsa detecção ele era inferior nas taxas de falso negativo (*outliers* não identificados). Por meio das simulações apresentadas foi possível notar que o MDP consegue obter resultados melhores nas taxas de falso negativo em dados com alta dimensões, mas as taxas de falsa detecção ficaram muito acima dos demais métodos comparados.

Veloso e Cirillo (2016) apresentaram um método utilizando componentes principais. Na verdade os autores propuseram um método baseado no que foi proposto por Filzmoser, Maronna e Werner (2008). Enquanto o método mais antigo ponderava os dados usando MAD e a mediana, Veloso e Cirillo (2016), apresentaram uma nova forma de padronização da amostra usando o teste de qui-quadrado de Pearson e o teste de correção de Yates. Por meio das simulações Monte Carlo os autores realizaram testes de significância para encontrar a correção de padronização mais significativa na detecção de *outliers*. Os autores fizeram a comparação entre os métodos com amostras sem ponderação e com ponderação usando a correção de Pearson e correção de Yates. Nos casos simulados a correção de Pearson obteve resultados melhores de desempenho. Porém, à medida que p aumentava o método perdia eficiência.

Palma e Gallo (2016) apresentaram uma versão do *Comedian* diferenciada para detecção de *outliers* em dados de composição (CoDa). Dados de composição possuem características restritas e precisam de um tratamento diferenciado. Dentre essas características os autores mencionam que a soma dos dados são sempre iguais a uma constante e por isso não permite trabalho no espaço euclidiano, portanto, deve ser utilizado um sub-espaço chamado *simplex*. Outro detalhe, os *outliers* não são visíveis como observações extremas no espaço amostral.

Para solucionar esse problema os autores aplicaram uma transformação nos dados para que fosse possível aplicar os métodos da distância euclidiana.

Barbosa, Pereira e Oliveira (2018) propuseram um método para detecção de *outliers* fazendo uma análise de agrupamento k-médias com o objetivos de agrupar observações análogas, em seguida compara as distâncias do centróide de cada grupo com a mediana das observações. O método ficou conhecido como Método de Análise de Cluster (CAM - Cluster Analysis Method). Nessa ocasião o novo método foi comparado apenas com a distância de Mahalanobis. Em 2020 os autores publicaram um novo artigo comparado-o com os métodos clássicos MVE e MCD. Neste último teste, o método proposto obteve a melhor taxa de detecção falsa e maior acurácia (taxa total de acertos do modelo), enquanto o MVE obteve maior taxa de sucesso (BARBOSA; DUARTE; MARTINS, 2020). Nessa questão é importante que os métodos aplicados não identifiquem observações que não são *outliers* como *outliers*, mas o principal objetivo é que *outliers* sejam identificados.

3 METODOLOGIA

Para realizações dos procedimentos de detecção de *outliers*, foram realizadas simulações. Os *outliers* foram gerados por meio de populações normais multivariadas contaminadas. Em seguida, foram realizados os testes de detecção de *outliers* usando $N = 2000$ repetições para estimar as taxas de sucesso e de falsa detecção.

3.1 Geração da Normal Multivariada Contaminada

Uma população normal contaminada é uma mistura de distribuições normais multivariadas (BARBOSA; PEREIRA; OLIVEIRA, 2018). Para construção da simulação da normal multivariada contaminada, primeiro deve-se gerar um vetor $\boldsymbol{X}$ com função densidade de probabilidade:

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

em que $\boldsymbol{X} = [X_1, X_2, ..., X_p] \in R^p$ (vetor com distribuição normal multivariada contaminada), $(1-\delta)$ é a probabilidade dos dados gerados terem distribuição N($\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$), sendo que δ está contido entre [0,1] e δ é a probabilidade dos dados terem distribuição N($\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$), sendo ainda que $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$ são as matrizes de covariâncias positivas definidas e $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ são os vetores de médias.

Amostras aleatórias da distribuição normal contaminada foram facilmente obtidas com ajuda do *software* (R Core Team, 2021), fazendo uso do pacote *MASS* usando a função *mvrnorm*, para geração dos vetores da normal multivariada. O algoritmo gera dados usando a função *runif* para obter um valor aleatório uniforme entre 0 e 1. A seguir deve ser verificado se os valores simulados são maiores ou menores que δ. Caso seja maior, a função *mvrnorm* gera valores com distribuição N($\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$) e caso seja menor os dados terão distribuição N($\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2$).

Com relação à contaminação dos dados, considerou-se diversas situações, sendo que o nível de contaminação é definido por δ, uma vez que $100(1-\delta)\%$ é o percentual de dados sem contaminação e $100\delta\%$ o percentual contaminado. Além disso, pretendeu-se avaliar também quanto houve de influência da contaminação da média e da covariância na detecção de *outliers*. Para os dados sem contaminação considerou-se distribuição $N(\boldsymbol{0}, \boldsymbol{I})$, e para os dados contamina-

dos distribuição $N(\xi \boldsymbol{u}, \lambda \boldsymbol{I})$, sendo $\boldsymbol{u} = (1, ..., 1)^\top$, com tamanho $p \times 1$ e $\boldsymbol{I}$ uma matriz identidade com dimensão $p \times p$. Essas simulações são próximas do que tem sido abordados em trabalhos de detecção de *outliers* com por exemplo (ROCKE; WOODRUFF, 1996), Filzmoser, Maronna e Werner (2008), Sajesh e Srinivasan (2012) e (VELOSO; CIRILLO, 2016).

Um exemplo de matriz de covariância e vetor de médias para o caso sem contaminação é:

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_{p\times 1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{e} \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{I}_{p\times p} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Estes valores para a população contaminada pode ser escrita da seguinte forma:

$$\boldsymbol{\mu}_2 = \xi \boldsymbol{\mu}_{p\times 1} = \begin{bmatrix} \xi \\ \xi \\ \vdots \\ \xi \end{bmatrix} \qquad \text{e} \qquad \boldsymbol{\Sigma}_2 = \lambda \boldsymbol{I}_{p\times p} = \begin{bmatrix} \lambda & 0 & \dots & 0 \\ 0 & \lambda & \dots & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \lambda \end{bmatrix}.$$

Considerou-se também situações com correlações em que a estrutura de simetria composta foi utilizada, cuja correlação comum foi representada por ρ. A matriz de covariância para dados correlacionados considerada é dada por

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \rho & \ddots & \rho \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

Para se obter a matriz de covariância contaminada ($\boldsymbol{\Sigma}_2$) basta multiplicar a matriz $\boldsymbol{\Sigma}_1$ por λ, da mesma forma que foi feito com a matriz identidade.

3.2 Identificando *outlier*'s usando *Comedian*

Inicialmente, a partir da amostra sendo analisada é aplicado o método *Comedian* de onde se obtém-se uma matriz definida positiva $\boldsymbol{S(X)}$ como uma estimativa robusta. Assim, em seguida, é obtida a distância de Mahalanobis de cada observação para o centro amostral robusto,

cuja finalidade é a detecção de *outliers*, que nesse caso é computada por:

$$\boldsymbol{RD}(\boldsymbol{x}_i, \boldsymbol{m}) = rd_i = (\boldsymbol{x_i} - \boldsymbol{m})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{m}), \qquad i = 1, 2, ..., n, \tag{3.1}$$

em que $\boldsymbol{m}$ é o valor médio robusto.

O valor de corte para verificar até que ponto uma observação deve ser considerada como um *outliers* ou não foi obtido por

$$cv = \frac{1,4826[\chi^2_{0,95;p} med(rd_1, ..., rd_n)]}{\chi^2_{0,5;p}}, \tag{3.2}$$

em que 1,4826 é o fator de correção que torna *MAD* imparcial na distribuição normal (HUBERT; VEEKEN, 2008).

Quando qualquer $\boldsymbol{RD}(\boldsymbol{x}_i, \boldsymbol{m}) > cv$, a observação correspondente, $\boldsymbol{x}_i$, foi considerada um *outlier*. As estimativas *comedian* foram computadas no pacote `robustbase`, do programa R Core Team (2021), além de ter sido computada a distância de Mahalanobis da equação (3.1), também foi possível obter a matriz de covariância com todas as transformações necessárias. De posse dos rd_i's foi feita a comparação com o valor de corte (cv) apresentado na equação (3.2) para cada observação.

3.3 Identificando *outlier*'s usando MVE, MCD e OGK

Para identificar os *outliers* usando os métodos MVE, CMD e OGK foram utilizadas as funções `cov.mve`, *cov.mcd* ou `cov.mcd` e `CovOGK`, sendo que as duas primeiras funções estão no pacote `MASS` e as duas últimas no pacote `robustbase`.

Com as funções `cov.mve` e `cov.mcd` foi possível obter a matriz de covariância com determinantes mínimos e os pontos de centro para calcular a distância de Mahanalobis. O ponto de corte usado nesses dois casos foi $\sqrt{\chi^2_{0,975;p}}$.

A função `CovOGK`, além dos pontos de centros e a matriz de covariância, retorna também as distâncias de Mahalanobis computadas. O ponto de corte para esse método utilizou a equação (2.28) apresentada, com $\alpha = 0,975$.

3.4 Identificando *outlier*'s PCOut

Usando o método PCOut, foram obtidas duas distâncias para os dados, sendo uma para locação e uma para dispersão. No PCOut ambas as distâncias foram usadas para identificação dos *outliers*. Atribuiu-se "0" para *outliers* sempre que $d_i \geqslant c$, em que c é definido por

$$c = med(d_1, ..., d_n) + 2,5 * MAD(d_1, ..., d_n). \tag{3.3}$$

Foram computados os pesos

$$w_{1i} = \begin{cases} 0, & d_i \geqslant c, \\ \left(1 - \left(\frac{d_i - M}{c - M}\right)^2\right)^2 & M < d_i < c, \\ 1 & d_i \leqslant M \end{cases} \tag{3.4}$$

em que $i = 1, ..., n$, M é o $\frac{1}{3}$ quantil da distância $d_1, ..., d_n$, ou seja é o ponto mínimo em que a observação não foi classificada como *outlier*. Sempre que $d_i \leqslant M$ atribui-se "1". Ambas as distâncias encontradas foram aplicadas na equação de (3.4), a qual pode ser chamada de equação de pesos. Por fim foi utilizado uma última equação que reuniu os pesos da primeira distância com os da segunda (FILZMOSER; MARONNA; WERNER, 2008), dada por

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2}, \tag{3.5}$$

em que o valor de s é uma constante igual a $0,25$. Essa constante foi introduzida porque algumas observações não *outliers* podem ter recebido peso zero em um dos passos. Com $s \neq 0$, somente são considerados *outliers* as observações com 0 em ambas as fases. Por fim, as observações classificadas como *outliers* foram aquelas cuja $w_i < 0,25$ (FILZMOSER; MARONNA; WERNER, 2008). O algoritmo utilizado foi *PCout*, que pode ser encontrado no pacote `mvoutlier` do programa R Core Team (2021).

3.5 Taxas de sucesso (TS) e de taxa de falsa detecção (TFD)

Ao gerar uma população normal contaminada foi criado um vetor h (classificação real) que contém 1 nas posições onde não há *outliers* e 0 onde há *outliers*. Além disso, todos os

algoritmos dos métodos propostos tem um vetor de binário de pesos w, onde 0 indica *outliers* e 1 indica que a observação é não *outlier*.

Para obter a taxa de sucesso confrontou-se todas posições do vetor h, com as posições correspondentes do vetor w. Dessa forma, bastou realizar uma contagem para a condição "$h_i = 0$ e $w_i = 0$", sendo que o resultado dessa contagem foi armazenado em uma variável denotada Ta. A taxa de sucesso, TS, foi obtida usando a equação $TS = \frac{Ta}{n_{out}}$ em que n_{out} é o número total de *outliers* gerados e Ta é o total de *outliers* identificados corretamente.

A taxa de falsa detecção ou taxa de falsos positivos, foi obtida comparando novamente os vetores h e w. Mas, dessa vez foram contadas a seguinte condição "$h_i = 1$ e $w_i = 0$", em que o resultado dessa contagem também foi armazenado em uma variável $Ta4$. A equação para taxa de falsa detecção, TFD, usada foi $TFD = \frac{Ta4}{n - n_{out}}$, em que n é o número de observações geradas e n_{out} é o total de observações que são *outliers*, $Ta4$ é o total de observações que foram identificadas como *outliers*, mas que de fato não eram.

Além disso, vale ressaltar que cada configuração de simulação foi repetida duas mil vezes, portanto as taxas obtidas corresponderam às médias de todas as simulações.

3.6 Simulações

Antes de descrever as simulações, é necessário informar que os dados simulados é uma tentativa de aproximação da realidade. Além disso, dados reais podem ter características que não foram ou não possam ser representadas nas simulações realizadas.

As simulações foram divididas em cinco cenários. Sendo o primeiro cenário realizado simulações sem a presença de *outliers*, ou seja, $\delta = 0$. O objetivo desse cenário foi verificar como os métodos comparados reagiram a medida que foram incluídas mais variáveis. Os dados simulados possui distribuição $N(\mathbf{0}, \boldsymbol{I})$ e foram usados $p = 2$, 5, 10, 20 e 50. O número de observações para esse cenário foi de $n = 100$. Nesse primeiro cenário a TS é nula, o melhor método detre os comparados será aquele que obter a menor TFD, com já mencionado anteriormente o processo será repetido $N = 2000$ vezes para cada teste.

O segundo cenário tem o objetivo de verificar se os métodos são eficazes para detectar *outliers* de dispersão, ou seja, mantém-se a média igual tanto para os dados normais quanto para os contaminados ($\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$). As alterações na distribuição contaminada ocorre somente na covariância. Foram utilizados as seguintes taxas de contaminação 10%, 30% e 40% e λ igual a 5. Os valores de p (dimensões) foram considerados neste caso em 5, 10, 20 e 50, o número de

observações igual a 100. Dado que $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ e os dados normais têm matriz de covariância igual a $\boldsymbol{I}$, o valor de λ deve ser maior que 1. Caso $\lambda \leq 1$ não haverá *outliers*, porque os dados normais já estão entre "0" e "1" . Portanto, não faz sentido colocar a covariância na parte contaminada com λ menor ou igual a 1.

No terceiro cenário o objetivo foi verificar qual método consegue maior TS e a menor TFD. Foram utilizadas contaminações iguais a 10%, 20%, 30% e 40%, $\xi = 5$ e 10, $\lambda = 0{,}1$, 0,5 ,1 e 3 e $p = 5$, 10, 20, 50, 100 e 500. Dessa forma foi possível observar se há alterações nos resultados dos métodos a medida que foram feitas as mudanças na parte contaminada, alterando: taxa de contaminação, média, variância e número de variáveis. Para as dimensões iguais a 5, 10, 20, 50 e 100 foram usados $n = 100$ observações. Para $p = 500$ foi utilizado $n = 1000$.

O quarto cenário foram simulados dados contaminados com $p > n$. A taxas de contaminação foram iguais ao realizado na segunda etapa. O mesmo vale para os valores de ξ (5 e 10) e λ (0,1, 0,5 ,1 e 5). Porém como p deve ser maior que n, foram adotados $p = 50$, 100 e 500, com $n = 40$ nos dois primeiros casos e 200 para o último caso. Esse último cenário tem como objetivo verificar se há uma melhora no desempenho dos métodos ao aproximar o número de observações com o número de variáveis.

No último cenário foram realizadas simulações com dados contendo correlações, ρ. Foram realizadas simulações somente com casos extremos, ou seja, δ iguais a (0, 40%), λ iguais a (0,1, 5) e $\xi = 5$. Foram simulados dados com ρ = (0,1, 0,5, 0,9). Sendo 0,1 para baixa correlação, 0,5 para média e 0,9 para alta correlação. Para os casos de alta dimensão foi utilizado $n = 40$ e $p = 50$, por último, $n = 40$ e $p = 100$. Todas as configurações consideradas estão apresentadas na Tabela (3.1).

Tabela 3.1 – Tabela das configurações consideradas nas simulações realizadas nas 5 cenários.

Cenário	Taxa de contaminação δ	Nº de Variáveis p	ξ	λ	Nº de observações n
01	0	2, 5, 10 e 50	0	0	100
02	0,1, 0,2, 0,3 e 0,4	5, 10, 20 e 50	0,0	5,00	100
03	0,1, 0,2, 0,3 e 0,4	5, 10, 20 e 50	5, 10	0,1, 0,5, 1, 5	100
	0,1, 0,2, 0,3 e 0,4	100 e 500	5	0,1, 0,5, 1, 5	1000
04	0,1, 0,2, 0,3 e 0,4	50, 100 e 200	5	0,1, 0,5, 1 e 5	40
	0,1, 0,2, 0,3 e 0,4	200	5	0,1, 0,5, 1 e 5	100
05	0,1 e 0,4	50 e 100	0 e 5	0,1 e 5	40

Fonte: Do autor (2022).

4 RESULTADOS E DISCUSSÕES

Nas próximas subseções são apresentados os resultados das taxas de sucesso e taxas de falsa detecção, para os métodos MVE, MCD, OGK, PCout e *comedian*, nos 5 cenários de simulações. Importante lembrar que na geração dos dados foi utilizado a distribuição da normal multivariada contaminada. Portanto, espera-se que quanto maior for o valor de ξ e λ melhores devem ser os resultados de cada método.

4.1 Aplicação dos métodos em dados sem *outliers*

Na Tabela 4.1 são apresentadas as taxas de falsa detecção (TFD) nas simulações realizadas. Pode-se observar que o melhor método foi o *Comedian*, que apresentou as TFD's menores em todas simulações realizadas, com apenas 2%, para $p = 2$, 1% para p = 5, 10 e 0% para p=50. Outro detalhe, interessante para ser observado, são as mudanças nos resultados do MVE e MCD, percebe-se que o aumento de p provocou um aumento nas TFD's. Enquanto no *Comedian* esta relação entre p e a TFD foi inversamente proporcional. Os métodos OGK e PCOUT foram os piores para p = 2, 5 e 10. Para $p = 50$ os piores foram os métodos MCD e MVE. As taxas de falsa detecção foram bastante elevadas nesta última situação.

Tabela 4.1 – TFD's médios dos métodos de detecção de *outliers* em $N = 2000$ simulações na ausência de *outliers* considerando tamanho de amostra n = 100

MÉTODO	TFD p=2	TFD p=5	TFD p=10	TFD p=50
MVE	0,03	0,05	0,09	0,23
MCD	0,03	0,04	0,08	0,24
OGK	0,10	0,11	0,11	0,13
PCOut	0,12	0,13	0,13	0,10
COMEDIAN	0,02	0,01	0,01	0,00

Fonte: Do autor (2022).

Os resultados apresentados por Barbosa, Duarte e Martins (2020) para dados sem contaminação ficaram bem próximos dos resultados obtidos aqui, embora os autores tenham usado correlação nos dados. Por exemplo, para $p = 5$, os autores conseguiram uma TFD de 7% para MVE e 3% para MCD, enquanto aqui foi obtido 5% para MVE e 4% para MCD. Os autores usaram o método MVE e MCD para comparar com o método por eles desenvolvido.

4.2 Detecção de *outliers* em dados alterando apenas a variância

Na Tabela 4.2 são apresentados os resultados das simulações. A princípio, foram considerados os resultados sobre as TS's. Observando δ ficou evidente que ele possui uma relação inversa com a TS para todos métodos. O aumento da dimensão p também afetou as TS's, para os métodos *Comedian*, PCOut e OGK de maneira positiva e os métodos MVE e MCD de maneira negativa. Por outro lado, ao observar a TFD, ficou claro que o aumento em p implica em um aumento das TFD's nos métodos MVE e MCD, enquanto que no OGK houve um pequeno aumento. No método PCOut houve uma queda na TFD de 1% à medida que p aumentava e no *Comedian*, praticamente todas as TFD's foram zero exceto a primeira simulação com $p = 5$ e $\delta = 0,1$. Vale salientar que, as melhores TS's foram obtidas com o método OGK e PCOut, enquanto as melhores TFD foram obtidas com o *Comedian*. Para $p = 5$ e $\delta = 0,4$ OGK atingiu 68% de acerto sendo o melhor método seguindo pelo PCOUut com 62%, enquanto o *Comedian* obteve apenas 38%. Aumentando p para 20 e mantendo a mesma porcentagem de contaminação ($\delta = 0,40$), o *Comedian* atingiu 86%, PCOut 96% e o OGK 98%, ou seja, quanto maior a dimensão melhores serão os resultados dos métodos. Isso ficou mais evidente quando p foi elevado para 50.

Tabela 4.2 – TS's e TFD's dos métodos de detecção de *outliers* nas simulações com *outliers* de dispersão com n = 100 e λ = 5.

p	δ	MVE		MCD		OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
5	0,1	0,72	0,03	0,71	0,03	0,84	0,08	0,81	0,09	0,61	0,01
	0,3	0,60	0,01	0,59	0,01	0,75	0,03	0,71	0,03	0,48	0,00
	0,4	0,50	0,00	0,51	0,00	0,68	0,02	0,62	0,02	0,38	0,00
10	0,1	0,93	0,06	0,93	0,05	0,96	0,08	0,95	0,08	0,83	0,00
	0,3	0,85	0,02	0,88	0,02	0,93	0,03	0,89	0,03	0,72	0,00
	0,4	0,76	0,01	0,85	0,01	0,87	0,01	0,81	0,01	0,59	0,00
20	0,1	0,98	0,17	1,00	0,19	1,00	0,09	0,99	0,08	0,97	0,00
	0,3	0,93	0,05	0,99	0,08	0,99	0,04	0,99	0,02	0,93	0,00
	0,4	0,85	0,02	0,94	0,02	0,98	0,01	0,96	0,01	0,86	0,00
50	0,4	0,56	0,05	0,56	0,04	1,00	0,02	1,00	0,00	1,00	0,00

Fonte: Do autor (2022).

A princípio os métodos MVE e MCD também tiveram ganho com o aumento da dimensão p, mas quando p foi elevado a 50 os resultados foram ruins chegando a 56% de TS

(TABELA 4.2). Sobre a TFD o *Comedian* mostrou-se melhor em todas as simulações apresentadas na Tabela 4.2, enquanto o PCOut foi o pior para $p = 5$, ao lado do OGK. Para $p = 10$ eles continuaram sendo os piores apresentando TFD's de 8%, 3% e 1%, para contaminações de 10%, 30% e 40%, respectivamente. Para p = 20 e 50, o MVE e o MCD apresentaram as piores TFD's e TS's nos dados simulados.

Por fim é interessante observar que a taxa de contaminação menor, (nesse caso de 10%), apresenta índices mais elevados de TFD. Isso porque em torno de 90% dos dados tem distribuição normal $N(0,I)$. Assim como foi observado na Tabela 4.1 para os métodos MVE e MCD, observou-se também na Tabela 4.2 que ops métodos tendem a ter um aumento na TFD a medida que o a dimensão p aumenta. Em contra partida, $\delta = 0,1$, foi o índice que todos os métodos obtiveram as melhores TS's. No entanto, quanto maior o p, menor será essa diferença na TS em relação às mudanças nos valores de δ. Em resumo, pode-se concluir que a influência da taxa real de *outliers* δ nas TS's está ligada diretamente à dimensão p dos dados. Quanto maior p, menor será a influência de δ nas TS's dos métodos PCOut, OGK e *Comedian*.

Em comparação com a pesquisa apresentada por Filzmoser, Maronna e Werner (2008), quando ξ = 0 e p = 10, além de se adotar λ = 5, verificou-se que os métodos OGK e MCD superam o PCOut, tanto em TS como em TFD. Nos resultados obtidos aqui, o PCOut obteve TS superior à do método MCD, porém inferior à do método OGK. Na verdade os resultados das duas pesquisa ficaram próximos. No trabalho dos autores, as taxas de TS do PCOut, OGK e MCD foram $91,16\%$, $93,31\%$ e $92,35\%$, enquanto aqui os resultados foram 95%, 96% e 93%, respectivamente. Estes autores não utilizaram o *Comedian* e o MVE nas comparações entre os métodos.

4.3 Detecção de *outliers* de locação e dispersão

Neste cenário, as simulações foram mais extensas que as anteriores, pois envolveram alterações em μ_2, ou seja, $\xi \neq 0$. Para esse novo cenário ξ assumiu valores de 5 e 10, enquanto λ assumiu valores de 0,10, 0,50, 1 e 5. As dimensões (valores de p) tiveram a seguinte sequência 5, 10, 20, 50, 100, 200 e 500 e δ foi sendo alterado com 10%, 20%, 30% e 40%. Para casos da dimensão p até 50 foi utilizado n= 100 e para casos de $p \geq 100$, n foi alterado para 1000. O objetivo dessa nova etapa foi verificar qual método foi o melhor, além de verificar também como as alterações na distribuição normal contaminada influenciam as TS's e as TFD's.

A primeira simulação dessa nova etapa foi com $\xi = 5$, $p = 5$ e $n = 100$, alterando λ e δ com os valores descritos acima. Mais uma vez cada simulação foi repetida 2000 vezes. Na Tabela 4.3, são apresentados os primeiros resultados obtidos. Em comparação com as taxas anteriores quando era mantido $\xi = 0$, as TFD's mais altas eram com $\delta = 0,10$. Para o caso de $\xi = 5$, mantendo a taxa de contaminação igual a 10%, as maiores TFD's foram obtidas com os maiores λ's. Na primeira linha da Tabela 4.3, todos os métodos conseguiram 100% de TS. Mas, observando as TFD´s o melhor método foi o *comedian* com apenas 1% de TFD, enquanto os métodos PCOut e OGK foram os piores com TFD de 7% e 8% respectivamente. Porém ao aumentar a taxa de contaminação mantendo $\lambda = 0,10$, a TS cai, principalmente nos métodos MVE, OGK e MCD chegando a 1%, 1% e 0%, respectivamente, com $\delta = 0,40$. Com esse δ, o PCOut se mostrou o melhor método com 52% de TS. Embora o PCOut tenha sido o melhor método com contaminação de 40% dados considerando $\lambda = 0,10$, a TFD chegou a 11%. Para as TFD's o *Comedian* obteve os melhores resultados.

Tabela 4.3 – TS's e TFD's dos métodos de detecção de *outliers* nas simulações com *outliers* para n = 100, $\xi = 5$ e $p = 5$.

λ	δ	MVE		MCD		OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
0,10	0,10	1,00	0,03	1,00	0,02	1,00	0,08	1,00	0,07	1,00	0,01
	0,20	0,91	0,03	0,80	0,05	0,90	0,07	1,00	0,03	0,97	0,00
	0,30	0,28	0,18	0,10	0,28	0,20	0,15	0,97	0,01	0,81	0,00
	0,40	0,01	0,39	0,00	0,57	0,01	0,41	0,52	0,11	0,17	0,02
0,50	0,10	1,00	0,03	1,00	0,02	1,00	0,08	1,00	0,07	1,00	0,01
	0,20	1,00	0,02	1,00	0,01	0,96	0,06	1,00	0,03	1,00	0,00
	0,30	0,90	0,02	0,97	0,40	0,06	1,00	1,00	0,01	0,95	0,00
	0,40	0,33	0,08	0,58	0,05	0,04	0,14	0,85	0,01	0,40	0,00
1	0,10	1,00	0,03	1,00	0,02	1,00	0,08	1,00	0,07	1,00	0,01
	0,20	1,00	0,02	1,00	0,01	0,99	0,06	1,00	0,03	1,00	0,00
	0,30	1,00	0,01	1,00	0,01	0,63	0,04	1,00	0,01	1,00	0,01
	0,40	0,79	0,01	0,96	0,01	0,15	0,06	0,89	0,00	0,56	0,00
5	0,40	0,97	0,00	1,00	0,00	0,83	0,01	0,93	0,00	0,95	0,00

Fonte: Do autor (2022).

Pode-se verificar que o aumento de λ tem impacto positivo nas TS's e TFD's para todos os métodos. O método OGK foi o que mais obteve queda na TS com $\lambda < 1$ e $\delta \geqslant 0,30$, enquanto o PCOut obteve as melhores TS's. Por fim, com $\lambda = 5$ e $\delta = 0,40$, todos os métodos ficaram

com TS acima de 90%, exceto OGK que atingiu 83%, o MCD foi o melhor método com 100% de acerto e zero de TFD, seguido pelo MVE com 97% de TS e zero de TFD (TABELA 4.3).

Na próxima configuração alterou-se ξ para 10 e manteve-se $p = 5$. As demais quantidades foram mantidas como no caso anterior. Como a média da parte contaminada está maior,é esperado que os resultados dos métodos sejam melhores. Na Tabela 4.4 mostrou que todos os métodos tiveram excelentes resultados para TS. Analisando as TFD's o método *Comedian* foi o melhor método para qualquer $\delta \leq 0,30$ com TFD iguais a zero e 1%. Nessa mesma faixa de δ, as maiores TFD's, ocorreram para os métodos OGK e PCOut. Para contaminações de 40%, o melhor método foi o PCOut quando $\lambda \leq 1$. Com $\lambda = 5$, todos os métodos atingiram 100% de TS e zero de TFD, com exceção do MVE que atingiu 96% de TS e OGK que atingiu 91%. Assim como aconteceu nas simulações com ξ = 5, a medida que λ aumentava os métodos apresentaram melhores resultados para TS e TFD, independentemente da taxa de contaminação.

Tabela 4.4 – TS's e TFD's dos métodos de detecção de *outliers* nas simulações com *outliers* com n = 100, $\xi = 10$ e $p = 5$.

λ	δ	MVE		MCD		OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
0,10	0,10	1,00	0,03	1,00	0,03	1,00	0,08	1,00	0,07	1,00	0,01
	0,20	1,00	0,02	0,99	0,02	1,00	0,07	1,00	0,03	1,00	0,01
	0,30	0,80	0,07	0,53	0,19	0,77	0,07	1,00	0,01	1,00	0,00
	0,40	0,15	0,35	0,03	0,56	0,13	0,28	0,97	0,00	0,78	0,00
0,5	0,10	1,00	0,03	1,00	0,03	1,00	0,08	1,00	0,07	1,00	0,01
	0,20	1,00	0,02	1,00	0,01	1,00	0,06	1,00	0,03	1,00	0,00
	0,30	1,00	0,01	1,00	0,01	0,90	0,05	1,00	0,01	1,00	0,00
	0,40	0,82	0,03	0,93	0,01	0,29	0,07	1,00	0,00	0,92	0,00
1	0,10	1,00	0,03	1,00	0,02	1,00	0,08	1,00	0,07	1,00	0,01
	0,20	1,00	0,02	1,00	0,01	1,00	0,06	1,00	0,03	1,00	0,00
	0,30	1,00	0,01	1,00	0,01	0,96	0,04	1,00	0,01	1,00	0,00
	0,40	0,82	0,03	0,94	0,01	0,27	0,07	1,00	0,00	0,92	0,00
5	0,40	0,96	0,00	1,00	0,00	0,91	0,01	1,00	0,00	1,00	0,00

Fonte: Do autor (2022).

Serão ressaltados a seguir alguns pontos que podem ser esperado para as próximas configurações. Nas simulações sem *outliers* ficou claro que com o aumento de p, os métodos *Comedian* e PCOut tendem a ter resultados melhores e os métodos MVE e MCD tiveram um aumento nas TFD. Nas simulações em que se manteve ξ = 0, percebeu-se que à medida que

p aumentava, independentemente das taxas de contaminação (δ), os métodos tenderam a ter resultados melhores. Por fim, nas simulações em que a parte contaminada difere da parte não contaminada tanto em média quanto em covariância ficou claro que quanto maior λ, maiores as TS's e menores as TFD's, independentemente dos valores de δ (isso também é válido para ξ). Com base nesses resultados mencionados, nas próximas configurações das simulações espera-se TS's e TFD's melhores.

Considerando $p = 10$ e $\xi = 5$, com as demais quantidades mantidas como anteriormente, verificou-se na Tabela 4.5 que para $\delta \leq 0,20$, as melhores TS's foram obtidas com o *Comedian* e com o PCOut. Com $\delta \geqslant 0,30$ o PCOut passa a ser o melhor método para qualquer λ. Observando $\lambda = 0,1$, percebe-se que para $\delta > 0,10$, o MVE, o MCD e o OGK perderam eficiência e a TS ficou mais baixa em cada aumento de δ. O *Comedian* passou a ser influenciado quando $\delta = 0,30$, em que apresentou TS de 88%. Já o PCOut atingiu 99% de TS e para $\delta = 0,40$, o *Comedian* ficou com apenas 22% de TS, enquanto o PCOut apresentou desempenho de 76%. Por outro lado, as melhores TFD's foram novamente obtidas com o *Comedian*. Com o aumento de λ, os métodos foram melhorando seus resultados. O MVE, por exemplo, de 3% de TS com $\lambda = 0,1$ passou a atingir, com $\lambda = 0,5$, 37%, embora ainda muito aquém do desejado. Para reafirmar o quanto a covariância influência nos métodos de detecção, foi realizado um teste com $\lambda = 10$ e δ = 0,4. Todos os métodos tiveram TS's maiores ou iguais a 99%.

Considerando p = 10 e ξ = 10, espera-se mais uma vez, com a média maior, resultados melhores, pois a distância entre os *outliers* e os não *outliers* é muito maior. Na Tabela 4.6 as TS's e TFD's são apresentados para esse caso. O *Comedian* superou os demais métodos, com $\lambda < 0,40$, independentemente dos valores de δ, atingindo 100% de TS e zero para as taxas de TFD. O pior método foi do MCD, que com $\delta > 0,10$, a TS reduziu e a TFD aumentou. Considerando $\delta = 0,40$ e $\lambda = 0,10$, o melhor método foi o PCOut com 100% de TS e zero de TFD, o segundo melhor método foi o *Comedian* que atingiu 85% de TS. Assim como observado anteriormente, à medida que λ aumenta todos os métodos apresentaram melhorias no desempenho. Para reafirmar essa condição foi realizado mais 2 simulações mantendo δ = 0,40, porém com λ = 5 e 10. Para o primeiro caso, os dois piores métodos foram o MVE e OGK que atingiram 96% e 98% de TS respectivamente, embora este desempenho seja muito satisfatório. Com λ = 10, o pior método foi o MVE que atingiu 99% de TS. Interessante observar mais uma vez que para contaminações iguais ou menores a 20%, as TFD's dos métodos MVE e MCD subiram à medida que λ decresceu.

Tabela 4.5 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 100, $\xi = 5$ e $p = 10$.

λ	δ	MVE		MCD		OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
0,1	0,1	1,00	0,07	0,95	0,07	1,00	0,09	1,00	0,07	1,00	0,00
	0,2	0,55	0,16	0,16	0,34	0,92	0,07	1,00	0,03	1,00	0,00
	0,3	0,03	0,39	0,00	0,59	0,19	0,20	0,99	0,01	0,88	0,00
	0,4	0,00	0,57	0,00	0,75	0,00	0,46	0,76	0,07	0,22	0,02
0,5	0,1	1,00	0,06	1,00	0,05	1,00	0,09	1,00	0,06	1,00	0,00
	0,2	0,94	0,05	0,97	0,04	0,99	0,07	1,00	0,03	1,00	0,00
	0,3	0,37	0,12	0,48	0,10	0,50	0,07	1,00	0,01	0,97	0,00
	0,4	0,03	0,22	0,04	0,20	0,04	0,19	0,96	0,00	0,47	0,00
1	0,1	1,00	0,06	1,00	0,05	1,00	0,09	1,00	0,07	1,00	0,00
	0,2	1,00	0,04	1,00	0,03	1,00	0,06	1,00	0,03	1,00	0,00
	0,3	0,92	0,03	0,97	0,02	0,71	0,04	1,00	0,01	0,99	0,00
	0,4	0,44	0,05	0,64	0,04	0,20	0,06	0,98	0,02	0,61	0,03
5	0,4	0,95	0,01	1,00	0,01	0,96	0,01	0,99	0,00	0,95	0,00
10	0,4	0,99	0,01	1,00	0,01	1,00	0,01	1,00	0,00	0,99	0,00

Fonte: Do autor (2022).

Como demonstrado, os métodos tenderam a apresentar resultados melhores considerando $\xi = 10$, portanto, foram evitadas simulações com $\xi = 10$, com a finalidade de exigir mais dos métodos de detecção. Dessa maneira, para as próximas simulações foram considerado apenas ξ = 5 e, além disso, aumentou-se a dimensão p. A próxima simulação apresentada na Tabela 4.7 contém p = 20 e as demais quantidades permaneceram as mesmas usadas nas simulações anteriores. Os efeitos já mencionados continuam com o mesmo padrão, sendo que o MVE e o MCD, quando $\delta < 0,30$, apresentaram TFD alta. O *Comedian* é o melhor método quando $\delta < 0,30$ para qualquer λ e PCOut o melhor para $\delta \geq 0,30$. Quanto maior λ melhores são os desempenhos dos métodos para todos os métodos tanto da TS quanto da TFD.

Foram apresentadas simulações com p = 5, 10 e 20 e ficou identificado que quanto mais as médias e as covariâncias da parte contaminada se distanciam dos dados normais melhores são os resultados de detecção (o que é realmente era esperado em teoria). Na Tabela 4.8 são apresentados os resultados considerando alterações em p mantendo λ = 0,1. O objetivo dessa análise foi identificar os efeitos do aumento da dimensão p nos métodos de detecção. O primeiro ponto importante observado foi que para até p = 5 com δ = 0,10, o MVE e MCD conseguiram

Tabela 4.6 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 100, $\xi = 10$ e $p = 10$.

λ	δ	MVE		MCD		OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
0,10	0,10	1,00	0,06	1,00	0,05	1,00	0,09	1,00	0,07	1,00	0,00
	0,20	0,92	0,07	0,48	0,25	1,00	0,07	1,00	0,03	1,00	0,00
	0,30	0,23	0,33	0,01	0,59	0,76	0,08	1,00	0,01	1,00	0,00
	0,40	0,00	0,57	0,00	0,74	0,08	0,39	1,00	0,00	0,85	0,00
0,50	0,10	1,00	0,06	1,00	0,05	1,00	0,09	1,00	0,07	1,00	0,00
	0,20	1,00	0,04	1,00	0,03	1,00	0,07	1,00	0,03	1,00	0,00
	0,30	0,82	0,06	0,88	0,05	0,92	0,05	1,00	0,01	1,00	0,00
	0,40	0,21	0,19	0,27	0,17	0,29	0,11	1,00	0,02	0,96	0,03
1	0,10	1,00	0,06	1,00	0,05	1,00	0,09	1,00	0,07	1,00	0,00
	0,20	1,00	0,04	1,00	0,04	1,00	0,07	1,00	0,03	1,00	0,00
	0,30	0,98	0,02	0,98	0,02	0,97	0,04	1,00	0,01	1,00	0,00
	0,40	0,65	0,04	0,69	0,03	0,53	0,04	1,00	0,00	0,98	0,00
5	0,40	0,96	0,01	1,00	0,01	0,98	0,02	1,00	0,00	1,00	0,00
10	0,40	0,99	0,01	1,00	0,01	1,00	0,00	1,00	0,00	1,00	0,00

Fonte: Do autor (2022).

Tabela 4.7 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 100, $\xi = 5$ e $p = 20$.

λ	δ	MVE		MCD		OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
0,1	0,1	0,77	0,23	0,40	0,31	1,00	0,10	1,00	0,06	1,00	0,00
	0,2	0,04	0,41	0,00	0,47	0,92	0,08	1,00	0,02	1,00	0,00
	0,3	0,00	0,52	0,00	0,56	0,22	0,22	1,00	0,01	0,92	0,00
	0,4	0,00	0,66	0,00	0,66	0,00	0,46	0,92	0,06	0,31	0,02
0,5	0,1	0,96	0,18	0,97	0,19	1,00	0,09	1,00	0,06	1,00	0,00
	0,2	0,33	0,27	0,33	0,29	0,99	0,07	1,00	0,02	1,00	0,00
	0,3	0,05	0,37	0,05	0,41	0,58	0,08	1,00	0,01	1,00	0,00
	0,4	0,05	0,43	0,05	0,49	0,04	0,25	1,00	0,00	0,60	0,00
1	0,1	1,00	0,18	1,00	0,18	1,00	0,09	1,00	0,06	1,00	0,00
	0,2	0,83	0,14	0,91	0,13	1,00	0,07	1,00	0,03	1,00	0,00
	0,3	0,38	0,20	0,44	0,19	0,84	0,04	1,00	0,01	1,00	0,00
	0,4	0,26	0,23	0,26	0,23	0,27	0,07	0,97	0,02	0,95	0,02

Fonte: Do autor (2022).

100% de TS com TFD abaixo de 4%. Mas com p =10 considerando o mesmo δ, o MCD atingiu 95% de TS, enquanto o MVE manteve os 100%, mas as TFD subiram para 7%. Para p = 20, o MCD atingiu apenas 40% de TS e o MVE atingiu 77%, e as TFD subiram para 31% e 23%, respectivamente. Observando $\delta \geq 0,20$ nos métodos MCD e MVE, com o aumento de p o resultados ficaram com TS's mais baixas e TFD's mais altas. Fazendo a mesma análise para o método OGK, notou-se que as TS's e TFD's ficaram semelhantes, independente da dimensão p. Com o aumento da contaminação, o método apresentou TS mais baixas e TFD mais altas, lembrando que λ = 0,10 está fixo. O método PCOut obteve excelentes resultados para todas as faixas de contaminação, além disso a TS tende a melhorar a medida que p aumenta e a TFD manteve-se igual, para $\delta < 0,30$ e caiu 1 ponto percentual para p = 20. As melhores TS´s foram obtidas com o PCOut e com o *comedian*, mas vale ressaltar que para contaminações iguais a 40% o *comedian* passa a obter TS's baixa enquanto o PCOut ainda consegue obter taxas maiores. As TFD's do *Comedian* foram as melhores em todos as simulações e à medida que p aumentava elas se mantiveram iguais e nenhuma delas acima de 2%. Em resumo, enquanto o aumento de p diminuiu a eficácia do MVE e MCD, melhorou os resultados do PCOut e *Comedian*. O método OGK, aparentemente, não sofreu nenhum efeito expressivo.

Tabela 4.8 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 100, $\xi = 5$, $\lambda = 0,1$.

p	δ	MVE		MCD		OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
5	0,1	1,00	0,03	1,00	0,02	1,00	0,08	1,00	0,07	1,00	0,01
10	0,1	1,00	0,07	0,95	0,07	1,00	0,09	1,00	0,07	1,00	0,00
20	0,1	0,77	0,23	0,40	0,31	1,00	0,10	1,00	0,06	1,00	0,00
5	0,2	0,91	0,03	0,80	0,05	0,90	0,07	1,00	0,03	0,97	0,00
10	0,2	0,55	0,16	0,16	0,34	0,92	0,07	1,00	0,03	1,00	0,00
20	0,2	0,04	0,41	0,00	0,47	0,92	0,08	1,00	0,02	1,00	0,00
5	0,3	0,28	0,18	0,10	0,28	0,20	0,15	0,97	0,01	0,81	0,00
10	0,3	0,03	0,39	0,00	0,59	0,19	0,20	0,99	0,01	0,88	0,00
20	0,3	0,00	0,52	0,00	0,56	0,22	0,22	1,00	0,01	0,92	0,00
5	0,4	0,01	0,39	0,00	0,57	0,01	0,41	0,52	0,11	0,17	0,02
10	0,4	0,00	0,57	0,00	0,75	0,00	0,46	0,76	0,07	0,22	0,02
20	0,4	0,00	0,66	0,00	0,66	0,00	0,46	0,92	0,06	0,31	0,02

Fonte: Do autor (2022).

Os padrões de respostas observados no parágrafo anterior também foram observados nos resultados apresentados por Sajesh e Srinivasan (2012) e por Cabana, Lillo e Laniado (2021). Embora os resultados do presente trabalho não seja exatamente os mesmos, para ambos os trabalhos, eles se aproximaram e os padrões também se confirmaram. Os métodos usados pelos autores foram o *Comedian*, o OGK e o MCD. Kunjunni e Abraham (2020) utilizaram os mesmos resultados apresentados por Sajesh e Srinivasan (2012) para verificar a eficiência relativa do novo método proposto por eles de detecção de *outliers*.

De acordo com os resultados das simulações, à medida que p aumentava, os métodos MVE e MCD pioravam, então para $p \geqslant 50$ eles foram descartados neste novo cenário de simulação. Além disso, com as simulações apresentadas nas a tabelas anteriores já é possível ter uma ideia do que esperar com o aumento de p. Na Tabela 4.9 são apresentados os resultados das simulações feita com $p = 50$, $n = 100$ e $\xi = 5$, os valores adotados de λ e δ foram mantidos. Mais uma vez o *Comedian* mostrou-se melhor para contaminação menor que 30%. Com a δ maior ou igual à 30% o melhor método foi o PCOut. O aumento de λ resultou em melhores taxas para todos os métodos. Dentre *Comedian*, PCOut e OGK, o OGK obteve pior desempenho.

Tabela 4.9 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 100, $\xi = 5$ e $p = 50$.

λ	δ	OGK		PCOut		COMEDIAN	
		TS	TFD	TS	TFD	TS	TFD
0,1	0,1	1,00	0,11	1,00	0,04	1,00	0,00
	0,2	0,95	0,09	1,00	0,02	1,00	0,00
	0,3	0,29	0,23	1,00	0,00	0,99	0,00
	0,4	0,00	0,50	0,97	0,05	0,38	0,01
0,5	0,1	1,00	0,11	1,00	0,04	1,00	0,00
	0,3	0,72	0,08	1,00	0,00	1,00	0,00
	0,4	0,09	0,33	0,97	0,03	0,85	0,04
1	0,1	1,00	0,11	1,00	0,05	1,00	0,00
	0,3	0,94	0,05	1,00	0,00	1,00	0,00
	0,4	0,42	0,07	0,98	0,02	0,95	0,01
5	0,4	0,99	0,02	0,99	0,00	1,00	0,00

Fonte: Do autor (2022).

Para as simulações com $p = 100$, foi descartado o método de detecção OGK. Além disso, o número de observações n foi elevado para 1000, sendo que as demais quantidades foram mantidas semelhantes as simulações anteriores. Nota-se através da Tabela 4.10 que com o aumento

de *n* houve uma queda na TS do *Comedian* para contaminação de 40%, considerando $\lambda = 0,10$. O método identificou apenas 2% dos *outliers* enquanto o PCOut identificou 100% dos *outliers*. Por outro lado, as TFD's ficaram em zero no *Comedian*, de qualquer maneira os resultados do *Comedian* com contaminação de 40% dos dados quando $\lambda = 0,1$ sempre ficaram baixos, embora quando *n* estava mantido em 100 observações ele apresentava melhoras com o aumento de *p*. Para as demais simulações os resultados seguiram os mesmos padrões anteriores, ou seja, λ maior, melhores as TS's e TFD's. Com $\delta \leqslant 0,20$, o PCOut apresentou TFD's entre 2% a 6%, enquanto o *Comedian* obteve zero. Portanto, a mesma análise se aplicou nessa simulação, para $\delta \leqslant 0,20$, o *Comedian* foi melhor porque atingiu 100% de TS e zero de TFD. Para simulações com δ = 30% os resultados de ambos os métodos ficaram muito próximos e para δ = 40% o PCOut foi o melhor método.

Tabela 4.10 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com *n* = 1000, $\xi = 5$ e $p = 100$

λ	δ	PCOut		COMEDIAN	
		TS	TFD	TS	TFD
0,1	0,1	1,00	0,04	1,00	0,00
	0,2	1,00	0,02	1,00	0,00
	0,3	1,00	0,00	0,99	0,00
	0,4	1,00	0,01	0,02	0,00
0,5	0,1	1,00	0,06	1,00	0,00
	0,2	1,00	0,02	1,00	0,00
	0,3	1,00	0,01	1,00	0,00
	0,4	1,00	0,00	0,96	0,00
1	0,1	1,00	0,06	1,00	0,00
	0,2	1,00	0,02	1,00	0,00
	0,3	1,00	0,01	1,00	0,00
	0,4	1,00	0,00	1,00	0,00
5	0,4	1,00	0,00	1,00	0,00

Fonte: Do autor (2022).

O próximo cenário considerou *p* = 500, ξ = 5, *n* = 1000, λ = 0,1 e 5. Para esse caso, o número de interações foi apenas 100 e considerou-se δ de 40% com λ = 0,1 e também δ de 10, 30 e 40% com λ = 5. Na Tabela 4.11 são apresentadas as taxas para essas simulações. O *Comedian* mais uma vez apresentou TS baixa para λ =0,10 com δ = 0,4. O PCOut conseguiu 100% de TS nessa condição. Para a simulação considerando $\lambda = 5$, o *Comedian* foi o melhor

método porque obteve TFD's iguais a zero. Vale ressaltar que para δ = 0,40 e λ = 5, os dois métodos atingiram 100% de TS e zero de TFD.

Tabela 4.11 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 1000, $\xi = 5$ e $p = 500$.

δ	λ	PCOut		COMEDIAN	
		TS	TFD	TS	TFD
0,4	0,1	1,00	0,00	0,04	0,00
0,3	5	1,00	0,01	1,00	0,00
0,1	5	1,00	0,04	1,00	0,00
0,4	5	1,00	0,00	1,00	0,00

Fonte: Do autor (2022).

Nas pesquisas apresentadas por Sajesh e Srinivasan (2012) e por Cabana, Lillo e Laniado (2021) taxas de contaminação até 30% foram utilizadas. Os resultados obtidos pelos autores são semelhantes aos resultados obtidos nessa pesquisa. Embora, para esse cenário de *outliers* de localização e dispersão os resultados demonstrados aqui foram levemente superiores, principalmente no MCD e *Comedian*. Considerando, por exemplo, λ = 0,1, $p = 5$ e com taxa de contaminação de 30%, os resultados dos autores Cabana, Lillo e Laniado (2021) conseguiram atingir uma TS de 57%, enquanto nesse trabalho foi obtido 81%. Os métodos em comum foram o *Comedian*, MCD e OGK. Os métodos PCOut e OGK usados por (FILZMOSER; MARONNA; WERNER, 2008) também apresentaram resultados inferiores aos demonstrados aqui, mas os autores usaram dados correlacionados iguais, ρ, a 0,5. De qualquer forma para essas mesmas quantidades fixadas no cenário, os autores conseguiram apenas $32,73\%$ de TS, enquanto aqui obteve-se 100%. Vale a pena ressaltar que os autores usaram $n = 1000$ e 500 repetições.

4.4 Detecção de *ouliters* com dimensão p maior que observações n

O próximo passo e último das simulações foi considerar casos em que $n > p$. Essa situação só é possível de ser avaliada utilizando os métodos *Comedian* e PCOut. A primeira simulação com essa nova condição foi realizada com p = 50 e n = 40. As demais quantidades foram mantidas com fixação de ξ = 5, λ = 0,1, 0,5 e 1 e δ = 0,1, 0,3 e 0,4. Na Tabela 4.12 são apresentados os resultados das TS's e TFD's. A princípio toda a dinâmica encontrada nas simulações anteriores se repetem para esse caso. O ponto que mais chamou atenção foi o aumento da TFD para as simulações considerando $\delta = 0,4$.

Tabela 4.12 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 40, $\xi = 5$ e $p = 50$.

λ	δ	PCOut		COMEDIAN	
		TS	TFD	TS	TFD
0,1	0,1	0,99	0,05	0,99	0,00
	0,3	0,99	0,02	0,96	0,02
	0,4	0,81	0,16	0,61	0,24
0,5	0,1	0,99	0,05	0,99	0,00
	0,3	1,00	0,01	0,98	0,00
	0,4	0,87	0,10	0,76	0,14
1	0,1	0,99	0,05	0,99	0,00
	0,3	0,99	0,01	0,99	0,00
	0,4	0,89	0,08	0,88	0,07

Fonte: Do autor (2022).

A próxima simulação foi realizada com p = 100, sendo que as demais quantidades foram mantidas como no caso anterior. Os resultados ficaram bem próximos aos apresentados na Tabela 4.12. Houve apenas uma queda de 1% nas TFD's dos dois métodos e um aumento de 1% nas TS's para algumas configurações tanto do *Comedian* quanto do PCOut. O *Comedian* continuou sendo o melhor para os casos de $\delta \leq 0,20$ para qualquer λ e o PCOut o melhor método para os casos de em que $\delta \geq 0,30$, para $\lambda \leq 0,5$. Outro detalhe interessante que pode ser observado, o *Comedian* tende a ter resultados melhores que o PCOut quando λ é maior. Nessa simulação por exemplo, (isso também pode ser conferido nas Tabelas 4.3 e 4.9) para λ igual a 5 e δ = 0,40, o *Comedian* conseguiu 98% de TS e zero de TFD, enquanto o PCOut atingiu 91% de TS e 2% de TFD, sendo que nas demais tabelas mencionadas isso também ficou evidente.

Na Tabela 4.14 são apresentados os dados das simulações realizadas com p = 200 e as demais quantidades foram mantidos iguais aos usados nas simulações anteriores. Os mesmos efeito encontrados nas outras simulações se repetiram. O *Comedian* é o melhor método para $\delta = 0,20$ independentemente de λ. Para $\delta \geqslant 0,30$, o melhor método foi o PCOut, porém até $\lambda = 1$. Para o teste de $\delta = 0,4$ e $\lambda = 5$, ambos os métodos atingiram 100% de TS e zero de TFD.

Por fim, o último cenário de simulação considerou n = 100, p = 200, sendo que as demais quantidades foram mantidas iguais aos valores já utilizados anteriormente. O objetivo foi verificar se com houve influência nas TS's e TFD's com o aumento de observações n. Com n maior, a TFD diminuiu em todos os métodos. O *comedian* se destaca mais uma vez por obter

Tabela 4.13 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 40, $\xi = 5$ e $p = 100$.

λ	δ	PCOut		COMEDIAN	
		TS	TFD	TS	TFD
0,1	0,1	0,99	0,04	0,99	0,00
	0,2	1,00	0,02	1,00	0,00
	0,3	0,99	0,02	0,95	0,03
	0,4	0,84	0,15	0,61	0,23
0,5	0,1	0,99	0,04	0,99	0,00
	0,3	0,99	0,01	0,99	0,01
	0,4	0,88	0,09	0,78	0,12
1	0,1	0,99	0,04	0,99	0,00
	0,2	1,00	0,02	1,00	0,00
	0,3	0,99	0,01	0,99	0,00
	0,4	0,87	0,09	0,86	0,08
5	0,4	0,91	0,02	0,98	0,00

Fonte: Do autor (2022).

Tabela 4.14 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 40, $\xi = 5$ e $p = 200$.

λ	δ	PCOut		COMEDIAN	
		TS	TFD	TS	TFD
0,1	0,1	0,97	0,04	0,97	0,00
	0,3	1,00	0,01	0,95	0,01
	0,4	0,87	0,12	0,64	0,15
0,5	0,1	0,99	0,05	0,99	0,00
	0,3	0,99	0,01	0,99	0,01
	0,4	0,87	0,09	0,79	0,13
1	0,1	0,99	0,04	0,99	0,00
	0,3	1,00	0,01	1,00	0,00
	0,4	1,00	0,00	0,94	0,00
5	0,4	1,00	0,00	1,00	0,00

Fonte: Do autor (2022).

as menores TFD's e o PCOut se destaca por obter melhores TS's quando $\delta \geq 0,30$ e o $\lambda \leqslant 0,5$. Além disso, se os resultados da Tabela 4.15 forem comparados com os da Tabela (4.14), pode-se verificar uma melhora expressiva na TS de ambos os métodos.

Tabela 4.15 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* com n = 100, $\xi = 5$ e $p = 200$.

λ	δ	PCOut		COMEDIAN	
		TS	TFD	TS	TFD
0,1	0,1	1,00	0,03	1,00	0,00
	0,2	1,00	0,01	1,00	0,00
	0,3	1,00	0,00	0,98	0,00
0,5	0,1	1,00	0,03	1,00	0,00
	0,3	1,00	0,00	1,00	0,00
	0,4	0,97	0,03	0,88	0,03
1	0,1	1,00	0,03	1,00	0,00
	0,3	1,00	0,00	1,00	0,00
	0,4	0,97	0,02	0,97	0,01
5	0,4	0,98	0,01	0,99	0,00

Fonte: Do autor (2022).

4.5 Detecção de *outliers* com dados correlacionados

Para verificar a detecção de *outliers* em dados correlacionados foram abortados três situações, sendo elas, dados sem *outliers*, dados com *outliers* e dados com alta dimensão com a presença de *outliers*. As simulações realizadas abordaram somente os casos extremos. Para dados com contaminação foi usado somente taxa de 40%, a matriz de covariância foi alterada com λ = 0,1 e 5. Para correlações foi utilizado o parâmetro ρ, assumindo valores de 0,1, para baixa correlação, 0,5 para média e 0,9 para alta correlação.

No primeiro caso foram simulado dados sem *outliers*. Portanto, o melhor método foi escolhido por meio da TFD. Os números de variáveis usados foram somente 5 e 50. Por meio da Tabela 4.16 foi possível observar que os métodos MVE e MCD não sofreram nenhuma influência dos dados correlacionados. Os métodos PCOut e OGK, sofreram influência nas TFD's somente quando p foi aumentado para 50. O *Comedian* foi o método mais sensível as correlações dos dados. Por outro lado, foi o método com melhor desempenho, apresentando novamente as melhores TFD's entre os métodos comparados.

No segundo caso simulado foram considerados dados com a presença de *outliers*, portanto, buscou-se avaliar as TS's e as TFD's. Foi considerado somente taxa de contaminação de 40% e alterações na covariância de 0,1 e 5 (λ), $n = 100$ e $p = 5$ e 50. Considerando $p = 5$, o MCD foi o método que obteve as melhores TS's enquanto λ = 5 para qualquer ρ. Para covari-

Tabela 4.16 – TFD's dos métodos de detecção de *outliers* nas simulações de dados correlacionado sem *outliers*, com n = 100

p	ρ	MVE	MCD	OGK	PCOut	COMEDIAN
5	0,1	0,05	0,04	0,10	0,12	0,01
	0,5	0,05	0,04	0,11	0,12	0,03
	0,9	0,05	0,04	0,11	0,12	0,06
50	0,1	0,25	0,24	0,14	0,10	0,00
	0,5	0,25	0,24	0,15	0,10	0,03
	0,9	0,25	0,24	0,14	0,11	0,11

Fonte: Do autor (2022).

ância de 0,1, as melhores TS's foram obtidas com o PCOut. O *Comedian* obteve as melhores TDF's. Para $p = 50$ os métodos MVE e MCD são os piores considerando qualquer cenário. As melhores TFD's foram obtidas com o *Comedian* para qualquer simulação, conforme resultados apresentados na Tabela 4.17. Considerando λ = 0,1, o PCOut obteve as maiores TS's, mas nenhum método foi satisfatório com esse cenário. Para λ = 5 com 50 dimensões, os métodos PCOut, OGK e *Comedian* tiveram excelentes resultados.

Tabela 4.17 – TS's e TFD's dos métodos de detecção de *outliers* nas simulações com dados correlacionados contendo a presença de *outliers*; com n = 100, $\xi = 5$ e $\delta = 0,40$.

p	ρ	λ	MVE		MCD		OGK		PCOut		COMEDIAN	
			TS	TFD	TS	TFD	TS	TFD	TS	TFD	TS	TFD
5	0,1	0,1	0,00	0,39	0,00	0,57	0,00	0,42	0,35	0,17	0,15	0,03
		5	0,96	0,00	1,00	0,00	0,81	0,01	0,88	0,00	0,75	0,00
	0,5	0,1	0,00	0,39	0,00	0,57	0,00	0,45	0,06	0,31	0,08	0,05
		5	0,87	0,00	0,93	0,00	0,78	0,01	0,77	0,01	0,65	0,00
	0,9	0,1	0,00	0,39	0,00	0,58	0,00	0,42	0,31	0,00	0,09	0,05
		5	0,78	0,00	0,84	0,00	0,76	0,01	0,67	0,01	0,60	0,00
50	0,1	0,1	0,00	0,42	0,00	0,42	0,00	0,49	0,53	0,22	0,08	0,02
		5	0,56	0,04	0,56	0,03	1,00	0,02	1,00	0,00	1,00	0,00
	0,5	0,1	0,00	0,42	0,00	0,42	0,00	0,50	0,01	0,30	0,01	0,05
		5	0,56	0,05	0,57	0,03	1,00	0,02	1,00	0,00	0,99	0,00
	0,9	0,1	0,00	0,42	0,00	0,42	0,00	0,50	0,12	0,08	0,02	0,08
		5	0,56	0,05	0,56	0,03	1,00	0,02	1,00	0,00	0,94	0,01

Fonte: Do autor (2022).

No último caso (Tabela 4.18), foram considerados dados simulados com alta dimensão. Foi fixado $n = 40$ e alterado $p = 50$ e 100, as demais quantidades foram mantidas com as mesmas alterações. Com relação as TFD's o *comedian* foi o melhor método exceto quanto ρ

= 0,9 e λ = 5, messa condição o PCOut obteve melhores TFD. As TS's ficaram bem divididas entre os dois métodos. O mais interessante foi que com ρ = 0,5 e 0,9 o *comedian* passou a ter TS's melhores que o PCOut quando λ = 0,1.

Tabela 4.18 – TS's e TFD's dos métodos de detecção de *outliers* para simulações com *outliers* em dados correlacionados com alta dimensão com n = 40, $\xi = 5$ e $\delta = 0,40$.

p	ρ	λ	PCOut		COMEDIAN	
			TS	TFD	TS	TFD
50	0,1	0,1	0,50	0,19	0,42	0,03
		5	0,99	0,00	1,00	0,00
	0,5	0,1	0,06	0,33	0,26	0,05
		5	0,99	0,01	0,97	0,01
	0,9	0,1	0,24	0,08	0,25	0,09
		5	1,00	0,01	0,91	0,03
100	0,1	0,1	0,43	0,18	0,42	0,00
		5	0,95	0,01	1,00	0,00
	0,5	0,1	0,03	0,29	0,27	0,03
		5	098	0,01	0,99	0,02
	0,9	0,1	0,19	0,10	0,23	0,09
		5	1,00	0,01	0,94	0,04

Fonte: Do autor (2022).

Para esse último cenário de dados correlacionados com alta dimensão usando taxa de contaminação de 40% ficou claro que o *Comedian* reage de maneira mais positiva que o PCOut na detecção de *outliers*, atingindo resultados melhores.

Por fim, para verificar se os métodos têm resultados melhores em dados correlacionados ou em dados sem correlação, algumas comparações foram realizadas. Nas análises apresentadas a seguir, foram considerados apenas os casos de λ = 0,1, com taxa de contaminação igual a 40%. Neste cenário foram obtidos os piores resultados em todos métodos.

O primeiro caso, foi dados sem a presença de *outliers*. Na Tabela 4.19 apresentam-se os resultados das TFD's, em que pode-se verificar que os métodos MVE, MCD, OGK e PCOut praticamente não sofreram alterações nos resultados, embora o método *Comedian* tenha oscilado um pouco mais, mas por outro lado, foi o melhor método, independente do número de variáveis.

O segundo caso, foram usados dados com *outliers* de localização e dispersão. Pela análise dos resultados apresentados na Tabela 4.20, verificou-se que os métodos MVE, MCD e OGK continuaram sem sofrer alterações (são independentes dos dados estarem correlacionado

Tabela 4.19 – Comparação das TFD's para dados correlacionados e dados independentes; $\lambda = 0{,}1$; Sem *outliers*

MÉTODO	p	ρ			DADOS SEM CORRELAÇÃO
		0,1	0,5	0,9	
MVE	5	0,05	0,05	0,05	0,05
MCD		0,04	0,04	0,04	0,04
OGK		0,10	0,11	0,11	0,11
PCOUT		0,12	0,12	0,12	0,13
COMEDIAN		0,01	0,03	0,06	0,01
MVE	50	0,25	0,25	0,25	0,25
MCD		0,24	0,24	0,24	0,24
OGK		0,14	0,15	0,14	0,13
PCOUT		0,10	0,10	0,11	0,10
COMEDIAN		0,00	0,03	0,11	0,00

Fonte: Do autor (2022).

ou não). Os métodos PCOut e *Comedian* tiveram variações nos resultados (mas foram novamente os melhores). Interessante observar que ambos os métodos perderam eficiência nos teste com dados correlacionados, principalmente o método PCOut com correlação de 0,5.

Tabela 4.20 – Comparação das TS's e TFD's para dados correlacionados e dados independentes; $\lambda = 0{,}1$; Taxa de contaminação 40%; dados com *outliers*

MÉTODO	p	ρ						DADOS SEM CORRELAÇÃO	
		0,1		0,5		0,9			
		TS	TFD	TS	TFD	TS	TFD	TS	TFD
MVE	5	0,00	0,39	0,00	0,39	0,00	0,39	0,01	0,39
MCD		0,00	0,57	0,00	0,57	0,00	0,58	0,00	0,57
OGK		0,00	0,42	0,00	0,45	0,00	0,42	0,01	0,41
PCOUT		0,35	0,17	0,06	0,31	0,31	0,00	0,52	0,11
COMEDIAN		0,15	0,03	0,08	0,05	0,09	0,05	0,17	0,02
OGK	50	0,00	0,49	0,00	0,50	0,00	0,50	0,00	0,50
PCOUT		0,53	0,22	0,01	0,30	0,12	0,08	0,95	0,08
COMEDIAN		0,08	0,01	0,01	0,05	0,02	0,05	0,38	0,01

Fonte: Do autor (2022).

Para os dados com alta dimensão, a comparação foi feita apenas com o PCOut e *Comedian*. As TS's foram inferiores nos dados correlacionados, mas as TFD's usando o *Comedian* foram melhores. Como pode ser observado na Tabela 4.21 e já comentado anteriormente o PCOut, perde mais eficiência quando $\rho = 0{,}5$ e 0,9, ou seja, com correlações mais altas (mediana e alta).

Tabela 4.21 – Comparação das TS's e TFD's para dados correlacionados e dados independentes; $\lambda = 0{,}1$; Taxa de contaminação 40%; dados com alta dimensão correlacionados

MÉTODO	p	ρ						DADOS SEM CORRELAÇÃO	
		0,1		0,5		0,9			
		TS	TFD	TS	TFD	TS	TFD	TS	TFD
PCOUT	50	0,50	0,19	0,06	0,33	0,24	0,08	0,81	0,16
COMEDIAN		0,42	0,03	0,26	0,05	0,25	0,09	0,61	0,24
PCOUT	100	0,43	0,18	0,04	0,29	0,19	0,10	0,84	0,15
COMEDIAN		0,42	0,01	0,28	0,03	0,24	0,09	0,61	0,23

Fonte: Do autor (2022).

5 CONSIDERAÇÕES GERAIS

Nessa dissertação foram observados os métodos que têm sido mais utilizados para comparação de detecção de *outliers*, que são o MVE, o MCD e o OGK. Além desses métodos foram utilizados o método *Comedian* e PCOut, para detecção de *outliers* em dados com $n > p$ e para dados com alta dimensão $p > n$.

Os resultados das simulações de dados sem a presença de *outliers* apontam para o *Comedian* como melhor método, pois foi o método que menos detecta falsos *outliers*. Vale ressaltar que, o uso da distribuição normal pode produzir *outliers* independente de contaminação ou não. Os métodos identificam essas observações, mas fica a critério do pesquisador verificar se o *outlier* trouxe prejuízo ou não para o modelo de interesse e cabe portanto ao pesquisador remover ou não a observação.

Para os casos de *outliers* com relação a dispersão, o método com maiores taxa de acerto foi o OGK, mas também foi o método com maiores taxas de falsa detecção. O *Comedian* foi o método que obteve as menores taxa de falsa detecção. O mais interessante nessas simulações foi que à medida que p aumentava, os métodos OGK, PCOut e *Comedian* apresentavam TS's melhores.

Para simulação com *outliers* de locação e dispersão, há fortes indícios que apontam o *Comedian* como melhor método quando considerado uma taxa de contaminação menor que 30%. Mas, à medida que a dispersão dos dados (matriz de covariância) fica maior, o *Comedian* fica melhor, mesmo com 40% de taxa de contaminação. Todavia, o PCOut também apresentou excelentes resultados nas simulações com $\delta = 0,40$. Como esperado quanto maior for a média e variância da distribuição normal contaminada melhores os resultados para todos os métodos. Portanto, eleger o melhor método é uma tarefa difícil, mas é possível apontar alguns caminhos para os pesquisadores, como por exemplo, aplicar pelo menos dois testes de detecção de *outliers*. Também é fundamental o pesquisador ter conhecimento dos dados que está trabalhando, assim quando um método apontar alguma observação como *outliers*, cabe a ele verificar se houve influência ou não dos *outliers* e de manter ou eliminar essa observação. Pelos resultados apresentados nessa dissertação fica sugerido utilizar o *Comedian* e PCOut.

Em simulações com alta-dimensão há mais uma vez indícios que o método *Comedian* apresenta resultados melhores. Porém mais uma vez, enquanto a taxa de contaminação foi de 40%, o melhor método foi o PCOut. Os resultados de ambos os métodos melhoraram com o aumento da variância. Também vale ressaltar que o aumento da variância tem um reflexo

mais positivo no *Comedian*, pois em alguns casos ele chegou a superar os resultado do PCOut com contaminação de 40%. Também é valido observar: existem indícios que para dados de alta-dimensão quanto mais n se aproxima de p melhores serão os resultados.

No último cenário foi observado que o *comedian* é bem sensível aos dados correlacionados, da mesma forma que houve redução de desempenho dos demais métodos. Porém essa interferência não prejudicou os resultados relativos do *Comedian*, pois continuou sendo o melhor método para estes casos também de dados correlacionados sem *outliers*. Nos dados correlacionados com a presença de *outliers*, o PCOut se destaca em alguns casos como melhor método, embora o *comedian* obteve resultados satisfatórios (quando λ =5).

As simulações, no geral, apontaram para alguns pontos interessantes que podem ajudar na detecção de *outliers*. O primeiro ponto é que os métodos PCOut e *Comedian* tendem a ter resultados melhores quando existem muitas variáveis, independente da quantidade de *outliers*, considerando $p < n$. Segundo ponto, para $p > n$, os resultados melhoram quando n se aproxima de p. Por fim, quando $p = 5$, os melhores métodos foram o MVE e MCD. Portanto, esses dois métodos tendem a ser melhores quando o número de variáveis é menor.

6 CONCLUSÕES

Dentre os métodos comparados pode-se concluir que para $p \leq 5$, os melhores métodos são MCD e MVE. Nos demais cenários os melhores métodos foram o *comedian* e PCOut, mas não é possível dizer qual o melhor métodos dentre os dois, pois em algumas situações o *comedian* é superior em outras o PCOut é superior. Pode-se concluir portanto que o ideal é aplicar ambos os testes e analisar os *outliers* apontados de forma individual.

Para finalizar, essa dissertação deixa-se claro que ainda há muito para percorrer no assunto de detecção de *outliers*. Existem outras simulações a serem testadas e a possibilidade de aperfeiçoar o método *Comedian* usando *boostrap* e/ou simulação de Monte Carlo. Também existe a possibilidade de realizar comparações com outros métodos de detecção. Por fim, esses apontamentos são apenas o começo para uma nova pesquisa.

REFERÊNCIAS

BARBOSA, J. J. Data-driven cluster analysis method: uma nova metodologia para detecção de outliers em dados multivariados. Universidade Federal de Viçosa, 2021.

BARBOSA, J. J.; DUARTE, A. R.; MARTINS, H. S. R. A performance evaluation in multivariate outliers identification methods. **Ciência e Natura**, v. 42, p. 16, 2020.

BARBOSA, J. J.; PEREIRA, T. M.; OLIVEIRA, F. L. P. de. Uma proposta para identificação de outliers multivariados. **Ciência e Natura**, v. 40, p. 40, jul 2018. ISSN 2179-460X. Disponível em: <https://periodicos.ufsm.br/cienciaenatura/article/view/29535>.

CABANA, E.; LILLO, R. E.; LANIADO, H. Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators. **Statistical Papers**, Springer, v. 62, n. 4, p. 1583–1609, 2021.

CHUNG, H. C.; AHN, J. Subspace rotations for high-dimensional outlier detection. **Journal of Multivariate Analysis**, Elsevier, v. 183, p. 104713, 2021.

FALK, M. On mad and comedians. **Annals of the Institute of Statistical Mathematics**, Springer, v. 49, n. 4, p. 615–644, 1997.

FERREIRA, D. F. **Estatística multivariada**. [S.l.]: Editora Ufla Lavras, 2008.

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

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

HADI, A. S. Identifying multiple outliers in multivariate data. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 54, n. 3, p. 761–771, 1992.

HAWKINS, D. M. **Identification of outliers**. [S.l.]: Springer, 1980. v. 11.

HUBERT, M.; DEBRUYNE, M. Breakdown value. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, v. 1, n. 3, p. 296–302, 2009.

HUBERT, M.; ROUSSEEUW, P. J.; AELST, S. V. High-breakdown robust multivariate methods. **Statistical science**, JSTOR, p. 92–119, 2008.

HUBERT, M.; VEEKEN, S. Van der. Outlier detection for skewed data. **Journal of Chemometrics: A Journal of the Chemometrics Society**, Wiley Online Library, v. 22, n. 3-4, p. 235–246, 2008.

KUNJUNNI, S. O.; ABRAHAM, S. T. Multidimensional outlier detection and robust estimation using sn covariance. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, p. 1–11, 2020.

MARONNA, R. A.; YOHAI, V. J. Robust and efficient estimation of multivariate scatter and location. **Computational Statistics & Data Analysis**, Elsevier, v. 109, p. 64–75, 2017.

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.

PALMA, M. 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, 2021. Disponível em: <https://www.R-project.org/>.

RO, K. et al. Outlier detection for high-dimensional data. **Biometrika**, Oxford University Press, v. 102, n. 3, p. 589–599, 2015.

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.

ROUSSEEUW, P. J.; DRIESSEN, K. V. A fast algorithm for the minimum covariance determinant estimator. **Technometrics**, Taylor & Francis Group, v. 41, n. 3, p. 212–223, 1999.

ROUSSEEUW, P. J.; HUBERT, M. Robust statistics for outlier detection. **Wiley interdisciplinary reviews: Data mining and knowledge discovery**, Wiley Online Library, v. 1, n. 1, p. 73–79, 2011.

ROUSSEEUW, P. J.; ZOMEREN, B. C. V. Unmasking multivariate outliers and leverage points. **Journal of the American Statistical association**, Taylor & Francis, v. 85, n. 411, p. 633–639, 1990.

SAJESH, T. A.; SRINIVASAN, M. R. Outlier detection for high dimensional data using the Comedian approach. **Journal of Statistical Computation and Simulation**, v. 82, n. 5, p. 745–757, may 2012. ISSN 0094-9655. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/00949655.2011.552504>.

VELOSO, M. V. de S.; CIRILLO, M. A. Principal components in the discrimination of outliers: A study in simulation sample data corrected by pearson's and yates´ s chisquare distance. **Acta Scientiarum. Technology**, Universidade Estadual de Maringa, v. 38, n. 2, p. 193–200, 2016.

WOODRUFF, D. L.; ROCKE, D. M. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. **Journal of the American Statistical Association**, Taylor & Francis, v. 89, n. 427, p. 888–896, 1994.

APÊNDICE A – EXEMPLOS DE APLICAÇÃO DOS MÉTODOS

Para exemplo de aplicação passo a passo dos métodos foi utilizado o mesmo conjunto de dados já apresentado no referencial teórico através da Tabela 2.1 para a distância de Mahalanobis. Foi incluído a variável$X2$ para obter um caso bi-variado. Esse conjunto de dados está apresentado novamente na Tabela 1.

Exemplo para MVE

Tabela 1 – Observações por indivíduos - Exemplo 01

Indivíduo (i)	X1	X2	DM MVE 1ª interação	DM MVE 2ª interação
1	2	25	9,77	1,33
2	9	29	2,44	2,48
3	3	23	5,81	0,68
4	6	21	2,77	1,33
5	4	24	3,06	0,18
6	5	24	1,33	0,09
7	8	25	1,33	1,42
8	5	26	1,27	0,02
9	6	26	0,19	0,14
10	7	30	1,33	1,33
11	7	28	0,27	0,74
12	20	21	100,33	38,22
13	20	25	89,08	34,54
14	20	23	94,27	36,17
15	50	15	918,77	309,49

Fonte: Do autor (2022).

1º Passo, retirar uma amostra aleatória de tamanho $p + 1$ (no R Core Team (2021), pode ser usado a função *sample*). O tamanho da amostra pode ser configurado na função do MVE.

$$J = \begin{pmatrix} 8 & 25 \\ 5 & 24 \\ 7 & 20 \end{pmatrix}$$

2º Passo, encontrar as médias e matriz de covariância. (no R Core Team (2021), pode ser usado a função *apply(x,2 mean)* e *var(x)*).

$$T_j = (6,67, \quad 23) \quad e \quad C_j = \begin{pmatrix} 2,33 & 0 \\ 0 & 7 \end{pmatrix}$$

3º Calcular a distância de Mahalanobis usando o vetor de média e matriz de variância encontrado no passo 2. (vetor de média será o centro e a matriz será a matriz de dispersão;(no R Core Team (2021), pode ser usado a função *mahalanobis*) O resultado está na tabela (1).

4º Passo, verificar se foi atendida a condição da equação 2.8:

$$\#\{i;(\boldsymbol{x}_i - T)^\top \boldsymbol{C}^{-1}(\boldsymbol{x}_i - T) \leq a^2\} \geq h \tag{1}$$

em que $h = (n+p+1)/2$ e $a^2 = \chi_{0.5,p}$. Nesse exemplo $h = 9$ e $\chi_{0.5,2} = 1,386294$. O tamanho de h e a podem ser configurados na função do MVE; Observe também que $(\boldsymbol{x}_i - T)^\top \boldsymbol{C}^{-1}(\boldsymbol{x}_i - T)$ é a distância de Mahalanobis

$$i; DM_i \leq 1,386294 \geq 9, \quad i = 1,2,...20 \tag{2}$$

Para esse exemplo, foi necessário ter pelo menos 9 pontos dentro da elipsoide. Na primeira amostragem a condição não foi atendida. O processo exige várias re-amostragens. Abaixo segue uma re-amostragem que obteve êxito.

$$J = \begin{pmatrix} 2 & 25 \\ 6 & 21 \\ 7 & 30 \end{pmatrix}$$

$$T_j = (5,00, \quad 25,33) \quad e \quad C_j = \begin{pmatrix} 7,00 & 3,00 \\ 3,00 & 20,33 \end{pmatrix}$$

5º Passo: calcular o determinante da matriz $\boldsymbol{C}_j$ que atendeu a condição e a mediana das distância de Mahalanobis.

$Det(C_j) * med(DM2) = vomule \quad da \quad elpsoide$

6º Encontrar outras amostras que atendem a equação (1). Selecionar a equação com menor volume. Na amostragem selecionada aplicar a equação (3). Supondo que a amostra vista anteriormente seja a melhor.

$$\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}_J \quad e \quad \boldsymbol{C}(\boldsymbol{x}) = (\chi^2_{0.5,p})^{-1} m_J \boldsymbol{C}_J \tag{3}$$

$$T(X) = (5,00, \quad 25,33) \quad e \quad C(X) = \begin{pmatrix} 5,05 & 2,16 \\ 2,16 & 14,66 \end{pmatrix}$$

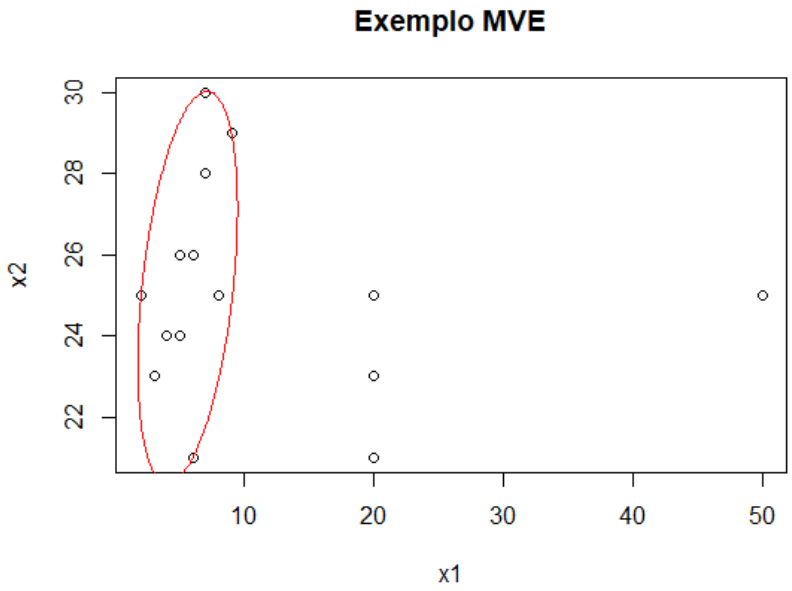
Agora aplica-se a distância de Mahalanobis. A proposta do MVE é obter a raiz quadrada da Distância de Mahalanobis, conforme visto no referencial teórico. Pode-se também aplicar a função de ponderação, atribuindo "0" para as observações que são *outliers* e "1" para observação que não são *outliers* (esse procedimento não foi realizado nesse exemplo).

Até aqui foram encontradas as estimativas robustas para ser utilizadas na função de Mahalanobis. As distâncias obtidas estão demonstradas abaixo.

DM = (1,36, 1,86, 0,97, 1,36, 0,51, 0,36, 1,40, 0,18, 0,45, 1,36, 1,01, 7,27, 6,92, 7,08, 20,71).

O valor de corte é definido por $\sqrt{\chi_{0,975,p}} = 2,716$ Segue o gráfico 1 da elipsóide, para esse conjunto de dados.

Figura 1 – Gráfico de dispersão dos dados mostrando a elipsoide - MVE



Fonte: Do autor (2022).

Exemplo para MCD

1º Passo: Retirar uma amostra H_0 de tamanho $(p+1)$ e encontrar o determinante. Para facilitar esse exemplo, optou-se em usar a mesma amostra utilizada no MVE.

$$H_0 = \begin{pmatrix} 2 & 25 \\ 6 & 21 \\ 7 & 30 \end{pmatrix}$$

$$T_j = (5,00, \quad 25,33) \quad e \quad S_j = \begin{pmatrix} 7,00 & 3,00 \\ 3,00 & 20,33 \end{pmatrix}$$

2º Passo: Obter as distâncias de Mahalanobis e colocar em ordem. (novamente utilizou-se a raiz quadrada das distâncias de Mahalanobis). Abaixo está relacionado a ordem conforme a posição das observações na matriz de X.

d_i = (8, 6, 9, 5, 3, 11, 4, 1, 10, | 7, 2, 13, 14, 12, 15)

3º Construir o conjunto H_1 contendo h observações, em ordem. Sendo que h pode ser $[n+p+1]/2$. Na função do MCD é possível determinar o tamanho de h. Nesse caso h = 9.

Tabela 2 – Observações por indivíduos - Exemplo 01

Indivíduo (i)	X1	X2
1	2	25
3	3	23
4	6	21
5	4	24
6	5	24
8	5	26
9	6	26
10	7	30
11	7	28

Fonte: Do autor (2022).

4º Passo: Calcular o determinante da matriz de covariância de H_1, Caso o determinante seja zero, deve-se incluir mais observações até que o determinante seja diferente de zero. Em seguida, calcular a distância de Mahalanobis e colocar novamente em ordem.

$$T_{h1} = (5,00, \quad 25,22) \quad e \quad S_{h1} = \begin{pmatrix} 3,00 & 2,25 \\ 2,25 & 7,19 \end{pmatrix}$$

$$det(h1) = 16,52$$

d_i = (8, 6, 9, 5, 3, 11, 10, 1, 7, | 4, 2, 13, 14, 12, 15)

Observe que dessa vez, a observação 4 ficou fora das 9 menores distâncias e a observação 7 entrou dentro das 9 menores. Novamente, calcular a média, variância e distância.

A ideia do algorítimo é fazer isso até que o determinante não seja mais reduzido. Isso deve ser feito com várias amostras. Será selecionado a matriz de covariância com o menor determinante. Depois de obtido, calcular a distância de Mahalanobis. Supondo que a amostra escolhida seja aquela que produziu o menor determinante.

$$T_{h1} = (5,22, \quad 25,66) \quad e \quad S_{h1} = \begin{pmatrix} 3,94 & 2,58 \\ 2,58 & 4,75 \end{pmatrix}$$

$$det(h1) = 12,0625$$

$$T_{h1} = (5,44, \quad 25,55) \quad e \quad S_{h1} = \begin{pmatrix} 5,27 & 3,22 \\ 3,22 & 3,77 \end{pmatrix}$$

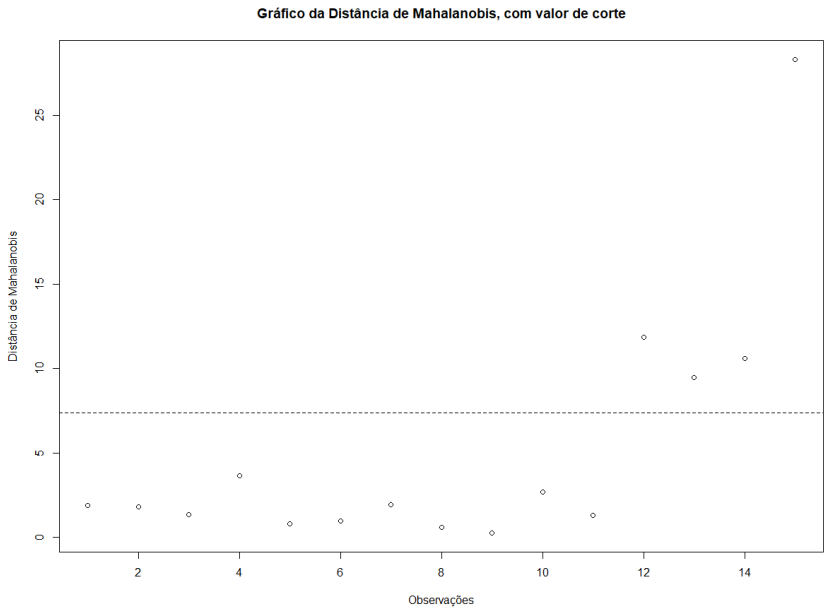
$$det(h1) = 9,55$$

Até aqui foram encontradas as estimativas robustas para ser utilizadas na função de Mahalanobis. As distâncias obtidas estão demonstradas abaixo.

d_i = (1,88 1,81 1,32 3,64 0,80 0,97 1,92 0,56 0,25 2,68 1,30 11,82 9,45 10,60 28,31)

O valor de corte é definido por $\sqrt{\chi_{0.975,p}} = 2,716$. Qualquer distância acima de 2,716 será considerada *outliers*.

Figura 2 – Gráfico com as distâncias encontradas e a linha de corte - MCD



Fonte: Do autor (2022).

Exemplo para *Comedian*

1º Passo: Obter a matriz $\boldsymbol{COM(X)}$ em seguida construir a matriz $\boldsymbol{\delta(X)}$. Abaixo segue a definição utilizada e os resultados encontrados. Para encontrar $\boldsymbol{\delta(X)}$ é necessário criar a matriz diagonal $\boldsymbol{D}$.

$$COM(X_i, X_j) = med((X_i - med(X_i))(X_j - med(X_j))), \tag{4}$$

$$COM(\boldsymbol{X}) = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

em que $\boldsymbol{D}$ a matriz diagonal dos elementos $1/MAD(X_i)$ sendo i = 1, 2,...,p (no R Core Team (2021) pode ser usado a função *diag* para criar a diagonal principal. Para encontrar o *MAD* de cada variável pode ser usado a função *mad*. Uma forma bem fácil de obter todas os *MAD*'s das variáveis é usar a função "apply(x, 2, mad)").

$$D = \begin{pmatrix} 0,33 & 0 \\ 0 & 0,67 \end{pmatrix}$$

$$\delta(\boldsymbol{X}) = \boldsymbol{D}COM(\boldsymbol{X})\boldsymbol{D}^{\top}, \tag{5}$$

$$\boldsymbol{\delta(X)} = \begin{pmatrix} 0,45 & 0 \\ 0 & 0,45 \end{pmatrix}$$

2º Passo: Transformar a matriz $\delta(\boldsymbol{X})$ em uma matriz semi-positiva definida. Para isso, obtém-se os autovetores de $\delta(\boldsymbol{X})$ e chamar de $\boldsymbol{E}$ (no R Core Team (2021), pode ser usado o a função *eigen*). Em seguida obter $\boldsymbol{Q} = D(\boldsymbol{X})^{-1}\boldsymbol{E}$ e $\boldsymbol{z}_i = \boldsymbol{Q}^{-1}\boldsymbol{X}_i$. Os dados encontrados para $\boldsymbol{z}$ estão na Tabela (3).

$$\boldsymbol{E} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

$$\boldsymbol{Q} = \begin{pmatrix} 0 & -2,96 \\ 1,48 & 0 \end{pmatrix}$$

Tabela 3 – z_i Obtida com $\boldsymbol{z_i = Q^{-1} X_i}$

Indivíduo (i)	z1	z2
1	16,86	-0,67
2	19,56	-3,03
3	15,51	-1,01
4	14,16	-2,02
5	16,18	-1,34
6	16,18	-1,68
7	16,86	-2,69
8	17,53	-1,68
9	17,53	-2,02
10	20,23	-2,36
11	18,88	-2,36
12	14,16	-6,74
13	18,86	-6,74
14	15,51	-6,74
15	16,86	-16,86

Fonte: Do autor (2022).

3º Passo: Obter as estimativas robustos $\boldsymbol{S(X)}$ e $\boldsymbol{m(X)}$.

$$S(\boldsymbol{X}) = \boldsymbol{Q\Gamma Q}^{\top} \qquad e \qquad m(\boldsymbol{X}) = \boldsymbol{Ql} \tag{6}$$

em que $\boldsymbol{\Gamma} = diag(MAD(\boldsymbol{Z}_1)^2, ..., MAD(\boldsymbol{Z}_p)^2)$ e $\boldsymbol{l} = (med(\boldsymbol{Z}_1), ..., med(\boldsymbol{Z}_1))^{\top}$.

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\boldsymbol{l} = (16,86, \quad -2,36)$$

As estimativas robustas são:

$$S(\boldsymbol{X}) = \begin{pmatrix} 8,79 & 0 \\ 0 & 2,19 \end{pmatrix}$$

$$m(\boldsymbol{X}) = (7, \quad 25)$$

Até aqui foram encontradas as estimativas robustas para ser utilizadas na função de Mahalanobis.

4º Passo: Calcular as distâncias de cada observação usando a equação (7). E em seguida, encontrar o valor de corte.

$$\boldsymbol{RD(x_i, m)} = rd_i = \boldsymbol{(x_i - m)}^{\top} \boldsymbol{S}^{-1} \boldsymbol{(x_i - m)}, \qquad i = 1, 2, ..., n, \tag{7}$$

RD = (2,84, 7,73, 3,63 ,7,39, 1,47, 0,90, 0,11, 0,90, 0,56, 11,37, 4,09, 26,50, 19,22, 21,04, 210,294)

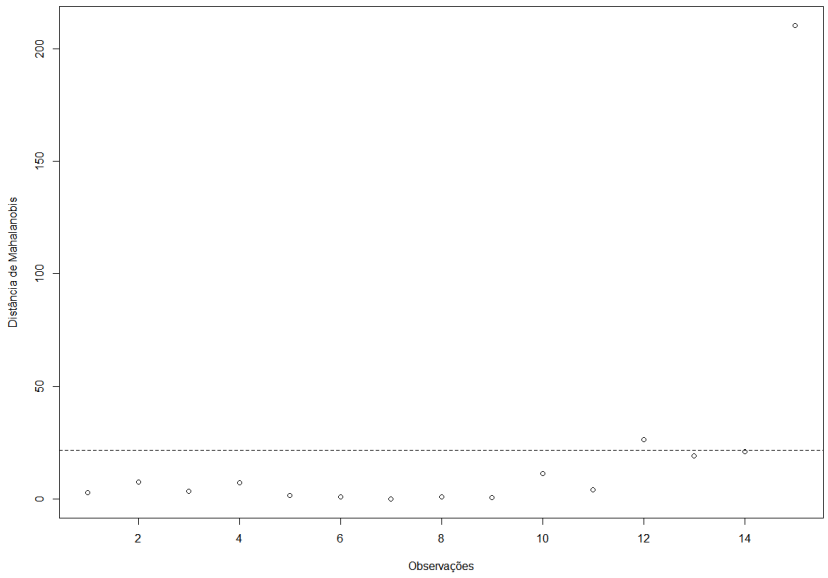
O valor de corte é obtido por meio da equação 8.

$$cv = \frac{1,4826[\chi^2_p(0.95) median(rd_1, ..., rd_n)]}{\chi^2_p(0.5)} \tag{8}$$

$cv = 21,79$

Observe que nesse caso, as observações 13 e 14 não foram detectados como *outliers*. Mas pelo gráfico é possível verificar que elas ficaram bem próxima da linha de detecção.

Figura 3 – Gráfico com as distâncias encontradas e a linha de corte - *Comedian*



Fonte: Do autor (2022).

Exemplo para OGK

1º Passo: definição do escalar robusto que será usado. Por padrão, o método OGK usa "τ *scale*". Definido da seguinte forma.

$$W_c(X) = \left(1 - \left(\frac{x}{c}\right)^2\right)^2 I(|x| \leqslant c) ep_c(x) = min(x^2, c^2)$$

Seja $X = x_1, ..., x_n$ uma amostra univariada então:

$\sigma_0 = MAD(X)$ e $W_i = W_{c1}\left(\frac{x_i - med(X)}{\sigma_0}\right)$

A estatística de locação e de dispersão (ou escala), são definidas por

$$\mu(\boldsymbol{X}) = \frac{\sum_i x_i w_i}{\sum_i w_i}$$

e

$$\sigma(\boldsymbol{X})^2 = \frac{\sigma_0^2}{n} \sum_i pc_2\left(\frac{x_i - \mu(X)}{\sigma_0}\right)$$

em que c_1 foi definido igual 4,5 e c_2 igual a 3, dessa forma o algoritmo obtém uma eficiência maior, com aproximadamente 80% (MARONNA; ZAMAR, 2002)(no R Core Team (2021), pode ser utilizado a função *scaleTau2*). Para o exemplo os valores obtidos foram: $\mu(X) =$ (5,94, 24,80) e $\sigma_0 =$ (3,69, 2,00)

2º Passo: construir Y usando $\boldsymbol{D} = diag(\sigma(\boldsymbol{X}_1), ..., \sigma(\boldsymbol{X}_p))$. $\boldsymbol{y}_i = \boldsymbol{D}^{-1}\boldsymbol{x}_i$. A tabela 4, apresenta todos os dados de $\boldsymbol{Y}$.

Tabela 4 – $\boldsymbol{y_i}$ Obtida com $\boldsymbol{y}_i = \boldsymbol{D}^{-1}\boldsymbol{x}_i$

Indivíduo (i)	y1	y2
1	0,54	12,46
2	2,43	14,45
3	0,81	11,46
4	1,62	10,46
5	1,08	11,96
6	1,35	11,96
7	2,16	12,46
8	1,35	12,96
9	1,62	12,96
10	1,89	14,95
11	1,89	13,95
12	5,41	10,46
13	5,41	12,46
14	5,41	11,46
15	13,52	12,46

Fonte: Do autor (2022).

3º Passo: Calcular a matriz de correlação $\boldsymbol{U} = [\boldsymbol{U}_{jk}]$. Sendo $U_{jk} = v(\boldsymbol{Y}_j, \boldsymbol{Y}_k)$. Usando o estimador de Gnanadesikan-Kettering como v. A equação do estimador de Gnanadesikan-Kettering está definida abaixo:

$$U_{jk} = \frac{1}{4}[\sigma(Y_j + Y_k)^2 - \sigma(Y_j - Y_k)^2], \quad j \neq k. \tag{9}$$

em que σ é um escalar. Nesse caso, como ja mencionado foi utilizado o "τ *scale*" (no R Core Team (2021) pode ser usada a função *covGK*).

$$\boldsymbol{U}_{jk} = \begin{pmatrix} 1 & -0,14 \\ -0,14 & 1 \end{pmatrix}$$

4º Passo: : Transformar a matriz $\boldsymbol{U}$ em uma matriz semi-positiva definida (mesmos passos aplicado no caso do Comedian). Os resultados obtidos segue abaixo:

$$\mathbf{T}(\mathbf{X}) = (\mathbf{7},\mathbf{13}, \quad \mathbf{25},\mathbf{73}) \quad \mathbf{e} \quad \mathbf{V}(\mathbf{X}) = \begin{pmatrix} 36,93 & -1,09 \\ -1,09 & 10,87 \end{pmatrix}$$

5º Passo: atribuir pesos e verificar o valor de corte. De acordo com a função de peso, será atribuído 1, para observações com distância menor que o valor de corte e 0 para observações com distância maior que o valor de corte. O valor de corte está definido na equação (10).

$$vc = \frac{\chi^2_p(B) med(d_1, ..., d_n)}{\chi^2_p(0.5)} \tag{10}$$

$$vc = 5,91627$$

distancias = (0,78 1,11 1,21 2,13 0,57 0,42 0,06 0,12 0,03 1,67 0,47 6,23 4,49 4,99 49,78)

Observe que, assim como o *Comedian*, as observações 13 e 14 ficaram abaixo do valor de corte, mas o método OGK possui mais um procedimento. Ponderar o valor de centro e a matriz de dispersão. Esse procedimento pode ser aplicado no *Comedian* para melhorar os resultados.

$$\boldsymbol{t_w} = \frac{\sum_i w_i \boldsymbol{x_i}}{\sum_i w_i} \qquad \boldsymbol{V_w} = \frac{\sum_i w_i (\boldsymbol{x_i} - \boldsymbol{t_w})(\boldsymbol{x_i} - \boldsymbol{t_w})^\top}{\sum_i w_i} \tag{11}$$

$$\boldsymbol{T(X)} = (5,63, \quad 25,54)$$

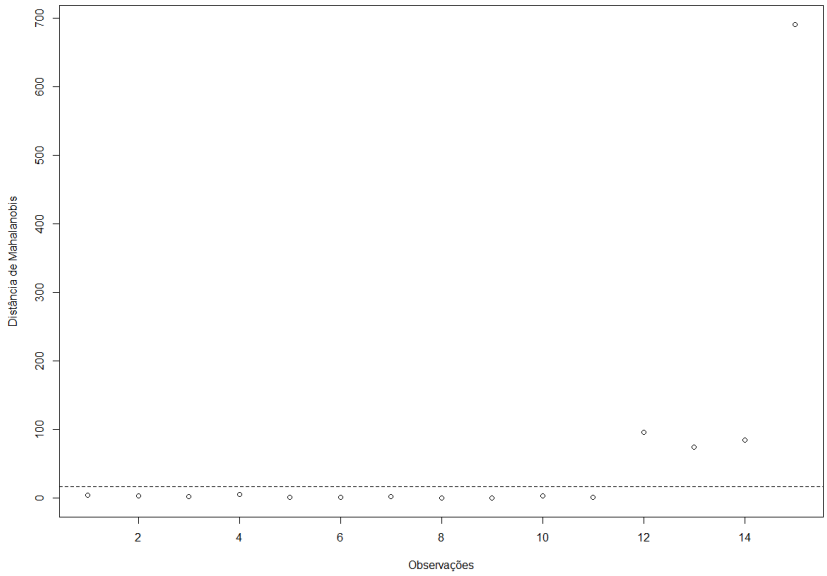
$$\boldsymbol{V(X)} = \begin{pmatrix} 4,04 & 2,74 \\ 2,74 & 6,42 \end{pmatrix}$$

Até aqui foram obtidas as estimativas robustas para serem aplicadas na equação de Mahalanobis. O resultado das distância e novo valor de corte segue abaixo:

distancias = (4,07 3,09 1,84 5,05 0,70 0,37 2,38 0,27 0,04 3,18 0,97 95,54 74,05 83,92 690,92)

$vc = 16,47$

Figura 4 – Gráfico com as distâncias encontradas e a linha de corte - OGK



Fonte: Do autor (2022).

Exemplo para PCOut

1º Passo: Reponderar a Matriz $\boldsymbol{X}$ conforme equação (12). A nova matriz reponderada está demonstrada na Tabela 5.

$$x_{ij}^{*} = \frac{x_{ij} - med(x_{1j},...,x_{nj})}{MAD(x_{1j},...,x_{nj})}, \qquad j = 1,...,p. \tag{12}$$

Tabela 5 – Dados ponderados de $\boldsymbol{X}$

Indivíduo (i)	x1	x2
1	-1,68	0,00
2	0,67	2,69
3	-1,34	-1,34
4	-0,33	-2,69
5	-1,01	-0,67
6	-0,67	-0,67
7	0,33	0,00
8	-0,67	0,67
9	-0,33	0,67
10	0,00	3,37
11	0,00	2,02
12	4,38	-2,69
13	4,38	0,00
14	4,38	-1,34
15	14,50	0,00

Fonte: Do autor (2022).

2º Passo: Verificar os autovalores que mais contribuem para a matriz de variância. Nesse caso temos apenas duas variáveis então, não vai haver redução de variáveis. Ainda assim, deve ser separado a matriz de autovetores e obter $\boldsymbol{Z} = \boldsymbol{X}^{*}\boldsymbol{V}$,. Em seguida, reponderar $\boldsymbol{Z}$ da mesma maneira que foi feito com $\mathbf{X}$. Os resultados para $\boldsymbol{Z}$ estão na Tabela 6.

Tabela 6 – Dados ponderados de $\boldsymbol{Z}$

Indivíduo (i)	z1	z2
1	-1,53	0,09
2	0,67	-2,02
3	-1,11	1,08
4	0,00	2,04
5	-0,81	0,56
6	-0,47	0,55
7	0,50	0,00
8	-0,55	-0,45
9	-0,21	-0,47
10	-0,05	-2,49
11	0,03	-1,49
12	4,77	1,81
13	4,60	-0,19
14	4,68	0,81
15	14,83	-0,67

Fonte: Do autor (2022).

3º Passo: Encontrar W_j ,e em seguida fazer $\frac{w_j}{\sum_i w_i}$ que é a probabilidade de ter *outlier* em cada variável. Quanto mais perto de 1 maior a chance de conter algum *outlier*. Depois, obter a $d1$ para os dados ponderados.

$$w_j = \mid \frac{1}{n}\sum_{i=1}^{n} \frac{(z^*_{ij} - med(z^*_{1j},...,z^*_{nj}))^4}{MAD(z^*_{1j},...,z^*_{nj})^4} - 3 \mid, \quad j = 1,...,p^* \tag{13}$$

$\frac{w_j}{\sum_i w_i} = (0,99, \quad 0,00)$.

Note que a probabilidade de ter um *outlier* na marginal z_1 é quase 100% e na marginal z_2 é zero. Para obter a distância é necessário utilizar as duas equações abaixo. Os resultados seguem apresentados na sequência.

$$RD_i = \sum_{j=1}^{p} \left(\frac{Zr_{ij} - \mu(Zr_j)}{\sigma(Zr_j)} \right)^2, \quad i = 1,...,n. \tag{14}$$

$$d_i = RD_i \frac{\sqrt{\chi^2_{p*0.5}}}{med(RD_1,...,RD_n)} \quad para \quad i = 1,...,n. \tag{15}$$

$d1_i$ =(2,68 1,17 1,93 0,00 1,41 0,82 0,88 0,97 0,38 0,08 0,06 8,33 8,03 8,18 25,89)

4º Passo: Obter as distâncias 02 e verificar os pesos. A segunda fase do algoritmo é semelhante a primeira, porém não faz uso do coeficiente de curtose para reponderar os componentes principais. Nesse caso, com os dados obtidos na equação (14) aplica-se a função da distância de Mahalanobis.

$d2_i$ = (1,16 1,61 1,17 1,54 0,75 0,54 0,38 0,54 0,39 1,89 1,13 3,87 3,49 3,60 11,25)

5º Passo: Por fim, atribuir construir w, seguindo as equações abaixo e juntar ambos os pesos. Observações com $w_i < 0,25$ serão consideradas *outliers*.

$$w_{1i} = \begin{cases} 0, & d_i \geqslant c, \\ \left(1 - \left(\frac{d_i - M}{c - M}\right)^2\right)^2 & M < d_i < c, \\ 1 & d_i \leqslant M \end{cases} \tag{16}$$

W_1 = (0,67 0,98 0,88 1,00 0,96 1,00 0,99 0,99 1,00 1,00 1,00 0,00 0,00 0,00 0,00)

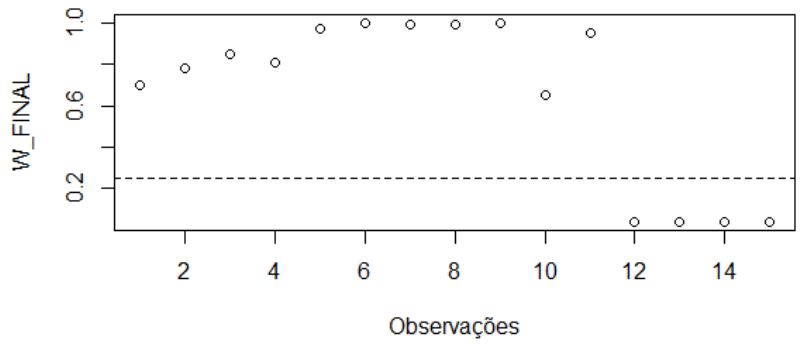
W_2 = (0,93 0,73 0,93 0,77 1,00 1,00 1,00 1,00 1,00 0,56 0,94 0,00 0,00 0,00 0,00)

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2} \tag{17}$$

w_i =(0,70 0,78 0,85 0,81 0,97 1,00 0,99 0,99 1,00 0,65 0,95 0,04 0,04 0,04 0,04)

Nesse caso as observações identificadas como *outliers* foram 12, 13, 14 e 15.

Figura 5 – Gráfico com W-final e linha de corte



Fonte: Do autor (2022).

Nesse caso os *outliers* ficam abaixo do valor de corte.

APÊNDICE B – COMANDOS USADOS NO R

Listing 1 – Algoritmo em R

```
#### pacotes utlizados ##################
library(robustbase)
library(MASS)
library(rrcov)
library(mvoutlier)



# Funcao para contar a quantidade de
# outliers identificados corretamente
# Conta tambem a quantidade de outliers
# identificados erroneamente.

verificador <- function(h,W,n)
{
  ta1 <- 0
  ta2 <- 0
  ta3 <- 0
  ta4 <- 0

  for (i in 1 : n)
  {
    #TAXA DE SUCESSO - OBSERVACOES QUE
    #SAO OUTLIERS IDENTIFICADAS
    #COMO OUTLIERS
    if (h[i] == 0 && W[i] == 0)
      ta1 <- ta1+1

    #TAXA DE ACERTO - OBSERVACOES QUE
    #SAO NORMAIS - IDENTIFICADAS
    #COMO NORMAIS
    if (h[i] == 1 && W[i] == 1)
      ta2 <- ta2 + 1
```

```
    #TAXA DE FRACASSO - OBSERVACOES QUE
    #SAO OUTLIERS - IDENTIFICADAS
    # COMO NORMAIS
    if (h[i] == 0 && W[i] == 1)
      ta3 <- ta3 + 1

    #TAXA DE FALSA DETECCAO - OBSERVACOES QUE
    #NAO SAO OUTLIERS IDENTIFICADAS
    #COMO OUTLIERS

    if (h[i] == 1 && W[i] == 0)
      ta4 <- ta4 + 1
  }

  return(list(ta1 = ta1, ta2 =ta2,ta3=ta3, ta4=ta4))
}

# Funcao que gera dados com distribuicao
#normal multivariada contaminada

rNCMWL <- function(n, delta, mu, mu2, Sigma, Sigma2)
{
  u <- runif(n)
  p <- nrow(Sigma)
  n1 <- length(u[u<= delta])
  n2 <- n - n1

  # vetor criado para identificar os outliers (0 <- outlier;
  #1 <-  observacao normal)
  pos <- c()
  for (i in 1 :n)
  {
    if (u[i]<= delta){
```

```
      pos[i]<- 0
    }else
    {
      pos[i] <- 1
    }

  }

  X <- matrix(0, n, p)
  if(n1 > 0) X[u <= delta, ] <- mvrnorm(n1, mu2, Sigma2)
  if(n2 > 0) X[u > delta, ] <- mvrnorm (n2, mu, Sigma)
  return(list("X" = X,"outliers" = n1,"normais"=n2,
              "posicao" =pos))
}


########################mve#########################

# Funcao para os testes do metodo MVE

sim_mve <- function(dados, nout, h)
{
  rej <- matrix(0, 3, 2)
  rownames(rej) <- c("MVE","Outliers (%)",
                     "Observacoes normais (%)")
  colnames(rej) <- c("Outliers detectados (%)",
                     "Outliers nao detectados (%)")

  ts_out <- 0
  ts_nor <- 0
  ts_fra <- 0
  tfd <- 0

  mve <- CovMve(dados)
```

```
  W <- mve$wt

  taxas <- verificador(h,W,n)

  ta1 <- taxas$ta1
  ta2 <- taxas$ta2
  ta3 <- taxas$ta3
  ta4 <- taxas$ta4

  if (nout > 0)
  {
    ts_out <- ta1/nout
    ts_nor <- ta2/(n - nout)
    ts_fra <- ta3/nout
    tfd <- ta4/(n - nout)
  }
  if (delta == 0 || nout == 0)
  {
    ts_out <- 0
    ts_nor <- ta2/(n - nout)
    ts_fra <- (1-ts_out)
    tfd <- ta4/(n - nout)
  }



  rej[2,1] <- ts_out
  rej[3,2] <- ts_nor
  rej[2,2] <- ts_fra
  rej[3,1] <- tfd

  return(rej)
}
```

```
##################MCD#####################################

# Funcao para os testes do metodo MCD

sim_mcd <- function(dados, nout, h)
{
  rej <- matrix(0, 3, 2)
  rownames(rej) <- c("MCD","Outliers (%)",
                     "Observacoes normais (%)")
  colnames(rej) <- c("Outliers detectados (%)",
                     "Outliers nao detectados (%)")

  ts_out <- 0
  ts_nor <- 0
  ts_fra <- 0
  tfd <- 0

  #if (p > n)
  #{
  # mcd <- rmdp(dados, alpha = 0.05)
  #  W <- mcd$wei
  #  W <- as.numeric(W)
  #}else{
    mcd <- CovMcd(dados)
    W <- mcd$wt
  #}

  taxas <- verificador(h,W,n)

  ta1 <- taxas$ta1
  ta2 <- taxas$ta2
  ta3 <- taxas$ta3
  ta4 <- taxas$ta4
```

```
  if (nout > 0)
  {
    ts_out <- ta1/nout
    ts_nor <- ta2/(n - nout)
    ts_fra <- ta3/nout
    tfd <- ta4/(n - nout)
  }
  if (delta == 0 || nout == 0)
  {
    ts_out <- 0
    ts_nor <- ta2/(n - nout)
    ts_fra <- (1-ts_out)
    tfd <- ta4/(n - nout)
  }

  rej[2,1] <- ts_out
  rej[3,2] <- ts_nor
  rej[2,2] <- ts_fra
  rej[3,1] <- tfd

  return(rej)
}

############# COMEDIAN ######################

# Funcaoo para os testes do metodo COMEDIAN

sim_comed <- function(dados, nout, h)
{
  rej <- matrix(0, 3, 2)
  rownames(rej) <- c("COMEDIAN","Outliers (%)",
                     "Observacoes normais (%)")
  colnames(rej) <- c("Outliers detectados (%)",
                     "Outliers nao detectados (%)")
```

```
ts_out <- 0
ts_nor <- 0
ts_fra <- 0
tfd <- 0
n <- nrow(dados)
p <- ncol(dados)

comed01 <- covComed(dados, n.iter = 5)
W <- comed01$weights

taxas <- verificador(h,W,n)

ta1 <- taxas$ta1
ta2 <- taxas$ta2
ta3 <- taxas$ta3
ta4 <- taxas$ta4

if (nout > 0)
{
  ts_out <- ta1/nout
  ts_nor <- ta2/(n - nout)
  ts_fra <- ta3/nout
  tfd <- ta4/(n - nout)
}
if (delta == 0 || nout == 0)
{
  ts_out <- 0
  ts_nor <- ta2/(n - nout)
  ts_fra <- (1-ts_out)
  tfd <- ta4/(n - nout)
}
```

```
  rej[2,1] <- ts_out
  rej[3,2] <- ts_nor
  rej[2,2] <- ts_fra
  rej[3,1] <- tfd

  return(rej)
}

########### OGK #################

# Funcao para os testes do metodo OGK

sim_ogk <- function(dados, nout, h)
{
  rej <- matrix(0, 3, 2)
  rownames(rej) <- c("OGK","Outliers (%)",
                     "Observacoes normais (%)")
  colnames(rej) <- c("Outliers detectados (%)",
                     "Outliers nao detectados (%)")

  ts_out <- 0
  ts_nor <- 0
  ts_fra <- 0
  tfd <- 0

  ogk<- CovOgk(dados)

  W <- ogk$raw.wt

  taxas <- verificador(h,W,n)

  ta1 <- taxas$ta1
  ta2 <- taxas$ta2
  ta3 <- taxas$ta3
```

```
  ta4 <- taxas$ta4

  if (nout > 0)
  {
    ts_out <- ta1/nout
    ts_nor <- ta2/(n - nout)
    ts_fra <- ta3/nout
    tfd <- ta4/(n - nout)
  }
  if (delta == 0 || nout == 0)
  {
    ts_out <- 0
    ts_nor <- ta2/(n - nout)
    ts_fra <- (1-ts_out)
    tfd <- ta4/(n - nout)
  }


  rej[2,1] <- ts_out
  rej[3,2] <- ts_nor
  rej[2,2] <- ts_fra
  rej[3,1] <- tfd

  return(rej)
}

############------ PCOUT --------------#########################

# Funcao para os testes do metodo PCOUT

sim_pcout <- function(dados, nout, h)
{
  rej <- matrix(0, 3, 2)
  rownames(rej) <- c("PCOUT","Outliers (%)",
```

```
                    "Observacoes normais (%)")
colnames(rej) <- c("Outliers detectados (%)",
                    "Outliers nao detectados (%)")

ts_out <- 0
ts_nor <- 0
ts_fra <- 0
tfd <- 0

pcout <- pcout(dados, makeplot = FALSE)
W <- pcout$wfinal01

taxas <- verificador(h,W,n)

ta1 <- taxas$ta1
ta2 <- taxas$ta2
ta3 <- taxas$ta3
ta4 <- taxas$ta4

if (nout > 0)
{
  ts_out <- ta1/nout
  ts_nor <- ta2/(n - nout)
  ts_fra <- ta3/nout
  tfd <- ta4/(n - nout)
}
if (delta == 0 || nout == 0)
{
  ts_out <- 0
  ts_nor <- ta2/(n - nout)
  ts_fra <- (1-ts_out)
  tfd <- ta4/(n - nout)
}
```

```
  rej[2,1] <- ts_out
  rej[3,2] <- ts_nor
  rej[2,2] <- ts_fra
  rej[3,1] <- tfd

  return(rej)
}

# Funcao que realizar todas as simulacoes
# com os parametros determinados
# apresentados os resultados para todos
# os metodos

todos <- function(B,n, delta,mu,mu2,Sigma,Sigma2)
{
  REJ_MVE1 <- 0
  REJ_MCD1 <- 0
  REJ_COMED1 <- 0
  REJ_OGK1 <- 0
  REJ_PCOUT1 <- 0
  cont <- 0

  for (j in 1: B)
  {
    X <- rNCMWL(n, delta, mu, mu2, Sigma, Sigma2)
    h <- X$posicao
    nout <-X$outliers
    dados <- X$X
    p <- ncol(dados)
    n <- nrow(dados)


    if( nout >=  (n- nout))
    {
```

```
    cont <- cont + 1

  }else{
    if (n > p)
    {

    #aplicacao MVE
    REJ_MVE <- sim_mve(dados, nout, h)
    REJ_MVE1 <- REJ_MVE1+REJ_MVE
    }

    # aplicacao MCD
    REJ_MCD <- sim_mcd(dados, nout, h)
    REJ_MCD1 <- REJ_MCD1+REJ_MCD

    # aplicacao OGK
    REJ_OGK <- sim_ogk(dados, nout, h)
    REJ_OGK1 <- REJ_OGK1+REJ_OGK

    # aplicacao PCOUT
    REJ_PCOUT <- sim_pcout(dados, nout, h)
    REJ_PCOUT1 <- REJ_PCOUT1+REJ_PCOUT

    # aplicacao COMEDIAN
    REJ_COMED <- sim_comed(dados, nout, h)
    REJ_COMED1 <- REJ_COMED1+REJ_COMED
  }
}
# Evitar que o MVE receba n > p.
if (n > p)
{
 REJ_MVE1 <- REJ_MVE1/(B- cont)
}else{
  REJ_MVE1 <- 0
```

```
  }
  # Resultado
  REJ_MCD1 <- REJ_MCD1/(B- cont)
  REJ_OGK1 <- REJ_OGK1/(B- cont)
  REJ_PCOUT1 <- REJ_PCOUT1/(B- cont)
  REJ_COMED1 <- REJ_COMED1/(B- cont)

  return(list(REJ_MVE1,REJ_MCD1,REJ_OGK1,REJ_PCOUT1,REJ_COMED1))
}

#EXEMPLO
n <- 1000
delta <- 0.40
p <- 50
lambda <-0.1
m <- 5


# dados normais
mu <- c(rep(0,p))
Sigma <-diag(1,p)

# dados contaminados
mu2 <- c(rep(m,p))
Sigma2 <-  Sigma*lambda

#Interacoes
B <- 1000

todos(B,n, delta,mu,mu2,Sigma,Sigma2)
```