**Alguns métodos robustos
para detectar
outliers multivariados**

Fabíola Rocha de Santana Girollo

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO DE MESTRE
EM
CIÊNCIAS

Área de Concentração: **Estatística**

Orientador: **Profa. Dra. Lúcia Pereira Barroso**

São Paulo, março de 2008

**Alguns métodos robustos para detectar
outliers multivariados**

Este exemplar corresponde à redação
final da dissertação devidamente corrigida e
defendida por Fabíola Rocha de Santana Giroldo
e aprovada pela Comissão Julgadora.

São Paulo, 07 de março de 2008.

Banca Examinadora:

- Profa. Dra. Lúcia Pereira Barroso (presidente) - IME-USP
- Prof. Dr. Rinaldo Artes - IBMEC
- Prof. Dr. Manoel Raimundo de Sena Junior - UFPE

*À Lúcia, minha orientadora, aos meus pais
e ao Rodrigo*

Agradecimentos

Em primeiro lugar, agradeço à minha orientadora, Lúcia, por ter aceitado seguir comigo nesse desafio, o que me deixou muito feliz. Lúcia, obrigada por toda ajuda, paciência e confiança que depositou em mim em todos os momentos, por sempre ter me acalmado nas horas difíceis, me mostrando que eu posso vencer mais essa etapa da minha vida.

Obrigada aos meus professores do IME por terem colaborado, em todos os aspectos, para minha formação.

Agradeço à minha mãe, Luzia, por ter me incentivado a encarar mais esse desafio, por sempre estar ao meu lado me apoiando, me alegrando e sempre torcendo para que eu cresça cada vez mais.

Ao Rodrigo, por sempre ter dado uma solução rápida e eficiente para os meus problemas, por ter tido paciência e pelo incentivo que me deu e me dá sempre a tudo que eu vou fazer.

Agradeço ao meu pai, Celso, por sempre me admirar, por ter me ensinado a ter mais paciência com os desafios da vida, me mostrando que sou capaz de superá-los.

Ao Billy e ao Nero por toda alegria que sempre me proporcionam, principalmente nos momentos difíceis.

Aos amigos do IME, agradeço por terem suavizado os meus estudos, em especial à Paula, Dani, Roberta, Eliane e ao David.

Agradeço à SD&W pela compreensão, disponibilizando uma flexibilidade no horário.

Tia Ana, Tio José, Fred e Gabriel, agradeço pelos momentos sempre alegres em volta da mesa. Obrigada tia Esther e tio Paulo por me oferecerem ajuda e por torcerem por mim.

Enfim, agradeço a Deus pela coragem e pela força.

Resumo

Observações atípicas ou *outliers* estão quase sempre presentes em qualquer conjunto de dados, seja ele grande ou pequeno. Isso pode ocorrer por erro no armazenamento dos dados ou por existirem realmente alguns pontos diferentes dos demais.

A presença desses pontos pode causar distorções nos resultados de modelos e estimativas. Por isso, a sua detecção é muito importante e deve ser feita antes do início de uma análise mais profunda dos dados. Após esse diagnóstico, pode-se tomar uma decisão a respeito dos pontos atípicos. Uma possibilidade é corrigi-los caso tenha ocorrido erro na transcrição dos dados. Caso sejam pontos válidos, eles devem ser tratados de forma diferente dos demais, seja com uma ponderação, seja com uma análise especial.

Nos casos univariado e bivariado, o *outlier* pode ser detectado analisando-se o gráfico de dispersão que mostra o comportamento de cada observação do conjunto de dados de interesse. Se houver pontos distantes da massa de dados, eles devem ser considerados atípicos. No caso multivariado, a detecção por meio de gráficos torna-se um pouco mais complexa porque a análise deveria ser feita observando-se duas variáveis por vez, o que tornaria o processo longo e pouco confiável, pois um ponto pode ser atípico com relação a algumas variáveis e não ser com relação a outras, o que faria com que o resultado ficasse mascarado.

Neste trabalho, alguns métodos robustos para detecção de *outliers* em dados multivariados são apresentados. A aplicação de cada um dos métodos é feita para um exemplo. Além disso, os métodos são comparados de acordo com o resultado que cada um apresentar para o exemplo em questão e via simulação.

Abstract

Unusual observations or outliers are frequent in any data set, if it is large or not. Outliers may occur by typing mistake or by the existence of observations that are really different from the others.

The presence of this observations may distort the results of models and estimates. Therefore, their detection is very important and it is recommended to be performed before any detailed analysis, when a decision can be taken about these atypical observations. A possibility is to correct these observations if the problem occurred with the construction of the data set. If the observations are correct, different strategies can be adopted, with some weights or with especial analysis.

In univariate and bivariate data sets, outliers can be detected analysing the scatter plot. Observations distant from the cloud formed by the data set are considered unusual. In multivariate data sets, the detection of outliers using graphics is more difficult because we have to analyse a couple of variables each time, which results is a long and less reliable process because we can find an observation that is unusual for one variable and not unusual for the others, masking the results.

In this work, some robust methods for detection of multivariate outliers are presented. The application of each one is done for an example. Moreover, the methods are compared by the results of each one in the example and by simulation.

Sumário

1	Introdução	1
2	Conceito de <i>Outlier</i>	3
3	Métodos para Detecção de <i>Outliers</i>	16
3.1	Método Baseado no MVE (Rousseeuw e Van Zomeren, 1990)	18
3.2	Método <i>The Forward Search</i> (Atkinson e Riani, 2004)	23
3.3	Método Baseado no MCD (Filzmoser et al., 2005)	26
3.4	Método Baseado na MED (Gao et al., 2005)	30
4	Aplicação em Dados Reais	32
4.1	Análise Descritiva	32
4.2	Resultado da Aplicação do Método Baseado no MVE	42
4.3	Resultado da Aplicação do Método <i>The Forward Search</i> (FS)	42
4.4	Resultado da Aplicação do Método Baseado no MCD	46
4.5	Resultado da Aplicação do Método Baseado na MED	49
4.6	Comparação dos Métodos	49
5	Aplicação em Dados Simulados	52
5.1	Mudança na Variância dos Dados Simulados	52
5.1.1	Detalhes da Simulação Envolvendo a Distribuição $N(0, 10)$	54
5.1.2	Comparação dos Métodos para os Dados Simulados	59
5.2	Mudança na Média dos Dados Simulados	60
5.2.1	Detalhes da Simulação Envolvendo a Distribuição $N(20, 1)$	60

	vii
5.2.2 Comparação dos Métodos para os Dados Simulados	67
5.3 Resultados	69
6 Conclusões	72
A Programas	73
A.1 Método Baseado no MVE	73
A.2 Método <i>The Forward Search</i>	76
A.3 Método Baseado no MCD	78
A.4 Método Baseado na MED	79
Referências Bibliográficas	82

Capítulo 1

Introdução

Em qualquer conjunto de dados, seja ele grande ou pequeno, é necessário que se faça, antes da aplicação de qualquer técnica estatística, uma análise muito cuidadosa de seus componentes. Isso é importante, pois podem existir objetos que não condizem com a distribuição do restante dos dados, isto é, pontos que se destacam quanto a uma ou mais variáveis do estudo e que podem prejudicar a modelagem estatística dos dados da amostra e, portanto, a análise do comportamento dos mesmos.

Esses pontos atípicos são conhecidos como *outliers*. No caso em que se observa apenas uma variável para cada objeto do conjunto de dados, esse *outlier* é univariado. Já quando se trata de um conjunto de dados com duas ou mais variáveis sendo analisadas, o ponto atípico é chamado de *outlier* multivariado.

Depois de feita essa primeira análise no conjunto de dados, é possível tomar uma decisão a respeito desses pontos atípicos para que eles não comprometam o ajuste de modelos estatísticos: aplicar uma técnica que seja robusta à presença dos mesmos (ou seja, uma técnica que não sofra influência desses pontos), excluí-los da amostra ou ponderá-los com um peso diferente dos outros pontos (a idéia é que eles tenham um menor peso para que não prejudiquem a análise em questão).

No Capítulo 2, introduzimos o conceito de *outliers* através de dois exemplos. Além disso, são apresentados alguns métodos para encontrar esses pontos atípicos em conjuntos de dados univariados e bivariados.

No Capítulo 3, apresentamos quatro métodos para encontrar pontos atípicos em conjuntos de dados com mais de duas variáveis, os chamados dados multivariados. Os métodos de Rousseeuw e Van Zomeren (1990), Atkinson e Riani (2004) e Filzmoser et al. (2005) utilizam a distância de

Mahalanobis com estimadores robustos de locação e escala para detectar os *outliers* multivariados. O método de Gao et al. (2005) consiste em um procedimento iterativo que utiliza os autovalores e autovetores da matriz de covariância dos dados para detectar os *outliers* multivariados.

No Capítulo 4, apresentamos um exemplo no qual são aplicados todos os métodos de detecção de *outliers* discutidos neste trabalho. Os métodos são comparados através dos resultados da aplicação de cada um deles no conjunto de dados de interesse.

No Capítulo 5, apresentamos os resultados obtidos da aplicação de cada um dos métodos de detecção de *outliers* a dados simulados. As conclusões são dadas em seguida.

Capítulo 2

Conceito de *Outlier*

Se o interesse é encontrar pontos atípicos em um conjunto de dados que possui apenas uma variável, pode-se fazer um gráfico de dispersão que mostra o valor da variável para cada objeto da amostra. Para ficar mais claro, vamos considerar, por exemplo, o número de filhos para uma amostra composta por 10 indivíduos, como mostra a Tabela 2.1 (dados hipotéticos):

Tabela 2.1: Número de filhos por indivíduo

Indivíduo	Número de Filhos
1	2
2	1
3	3
4	1
5	4
6	12
7	3
8	2
9	1
10	20

Para esse conjunto de dados, pode-se construir o seguinte gráfico de dispersão para a visualização de pontos atípicos:

Figura 2.1: Gráfico de dispersão do número de filhos para cada indivíduo da amostra

A Figura 2.1 mostra que os indivíduos 6 e 10 são prováveis candidatos a *outlier* univariado, pois eles estão afastados dos demais pontos quando observamos a variável “Número de Filhos”.

Além do método gráfico, também é possível identificar um *outlier* univariado através do cálculo da distância Euclidiana do i -ésimo indivíduo à média do grupo, dada pela expressão:

$$ED_i = \sqrt{(x_i - \bar{x})^2} \quad (2.1)$$

em que $i = 1, \dots, n$, n é o número de indivíduos da amostra, ED_i é a distância Euclidiana para o indivíduo i , x_i é o valor da variável para o indivíduo i e $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ é a média amostral do conjunto de dados de interesse.

No exemplo da Tabela 2.1, temos que $\bar{x} = 4,9$. A partir disso, pode-se calcular a distância Euclidiana de cada indivíduo da amostra à média. A Tabela 2.2 mostra o resultado.

Observando-se a Tabela 2.2, percebe-se que os indivíduos 6 e 10 são os que possuem as maiores distâncias Euclidianas em relação à média da amostra. Isso significa que eles estão mais afastados do conjunto de dados e podem ser considerados pontos atípicos.

Como a média amostral é afetada por *outliers*, pode-se calcular a distância Euclidiana da seguinte forma para amenizar o efeito desses pontos:

1. A observação i é excluída da amostra
2. Calcula-se a média amostral sem a observação i

Tabela 2.2: Distância Euclidiana para cada indivíduo

Indivíduo	Distância Euclidiana
1	2,9
2	3,9
3	1,9
4	3,9
5	0,9
6	7,1
7	1,9
8	2,9
9	3,9
10	15,1

3. Calcula-se a distância Euclidiana para a observação i , utilizando-se a média calculada no passo 2
4. Repetem-se os passos 1, 2 e 3 para todos os indivíduos da amostra.

Para o exemplo acima, seja $\bar{x}^{(i)}$ a média amostral sem o indivíduo i . A Tabela 2.3 mostra o cálculo de $\bar{x}^{(i)}$ para $i = 1, \dots, n$.

Calcula-se, então, a distância Euclidiana de cada indivíduo i da amostra, utilizando-se a média $\bar{x}^{(i)}$. Se o indivíduo i for um ponto atípico, ao retirá-lo da amostra, a média poderá assumir um valor muito diferente (bem mais alto ou bem mais baixo) do que o da amostra completa. Portanto, a distância desse ponto será um valor alto com relação aos demais pontos.

O resultado para o conjunto de dados da Tabela 2.1 encontra-se na Tabela 2.4.

Pela Tabela 2.4, os indivíduos 6 e 10 são os que possuem maiores distâncias Euclidianas em relação à média da amostra quando excluídos do cálculo dessa média e são os que apresentam as maiores diferenças quando os valores das Tabelas 2.4 e 2.2 são comparados, ou seja, as exclusões desses pontos são as que mais afetam as distâncias Euclidianas. Esse fato é mais um indicativo de que os indivíduos 6 e 10 são prováveis candidatos a *outliers* univariados.

Outra forma de se encontrar pontos atípicos em dados unidimensionais é através da construção

Tabela 2.3: Média amostral com a exclusão do indivíduo i do conjunto de dados

i	$\bar{x}^{(i)}$
1	5,2
2	5,3
3	5,1
4	5,3
5	5,0
6	4,1
7	5,1
8	5,2
9	5,3
10	3,2

Tabela 2.4: Distância Euclidiana para cada indivíduo i da amostra após a exclusão do mesmo do conjunto de dados

Indivíduo	Distância Euclidiana
1	3,2
2	4,3
3	2,1
4	4,3
5	1,0
6	7,9
7	2,1
8	3,2
9	4,3
10	16,8

do chamado *boxplot* (Bussab e Morettin, 2006). A Figura 2.2 mostra o *boxplot* para o conjunto de dados da Tabela 2.1.

Figura 2.2: *Boxplot* para o número de filhos por indivíduo

Observando-se o gráfico apresentado na Figura 2.2, pode-se notar que os pontos 6 e 10 estão em destaque com relação aos outros pontos. Portanto, são prováveis candidatos a pontos atípicos e o *boxplot* é mais uma forma de encontrar esse tipo de observação.

No caso bivariado, podemos repetir o procedimento gráfico para encontrar os pontos atípicos do conjunto de dados de interesse.

Como ilustração, será considerado um conjunto de dados bivariados que relaciona o número médio de horas trabalhadas por dia com a idade para 20 indivíduos, de acordo com a Tabela 2.5 (dados hipotéticos).

Seguindo o mesmo raciocínio do caso univariado, pode-se construir o gráfico apresentado na Figura 2.3 para verificar a existência de pontos atípicos.

Observando-se a Figura 2.3, conclui-se que o indivíduo 10 pode ser considerado um candidato a *outlier* bivariado, pois ele está afastado da nuvem de pontos e possui valores altos para ambas as variáveis de estudo. Além disso, os indivíduos 1 e 6 também podem ser considerados pontos atípicos.

Tabela 2.5: Número médio de horas trabalhadas por dia x idade para 20 indivíduos

Indivíduo	Horas de Trabalho	Idade
1	2	60
2	1	35
3	7	32
4	2	32
5	4	41
6	15	30
7	3	40
8	3	35
9	1	30
10	20	50
11	6	39
12	5	38
13	6	41
14	5	42
15	6	41
16	5	42
17	6	42
18	5	44
19	7	45
20	8	46

Figura 2.3: Número médio de horas trabalhadas por dia x idade para 20 indivíduos

No caso do indivíduo 1, ele é um *outlier* com relação à variável “Idade”. Já no caso do indivíduo 6, ele é classificado como um ponto atípico, considerando-se a variável “Horas de Trabalho”. Possivelmente o indivíduo 3 é um *outlier*, mas não foi detectado por haver pontos mais extremos.

De maneira análoga ao caso univariado, calcula-se a distância Euclidiana para os dados bidimensionais através da expressão:

$$ED_i = \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2} \quad (2.2)$$

em que $i = 1, \dots, n$, n é o número de indivíduos da amostra, ED_i é a distância Euclidiana do indivíduo i com relação às médias das variáveis X e Y , x_i é o valor da variável X para o indivíduo i , y_i é o valor da variável Y para o indivíduo i , $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ é a média amostral da variável X e $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ é a média amostral da variável Y do conjunto de dados de interesse. No exemplo da Tabela 2.5, X é o número médio de horas trabalhadas por dia e Y é a idade.

Ao analisar a Tabela 2.6, nota-se que os indivíduos 1, 6 e 10 possuem os maiores valores para a distância Euclidiana em relação às médias das variáveis em questão.

Como já foi visto no caso univariado, a média amostral é afetada por *outliers*. Logo, deve-se calcular a distância Euclidiana da seguinte forma para amenizar o efeito desses pontos:

1. A observação i é excluída da amostra
2. Calcula-se a média amostral para as variáveis horas de trabalho e idade sem a observação i

Tabela 2.6: Distância Euclidiana para os 20 indivíduos da Tabela 2.5

Indivíduo	Distância Euclidiana
1	20,1
2	7,1
3	8,3
4	9,1
5	2,0
6	13,7
7	2,9
8	6,0
9	11,3
10	17,2
11	1,3
12	2,4
13	0,8
14	1,9
15	0,8
16	1,9
17	1,8
18	3,8
19	4,9
20	6,1

3. Calcula-se a distância Euclidiana para a observação i , utilizando-se a média calculada no passo 2

4. Repetem-se os passos 1, 2 e 3 para todos os indivíduos da amostra.

Para o exemplo acima, sejam $\bar{x}^{(i)}$ e $\bar{y}^{(i)}$ as médias amostrais das variáveis horas de trabalho e idade, respectivamente, sem o indivíduo i . A Tabela 2.7 mostra o cálculo de $\bar{x}^{(i)}$ e $\bar{y}^{(i)}$ para $i = 1, \dots, n$ e a distância Euclidiana ($ED^{(i)}$) para cada um dos indivíduos.

Observando-se a Tabela 2.7, é possível encontrar mais indícios de que os indivíduos 1, 6 e 10 são pontos atípicos deste conjunto de dados, pois quando cada um desses pontos é retirado, a distância Euclidiana deles em relação à média é maior do que a do restante dos dados.

Outra forma para se detectar *outliers* em um conjunto de dados com mais de uma variável é utilizar a distância de Mahalanobis em relação à média das variáveis, calculada para cada elemento da amostra, dada pela seguinte expressão:

$$MD_i = \sqrt{(x_i - \bar{x})^\top S_x^{-1} (x_i - \bar{x})} \quad (2.3)$$

em que $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, de dimensão p , é o vetor de médias do conjunto de dados X e $S_x = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$ é a matriz $p \times p$ de covariância amostral de X .

Para o exemplo apresentado na Tabela 2.5, calcula-se a distância de Mahalanobis para cada indivíduo. Os resultados seguem na Tabela 2.8.

Observando-se a Tabela 2.8, é possível notar que os indivíduos com as maiores distâncias de Mahalanobis são 1, 6 e 10, como já havia sido observado na Figura 2.3. Nota-se também que o indivíduo 9 apresentava alta distância Euclidiana, pois com exceção dos três *outliers* é o que mais se afasta da média. Entretanto, com base na distância de Mahalanobis ele já não se destaca, pois se localiza na nuvem de pontos típicos, seguindo a mesma relação entre horas trabalhadas e idade.

Como a distância de Mahalanobis também é uma medida afetada por pontos atípicos, pode-se repetir o procedimento realizado para a distância Euclidiana no modelo unidimensional, mas agora utilizando a distância de Mahalanobis:

1. A observação i é excluída da amostra

2. Calcula-se o vetor de média amostral e a matriz de covariância sem a observação i

Tabela 2.7: Cálculo de $\bar{x}^{(i)}$, $\bar{y}^{(i)}$ e $ED^{(i)}$ para os 20 indivíduos da Tabela 2.5

Indivíduo	$\bar{x}^{(i)}$	$\bar{y}^{(i)}$	$ED^{(i)}$
1	6,1	39,2	21,2
2	6,1	40,5	7,5
3	5,8	40,7	8,8
4	6,1	40,7	9,6
5	5,9	40,2	2,1
6	5,4	40,8	14,5
7	6,0	40,3	3,0
8	6,0	40,5	6,3
9	6,1	40,8	11,9
10	5,1	39,7	18,1
11	5,8	40,3	1,3
12	5,9	40,4	2,5
13	5,8	40,2	0,8
14	5,9	40,2	2,0
15	5,8	40,2	0,8
16	5,9	40,2	2,0
17	5,8	40,2	1,8
18	5,9	40,1	4,0
19	5,8	40,0	5,1
20	5,7	39,9	6,5

Tabela 2.8: Distância de Mahalanobis para os 20 indivíduos

Indivíduo	Distância de Mahalanobis
1	3,04
2	1,22
3	1,23
4	1,34
5	0,44
6	2,68
7	0,63
8	0,90
9	1,67
10	3,25
11	0,18
12	0,34
13	0,11
14	0,33
15	0,11
16	0,33
17	0,24
18	0,59
19	0,68
20	0,88

3. Calcula-se a distância de Mahalanobis para a observação i , utilizando-se a média e a matriz de covariância calculadas no passo 2
4. Repetem-se os passos 1, 2 e 3 para todos os indivíduos da amostra.

Para o exemplo da Tabela 2.5, seja $MD^{(i)}$ a distância de Mahalanobis em relação à média sem o indivíduo i . A Tabela 2.9 mostra o cálculo de todas essas distâncias para $i = 1, \dots, n$.

Observando-se a Tabela 2.9, confirma-se a conclusão obtida pela Figura 2.3 sobre os pontos 1, 6 e 10 do conjunto de dados que está sendo analisado. De fato, as observações 1, 6 e 10 podem ser consideradas *outliers* bivariados. Isso acontece porque esses indivíduos possuem as maiores distâncias de Mahalanobis quando são excluídos da amostra.

Quando mais do que duas variáveis são observadas para cada objeto do conjunto de dados de interesse, torna-se mais difícil a visualização dos pontos atípicos através dos métodos gráficos convencionais como, por exemplo, *boxplots* e gráficos de dispersão. Além disso, através dos exemplos anteriores, conclui-se que tanto a distância Euclidiana como a de Mahalanobis são influenciadas por pontos atípicos. Por isso, são necessários métodos robustos para a detecção de *outliers* multivariados.

Tabela 2.9: Distância de Mahalanobis para cada indivíduo i da amostra após excluí-los do conjunto de dados

Indivíduo	$MD^{(i)}$
1	4,47
2	1,30
3	1,32
4	1,45
5	0,45
6	3,53
7	0,65
8	0,94
9	1,86
10	5,16
11	0,19
12	0,35
13	0,11
14	0,34
15	0,11
16	0,34
17	0,25
18	0,61
19	0,71
20	0,92

Capítulo 3

Métodos para Detecção de *Outliers*

Neste capítulo, alguns métodos robustos para detecção de *outliers* multivariados são apresentados. São eles: método baseado no estimador robusto MVE (*Minimum Volume Ellipsoid* - Elipsóide de Volume Mínimo) de Rousseeuw e Van Zomeren (1990), método iterativo *The Forward Search* de Atkinson e Riani (2004), método baseado no estimador robusto MCD (*Minimum Covariance Determinant* - Covariância de Determinante Mínimo) de Filzmoser et al. (2005) e método iterativo baseado na MED (*Max-Eigen Difference* - Máxima Auto-Diferença) de Gao et al. (2005), baseado nos autovalores e autovetores da matriz de covariância do conjunto de dados de interesse. Todos eles, exceto o método de Gao et al. (2005), se baseiam na construção da distância de Mahalanobis robusta para detectar pontos atípicos em dados multivariados, o que produz a necessidade de que esses dados tenham distribuição aproximadamente Normal.

Define-se um conjunto de dados multivariados como sendo a matriz $X = \{x_1, \dots, x_n\}$ de dimensão $p \times n$, em que n é o número de objetos e p é o número de variáveis observadas para cada um deles. Seja $x_i, i = 1, \dots, n$, a i -ésima coluna de X , ou seja, um vetor de dimensão p .

Nos métodos clássicos, para se encontrar *outliers* em um conjunto de dados multivariados, calcula-se a distância de Mahalanobis em relação à média para cada ponto $x_i, i = 1, \dots, n$:

$$MD_i = \sqrt{(x_i - \bar{x})^\top S_x^{-1} (x_i - \bar{x})}, \quad (3.1)$$

em que $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, de dimensão p , é o vetor de médias do conjunto de dados X e $S_x = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$ é a matriz $p \times p$ de covariância amostral de X . Nesse caso, define-se o objeto i como sendo *outlier* se a distância de Mahalanobis em relação à média para esse ponto for grande se comparada com a MD calculada para os demais pontos da amostra. Se a distribuição dos

dados for aproximadamente Normal, os valores da distância de Mahalanobis ao quadrado MD_i^2 possuem aproximadamente distribuição qui-quadrado com p graus de liberdade (χ_p^2). Pode-se, então, classificar uma observação como atípica comparando-se o quadrado da sua distância de Mahalanobis com um quantil da distribuição qui-quadrado escolhido.

A desvantagem de se utilizar a distância de Mahalanobis é que $\bar{x}$ e S_x são estimadores que sofrem a influência de pontos atípicos, se estes existirem no conjunto de dados de interesse.

De acordo com Becker e Gather (1999), considerando um conjunto de dados da distribuição Normal p -variada $N(\mu, \Sigma)$, sendo $\mu \in \mathbb{R}^p$ a média e $\Sigma \in \mathbb{R}^{p \times p}$, positiva definida, a matriz de covariância dos dados, pode-se classificar um ponto como sendo atípico caso ele pertença à região α -outlier dada por:

$$out(\alpha, \mu, \Sigma) := \left\{ x \in \mathbb{R}^p : (x - \mu)^\top \Sigma^{-1} (x - \mu) > \chi_{p; 1-\alpha}^2 \right\}, \quad (3.2)$$

para $\alpha \in (0, 1)$ e sendo $\chi_{p; 1-\alpha}^2$ o quantil $(1 - \alpha)$ da distribuição qui-quadrado com p graus de liberdade, χ_p^2 . Isto significa dizer que se a observação analisada não estiver contida na elipse determinada por μ , Σ e $\chi_{p; 1-\alpha}^2$, ela pode ser considerada um α -outlier.

Então, pode-se calcular a probabilidade de uma observação do conjunto de dados X pertencer à região (3.2) através da fórmula (3.3):

$$P(x_i \in out(\alpha, \mu, \Sigma)) = \alpha \text{ para } x_i \sim N(\mu, \Sigma). \quad (3.3)$$

Ainda de acordo com Becker e Gather (1999), a região α -outlier pode ser ajustada ao tamanho da amostra. Isto pode ser feito incluindo-se a condição de que, sob o modelo, com probabilidade $1 - \alpha$, nenhuma observação pertença à região $out(\alpha_n, \mu, \Sigma)$. Em outras palavras, considerando uma amostra de tamanho n , a região $out(\alpha_n, \mu, \Sigma)$ pode ser escolhida de acordo com a seguinte condição:

$$P(x_i \in \mathbb{R}^p \setminus out(\alpha_n, \mu, \Sigma), i = 1, \dots, n) = 1 - \alpha. \quad (3.4)$$

Isso significa dizer que, considerando-se o tamanho da amostra, a probabilidade de nenhuma observação pertencer à região α_n -outlier é $(1 - \alpha) = (1 - \alpha_n)^n$, ou seja, $\alpha_n = 1 - (1 - \alpha)^{1/n}$. A expressão (3.5) representa a região α_n -outlier:

$$out(\alpha_n, \mu, \Sigma) := \left\{ x \in \mathbb{R}^p : (x - \mu)^\top \Sigma^{-1} (x - \mu) \geq c(\alpha_n, n, p) \right\}. \quad (3.5)$$

Como os parâmetros μ e Σ são desconhecidos, a região dada por (3.5) é desconhecida. Portanto, faz-se necessária a utilização de métodos de estimação robusta para a obtenção desses valores.

De acordo com o que já foi dito anteriormente, alguns desses métodos são apresentados no presente capítulo.

3.1 Método Baseado no MVE (Rousseeuw e Van Zomeren, 1990)

Como $\bar{x}$ e S_x são afetados pela presença de *outliers*, no método proposto no trabalho de Rousseeuw e Van Zomeren (1990), esses estimadores são substituídos por estimadores robustos de locação e escala, denotados, respectivamente, por $T(X)$ e $C(X)$ ou simplesmente T e C , em que X é a matriz de dados $p \times n$ a ser estudada. Calcula-se, então, a distância de Mahalanobis robusta através da expressão:

$$RD_i = \sqrt{(x_i - T(X))^T C(X)^{-1} (x_i - T(X))}. \quad (3.6)$$

Para calcular os estimadores $T(X)$ e $C(X)$, os autores sugerem o uso do estimador MVE (*Minimum Volume Ellipsoid* - Elipsóide de Volume Mínimo) introduzido por Rousseeuw (1985) e definido adiante. Sugerem também que o mesmo seja utilizado quando houver, no mínimo, cinco observações por variável, ou seja, $n/p > 5$. Mais informações sobre o estimador robusto MVE podem ser encontradas em Davies (1992).

A eficiência desses estimadores pode ser comprovada através do gráfico da Figura 3.1 (Rousseeuw e Van Zomeren, 1990, p.634) que compara a distância de Mahalanobis clássica (representada pela linha pontilhada e obtida utilizando-se as matrizes de média e de covariância amostral dos dados) com a distância de Mahalanobis robusta (representada pela linha cheia e obtida utilizando-se os estimadores MVE de locação e escala) para 28 espécies de animais. Os dados originais deste exemplo podem ser encontrados em Rousseeuw e Leroy (1987, p.57). Observando-se esse gráfico, é possível notar que a distância de Mahalanobis clássica é afetada pela presença de pontos atípicos, como as espécies 6, 14, 16, 17 e 25 desse conjunto de dados.

De acordo com Rousseeuw e Van Zomeren (1990), define-se o estimador MVE como o par $(T(X), C(X))$, sendo $T(X)$ um vetor de dimensão p e $C(X)$ uma matriz $p \times p$ positiva-semidefinida tal que seu determinante é mínimo sujeito à seguinte condição:

$$\# \left\{ i; (x_i - T(X))^T C(X)^{-1} (x_i - T(X)) \leq a^2 \right\} \geq h, \quad (3.7)$$

em que $\#$ indica o número de elementos do conjunto, $h = [(n + p + 1) / 2]$ e $[q]$ é a parte inteira de q .

Figura 3.1: Gráfico do log do peso do cérebro *versus* o log do peso do corpo para 28 espécies de animais

Esta condição garante que, dado a^2 , o número de elementos contidos na elipse determinada por $T(X)$ e $C(X)$ seja pelo menos h .

Podemos dizer, então, que $T(X)$ é o centro do elipsóide de volume mínimo que cobre pelo menos h pontos e $C(X)$ é a estrutura de covariância do mesmo elipsóide. O número a^2 é uma constante fixada, por exemplo, o quantil de ordem 0,50 da distribuição qui-quadrado com p graus de liberdade, ou seja, $\chi^2_{p;0,50}$, se é esperado que a maioria dos dados possui distribuição Normal (Rousseeuw e Van Zomeren, 1990). No caso em que se assume que os dados são uma amostra de uma distribuição elíptica $P_{\mu,\Sigma}$ com densidade $(\det \Sigma)^{-1/2} f \left[\left\{ (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}^{1/2} \right]$, uma escolha natural para a é o valor para o qual $P_{\mu,\Sigma} \left\{ (X - \mu)^\top \Sigma^{-1} (X - \mu) \leq a^2 \right\} = \int_{\|x\| \leq a} f(\|x\|) dx$ é igual a 1/2 (Lopuhaä e Rousseeuw, 1991). Essa constante não tem influência no valor de $T(X)$, que é a estimativa MVE de localização, mas determina a magnitude de $C(X)$, que é a estimativa MVE da covariância.

A obtenção do ponto de ruptura de um estimador é importante, pois, de acordo com Lopuhaä e Rousseeuw (1991), mede-se a robustez de T e C por meio do ponto de ruptura através da substituição de amostras finitas. O ponto de ruptura de um estimador de localização T de um conjunto de dados X é definido como a menor fração m/n de *outliers* que pode tornar a diferença entre as estimativas não limitada, ou seja,

$$\varepsilon^*(T, X) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{Y_m} \|T(X) - T(Y_m)\| = \infty \right\}, \quad (3.8)$$

em que o supremo é calculado para todas as possíveis amostras corrompidas Y_m que são obtidas a partir de X substituindo-se m pontos de X por valores arbitrários. O ponto de ruptura de um estimador de covariância C de um conjunto de dados X é definido como a menor fração m/n de *outliers* que pode ou deixar o maior autovalor $\lambda_1(C)$ não limitado, ou deixar o menor autovalor $\lambda_p(C)$ arbitrariamente perto de zero, isto é,

$$\varepsilon^*(C, X) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : \sup_{Y_m} D(C(X), C(Y_m)) = \infty \right\}, \quad (3.9)$$

em que o supremo é calculado para todas as possíveis amostras corrompidas Y_m como em (3.8) e $D(A, B) = \max \left\{ |\lambda_1(A) - \lambda_1(B)|, |\lambda_p(A)^{-1} - \lambda_p(B)^{-1}| \right\}$, com $\lambda_1(A) \geq \dots \geq \lambda_p(A)$ sendo os autovalores ordenados da matriz A .

Para determinar o ponto de ruptura ε^* das estimativas MVE de locação, $T(X)$, e covariância, $C(X)$, para um conjunto de dados X , com $n \geq p + 1$, consideram-se dois casos. Se $p = 1$, $\varepsilon^*(T, X) = [(n + 1)/2]/n$ e $\varepsilon^*(C, X) = [n/2]/n$, em que $\varepsilon^*(T, X)$ é o ponto de ruptura de $T(X)$ e $\varepsilon^*(C, X)$ é o ponto de ruptura de $C(X)$. Se $p \geq 2$, $\varepsilon^*(T, X) = \varepsilon^*(C, X) = [(n - p + 1)/2]/n$ (ver demonstração em Lopuhaä e Rousseeuw, 1991, pp.236 e 237).

Como estimadores robustos $T(X)$ e $C(X)$, pode-se utilizar a média ponderada e a matriz de covariância ponderada, dadas respectivamente por:

$$T_1(X) = \left(\sum_{i=1}^n w_i \right)^{-1} \sum_{i=1}^n w_i x_i \text{ e } C_1(X) = \left(\sum_{i=1}^n w_i - 1 \right)^{-1} \sum_{i=1}^n (x_i - T_1(X))(x_i - T_1(X))^{\top}, \quad (3.10)$$

em que $w_i = w(RD_i)$ depende das distâncias robustas dadas em (3.6). O ponderador $w_i = w(RD_i)$ assume valores menores para pontos da amostra com valores relativamente mais altos para a distância de Mahalanobis, os chamados *outliers*. Pode-se mostrar que $T_1(X)$ e $C_1(X)$ tem o mesmo ponto de ruptura que $T(X)$ e $C(X)$, respectivamente (seção 5 de Lopuhaä e Rousseeuw, 1991).

Para encontrar o estimador MVE, os autores propõem dois algoritmos de aproximação:

1. Algoritmo de reamostragem (Rousseeuw e Leroy, 1987): sorteiam-se sub-amostras com $p + 1$ observações diferentes, sendo cada uma delas indexada por um índice J . Para cada sub-amostra, calcula-se a média e a matriz de covariância da seguinte forma:

$$T_J = \frac{1}{p+1} \sum_{i \in J} x_i \text{ e } C_J = \frac{1}{p} \sum_{i \in J} (x_i - T_J)(x_i - T_J)^{\top}, \quad (3.11)$$

em que C_J é não singular.

Para que o elipsóide correspondente contenha exatamente h pontos, deve-se construir um fator de expansão ou contração dado por (3.12).

$$m_J^2 = \left\{ (x_i - T_J)^\top C_J^{-1} (x_i - T_J) \right\}_{h:n}. \quad (3.12)$$

O volume do elipsóide resultante é proporcional a

$$(\det(m_J^2 C_J))^{1/2} = (\det(C_J))^{1/2} (m_J)^p. \quad (3.13)$$

Isso deve ser repetido para algumas amostras e, depois disso, o elipsóide com menor função objetiva (3.13) é escolhido. Para esse melhor subconjunto J , calculamos

$$T(X) = T_J \text{ e } C(X) = (\chi_{p,0,50}^2)^{-1} m_J^2 C_J \quad (3.14)$$

como uma aproximação do estimador MVE seguido pelo passo da reponderação em (3.10), em que $\chi_{p,0,50}^2$ é a mediana da distribuição qui-quadrado com p graus de liberdade.

O número de sub-amostras necessárias deve ser tal que a probabilidade de que pelo menos uma de m sub-amostras seja “boa” é quase 1, como pode ser visto em Rousseeuw e Leroy (1987), segundo os quais uma sub-amostra é “boa” se consiste de p observações não-outliers da amostra, que contém uma fração de pontos acima de ϵ^* . A expressão para essa probabilidade, assumindo-se que n/p é grande, é

$$1 - (1 - (1 - \epsilon^*)^p)^m. \quad (3.15)$$

Como é necessário que a probabilidade (3.15) seja próxima de 1, pode-se determinar m para valores de p e ϵ^* dados (ver exemplo em Rousseeuw e Leroy, 1987, p. 198).

Para pequenas amostras, devemos ter $C(X) = (\chi_{p,0,50}^2)^{-1} c_{n,p}^2 m_J^2 C_J$. Por estudos de simulação, sabe-se que $c_{n,p}^2 = (1 + 15/(n - p))^2$ é um fator de correção razoável para pequenas amostras (Rousseeuw e Van Zomeren, 1990).

As estimativas aproximadas do MVE podem ser utilizadas como soluções iniciais como base para o desenvolvimento da reponderação. Podemos escolher w_i por meio da seguinte regra:

$$w_i = \begin{cases} 1 & \text{se } (x_i - T(X))^\top C(X)^{-1} (x_i - T(X)) \leq c \\ 0 & \text{caso contrário} \end{cases} \quad (3.16)$$

em que o valor de corte c pode ser escolhido como $\chi_{p;0,975}^2$. Agora pode-se calcular os estimadores ponderados dados em (3.10).

2. Algoritmo de projeção - é uma variante de um algoritmo de Gasko e Donoho (1982): para cada ponto x_i consideramos

$$u_i = \max_{\mathbf{v}} \frac{|x_i^\top \mathbf{v} - L(x_1^\top \mathbf{v}, \dots, x_n^\top \mathbf{v})|}{S(x_1^\top \mathbf{v}, \dots, x_n^\top \mathbf{v})} \quad (3.17)$$

sendo $\mathbf{v} = (v_1, \dots, v_n)$, em que $v_j = (x_j - T_j(X)) / |x_j - T_j(X)|$, $T_j(X)$ é o j -ésimo elemento do vetor de medianas dos dados (Gasko e Donoho, 1982). Para calcular essas estimativas, pode-se considerar qualquer conjunto de números $z_1 \leq z_2 \leq \dots \leq z_n$ e determinar a sua menor metade escolhendo-se a menor entre as diferenças $z_h - z_1, z_{h+1} - z_2, \dots, z_n - z_{n-h+1}$, sendo z_h o ponto central do conjunto de dados escolhido.

Se $z_j - z_{j-h+1}$ é a menor diferença obtida, atribuímos a L o valor do ponto médio da metade correspondente,

$$L(z_1, \dots, z_n) = (z_j - z_{j-h+1}) / 2 \quad (3.18)$$

e a S o comprimento da mesma ponderado por um fator $c(n)$

$$S(z_1, \dots, z_n) = c(n) (z_j - z_{j-h+1}), \quad (3.19)$$

no qual $c(n)$ depende do tamanho da amostra.

A expressão (3.17) corresponde à versão unidimensional da distância robusta RD_i aplicada às projeções $x_i^\top \mathbf{v}$ dos dados x_i na direção $\mathbf{v}$. Devemos selecionar algumas direções a serem testadas, já que não é possível calcular a expressão (3.17) para todas. Escolhemos todas as $\mathbf{v}$ direções da forma $x_l - M$, em que $l = 1, \dots, n$ e M é o vetor de medianas:

$$M = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_p). \quad (3.20)$$

Neste algoritmo, atualiza-se um vetor $(u_i)_{i=1, \dots, n}$ para $l = 1, \dots, n$. Os u_i 's finais são aproximações de RD_i , que podem ser plotadas ou utilizadas para reponderação como em (3.10).

Depois de se calcular o quadrado da distância de Mahalanobis robusta para cada ponto da amostra, compara-se cada um destes valores com um valor de corte escolhido. Esse valor pode ser $\chi_{p;1-\alpha}^2$, em que α é o quantil da distribuição qui-quadrado. Se o valor de RD_i^2 for maior do que esse valor de corte, define-se o ponto como *outlier* multivariado.

Uma outra possibilidade para se obter o estimador MVE para os parâmetros de localização e escala, além dos dois algoritmos de aproximação propostos pelos autores, é a função “*cov.mve*”. Esta função foi desenvolvida no *software* estatístico R (<http://www.r-project.org/>) por B. D. Ripley e faz uma busca aproximada de um subconjunto de tamanho $[(n + p + 1) / 2]$ que determina um elipsóide de menor volume. A primeira estimativa para os parâmetros de localização e escala serão, respectivamente, a média e a matriz de covariância reescalada desse subconjunto. Após isso, essa função classifica alguns pontos como sendo “bons”, isto é, pontos que pertencem à elipse de confiança de 97,5%, sob condições de normalidade. Esses pontos escolhidos como “bons” são utilizados para calcular as estimativas finais dos parâmetros de localização e escala que são, respectivamente, a média e a matriz de covariância reescalada dos mesmos.

O programa com a aplicação deste método para encontrar pontos atípicos em um conjunto de dados encontra-se no Apêndice A. Os estimadores MVE de localização e escala são utilizados para o cálculo da distância de Mahalanobis robusta para cada observação do conjunto de dados. Depois disso, esse valor ao quadrado é comparado com um quantil da distribuição qui-quadrado com p graus de liberdade, χ_p^2 , e classifica-se a observação como *outlier* multivariado se o valor da distância ao quadrado for superior ao valor do quantil escolhido. O resultado deste programa é um vetor que indica se a observação i , $i = 1, \dots, n$, é um *outlier*, caso em que a posição i deste vetor assume valor 1. Se a observação não é um ponto atípico, a posição i do vetor é zero.

3.2 Método *The Forward Search* (Atkinson e Riani, 2004)

O método proposto por Atkinson e Riani (2004) é um método gráfico que se baseia na distância de Mahalanobis e pode ser utilizado para detectar *outliers* (o que requer que os dados tenham distribuição aproximadamente Normal), encontrar grupos de observações semelhantes nos dados (os chamados *clusters*) ou alguma estrutura que não a normalidade dos dados que estão sendo estudados. Além disso, de acordo com Atkinson e Riani (2004), esse é um poderoso método estatístico robusto para explorar a relação entre dados e modelos ajustados.

De acordo com Atkinson e Riani (2004), este método pode ser utilizado também em regressão linear e não linear e modelos lineares generalizados, como pode ser encontrado em Atkinson e Riani (2000). Ainda de acordo com os autores, pode-se utilizar o método *The Forward Search* para encon-

trar transformações nos dados para satisfazer as suposições de normalidade multivariada. Atkinson e Riani (2000) usam o método para encontrar transformações satisfatórias para uma única variável, se existir, e as observações que influenciam na escolha. Riani e Atkinson (2001) utilizam o método para transformações multivariadas e Atkinson e Riani (2002) traz informações sobre transformações univariadas em regressão.

Como pode ser visto em Atkinson e Riani (2004), “este método ordena as observações por proximidade ao modelo assumido, começando por um pequeno subconjunto dos dados e aumentando o número de observações m utilizadas para o ajuste do modelo. *Outliers* e subconjuntos pequenos não identificados são incluídos no fim da busca”. Para que este método possa ser aplicado, os dados devem ter distribuição aproximadamente Normal, pois ele compara a distância de Mahalanobis ao quadrado com os quantis da distribuição qui-quadrado.

Como já foi dito anteriormente, a distância de Mahalanobis é dada pela expressão (3.1) e o seu quadrado pode ser escrito de forma análoga como:

$$d_i^2 = (y_i - \bar{x})^\top S_x^{-1} (y_i - \bar{x}), i = 1, \dots, n, \quad (3.21)$$

em que $\bar{x}$ é o vetor p -dimensional de médias das observações e S_x , de dimensão $p \times p$, é o estimador não viesado da matriz de covariância populacional.

Inicia-se o método selecionando-se m observações centrais do conjunto de dados. A partir da escolha dessas observações, os estimadores $\hat{\mu}(m)$ e $\hat{\Sigma}(m)$ de μ e Σ , respectivamente, são calculados. Depois disso, é possível obter o quadrado da distância de Mahalanobis para as n observações do conjunto de dados, conforme expressão abaixo:

$$d_i^2(m) = (y_i - \bar{x}(m))^\top S_x^{-1}(m) (y_i - \bar{x}(m)), i = 1, \dots, n. \quad (3.22)$$

As m observações do conjunto inicial podem ser escolhidas através da construção de *boxplots* robustos bivariados de Zani, Riani e Corbellini (1998). Isso é feito para excluir pontos atípicos encontrados em qualquer gráfico bidimensional do conjunto de dados. O método de construção de *boxplots* robustos bivariados é composto por três passos (Zani, Riani e Corbellini, 1998):

1. Definição da região interna análoga ao intervalo interquartílico no caso univariado. No caso bivariado, corresponde a uma região centrada em um estimador robusto de locação, que contém uma porcentagem fixada dos dados.

2. Definição do centróide robusto bivariado. Dentre as maneiras possíveis de encontrar esse centróide, Zani, Riani e Corbellini (1998) sugerem que ele seja encontrado utilizando-se a média aritmética das observações que pertencem à região interna.
3. Definição da região externa cujo contorno discrimina entre “boas” e “más” observações no conjunto de dados de interesse.

O passo seguinte consiste em um procedimento iterativo. Depois de escolhidas as m observações, o subconjunto ótimo com m observações produz n valores de $d_i^2(m)$. Ordenam-se essas distâncias ao quadrado e as $m + 1$ observações com os menores valores de $d_i^2(m)$ formarão o novo subconjunto. Embora, em geral apenas uma observação seja incluída no subconjunto em cada passo desse método, pode acontecer que um ou mais pontos dos dados sejam incluídos enquanto um ou mais são retirados (Atkinson e Riani, 2004).

Como as observações mais próximas são as que formam o novo subconjunto em cada passo do procedimento, os *outliers* e pequenos subconjuntos de dados são incluídos apenas nos passos finais. O procedimento termina quando o novo subconjunto for composto pelas n observações do conjunto de dados de interesse.

Para que a detecção de pontos atípicos no conjunto de dados seja mais eficiente, pode-se calcular as distâncias escalonadas, dadas por:

$$d_i(m) \times \{|S(m)|/|S(n)|\}^{1/2p}, \quad (3.23)$$

em que p é a dimensão das observações e $S(n)$ é a estimativa de Σ no final do procedimento. Essas distâncias escalonadas dão maior peso para os passos finais do procedimento iterativo, que são justamente aqueles nos quais serão introduzidos pontos atípicos dos dados.

Após calculadas essas distâncias de Mahalanobis escalonadas, é possível construir um gráfico que mostra, para cada observação do conjunto de dados, em cada passo do procedimento, como se comporta a distância de Mahalanobis escalonada. A partir desse gráfico, é possível detectar pontos atípicos ou *outliers* multivariados. Uma variação muito grande em algumas curvas pode ser um indício de que foi incluído um *outlier* no conjunto, naquele passo.

Como já foi dito anteriormente, o método pode também ser usado para formar agrupamentos. De acordo com Atkinson e Riani (2004), “se há *clusters*, a parte confirmatória da análise inclui

uma busca ajustando um modelo individual para cada *cluster*. Para todas as observações não classificadas, calcula-se a distância com relação ao centro de cada *cluster*. Observações são incluídas no subconjunto do *cluster* do qual elas estão mais próximas e as distâncias para todos os *clusters* são monitoradas”.

Para que um *cluster* uniforme não receba observações de um *cluster* disperso, a distância de Mahalanobis deve ser padronizada, utilizando-se a variância de cada *cluster* individualmente (Atkinson e Riani, 2004). Sendo assim, o quadrado da distância de Mahalanobis no passo m para a i -ésima observação do g -ésimo grupo é dada por:

$$d_{gi}^2(m) = \{y_{gi} - \bar{x}_g(m)\}^\top S_g^{-1}(m) \{y_{gi} - \bar{x}_g(m)\}, \quad (3.24)$$

com μ e Σ sendo estimados para cada grupo. De acordo com Atkinson e Riani (2004), o uso da distância padronizada dada pela expressão (3.25) elimina o efeito das variâncias diferentes entre cada grupo de observações.

$$d_{gi}(m) = d_{gi}^2(m) |S_{gm}|^{1/2p}. \quad (3.25)$$

A função “fwdmv.init”, que faz parte do programa “Rfwdmv”, é a implementação desse método no *software* estatístico R. O programa com a aplicação desse método encontra-se no Apêndice A e fornece, dentre outras informações, a distância de Mahalanobis (escalonada ou não, dependendo da opção que se pede na construção da função) com relação à média para cada indivíduo e qual ponto foi incluído na amostra em cada uma das etapas do método. Além disso, é possível construir, utilizando-se a função “fwdmvDistancePlot” desse programa “Rfwdmv”, o gráfico da distância de Mahalanobis para cada observação e para cada passo do processo iterativo.

Como já foi dito acima, esse é um método gráfico. Para a detecção de *outliers*, o gráfico a ser construído é a distância de Mahalanobis escalonada para cada indivíduo em cada um dos passos (entrada de uma nova observação na amostra). O ponto será considerado atípico se for constatada uma variação muito grande na distância de Mahalanobis de todos os indivíduos no momento da entrada deste ponto no conjunto, como já foi dito anteriormente.

3.3 Método Baseado no MCD (Filzmoser et al., 2005)

De acordo com Filzmoser (2005), “esse método segue a idéia de Gervini (2003) para aumentar a eficiência das estimativas robustas de locação e escala multivariadas”.

Nesse método, o procedimento MCD (*Minimum Covariance Determinant* - Covariância de Determinante Mínimo) é utilizado para determinar estimadores robustos de locação e escala, de forma a tornar a distância de Mahalanobis uma medida robusta. O estimador MCD, que é assintoticamente Normal (Butler, Davies e Jhun, 1993), é determinado pelo subconjunto de observações de tamanho h (dentre as n observações do conjunto de dados) que minimiza o determinante da matriz de covariância amostral, calculada apenas para esses h pontos. O estimador de locação é a média desses h pontos e o de escala é proporcional à matriz de covariância desses pontos. A escolha de h determina quão robusto é o estimador.

No caso em que $h \approx 0,75n$, o ponto de ruptura do estimador MCD é aproximadamente 25%, isto é, $(n - h)/n$. O ponto de ruptura é a fração de *outliers* que levará a estimadores distantes do valor verdadeiro quando ultrapassada.

Uma forma possível para encontrar os estimadores de locação e escala pelo método MCD é utilizando o algoritmo FAST-MCD proposto por Rousseeuw e Van Driessen (1999), que permanece eficiente mesmo quando se trabalha com conjuntos de dados com muitas observações e/ou muitas variáveis. Para pequenas amostras, esse algoritmo fornece estimativas exatas de locação e escala, enquanto que para conjunto de dados com maior número de observações, o resultado é mais acurado do que se comparado com outros algoritmos (Rousseeuw e Van Driessen, 1999).

Para implementar o método MCD (Rousseeuw e Van Driessen, 1999), dado um subconjunto H_t de tamanho h ou o par T_t, S_t , deve-se aplicar os seguintes passos para $t = 1, 2, 3, \dots$:

1. Calcular as distâncias de Mahalanobis $d_t(i)$, utilizando-se T_t e S_t , para $i = 1, \dots, n$
2. Ordenar as distâncias calculadas no passo anterior, o que produz uma permutação π para a qual $d_t(\pi(1)) \leq d_t(\pi(2)) \leq \dots \leq d_t(\pi(n))$.
3. Considerar $H_{t+1} := \{\pi(1), \pi(2), \dots, \pi(h)\}$.
4. Calcular $T_{t+1} := \text{média}(H_{t+1})$ e $S_{t+1} := \text{cov}(H_{t+1})$.

Como pode ser visto em Rousseeuw e Van Driessen (1999), a repetição dos passos de 1 a 4 torna o processo iterativo. Se $\det(S_{t+1}) = 0$ ou $\det(S_{t+1}) = \det(S_t)$, o processo é concluído, caso contrário, os passos de 1 a 4 são aplicados novamente ao novo conjunto de observações, que produzirão $\det(S_{t+2})$ e assim por diante. A sequência dada por $\det(S_1) \geq \det(S_2) \geq \det(S_3) \geq \dots$ é

não-negativa e, portanto, convergente (ver Teorema 1 de Rousseeuw e Van Driessen, 1999). Como a quantidade de subconjuntos de tamanho h é finita, existe um índice m para o qual $\det(S_m) = 0$ ou $\det(S_m) = \det(S_{m-1})$, o que significa que se atingiu a convergência. A convergência é atingida porque, mesmo que os passos acima fossem repetidos para (T_m, S_m) , o determinante não seria reduzido.

Rousseeuw e Van Driessen (1999) sugerem que, para diminuir o tempo computacional gasto na execução do método, sejam aplicadas apenas 2 vezes a seqüência de passos de 1 a 4 e retidos apenas os 10 melhores subconjuntos H_3 que serão utilizados até atingir a convergência.

Para que seja possível aplicar os passos de 1 a 4 apresentados anteriormente, primeiramente deve-se decidir como construir os subconjuntos H_1 iniciais. Ainda de acordo com o artigo de Rousseeuw e Van Driessen (1999), há duas possibilidades para essa escolha:

1. Selecionar aleatoriamente um subconjunto H_1 de tamanho h .
2. Selecionar aleatoriamente um subconjunto J de tamanho $p + 1$, em que p é o número de variáveis do conjunto de dados, e, então, calcular $T_0 := \text{média}(J)$ e $S_0 := \text{cov}(J)$. Se $\det(S_0) = 0$, adicionam-se observações ao subconjunto J até que $\det(S_0) > 0$. Então, calculam-se as distâncias $d_0^2(i) := (x_i - T_0)^\top S_0^{-1} (x_i - T_0)$ para $i = 1, \dots, n$. Depois disso, essas distâncias são ordenadas de tal forma que $d_0(\pi(1)) \leq \dots \leq d_0(\pi(n))$ e $H_1 := \{\pi(1), \dots, \pi(h)\}$.

Uma forma de diminuir o tempo gasto para obter o estimador MCD é dividir o conjunto com todas as observações em subconjuntos para que não seja necessário fazer todos os cálculos para os dados como um todo. Por exemplo, para $n > 1500$, o algoritmo divide o conjunto de dados em 5 subconjuntos de tamanho 300 (em que as observações são escolhidas uma a uma de forma aleatória, sem reposição), que não se sobrepõem e, juntos, formam o conjunto de tamanho 1500, que é um subconjunto do conjunto de dados de tamanho n . Ver maiores detalhes em Rousseeuw e Van Driessen (1999).

Aplicando-se o método MCD, obtemos a distância de Mahalanobis robusta dada por (3.6), em que $T(X)$ e $C(X)$ são os estimadores de locação e escala, respectivamente, obtidos por esse método.

Como já foi dito anteriormente, pode-se comparar o quadrado da distância de Mahalanobis robusta com um determinado quantil da distribuição qui-quadrado com p graus de liberdade (χ_p^2), em que p é o número de variáveis que estão sendo estudadas. Se a distância for maior do que esse quantil, o ponto é um candidato a *outlier*. É possível construir, a partir do quantil escolhido, uma

Figura 3.2: Gráfico do $\log(\text{Be})$ e $\log(\text{Sr})$

elipse de tolerância para verificar quais pontos não estão sendo cobertos pela mesma. Esses pontos serão considerados pontos atípicos.

Por meio de um exemplo real de Filzmoser et al. (2005), reproduzido na Figura 3.2 mostra-se que a elipse de tolerância construída com os estimadores de locação e escala obtidos pelo método MCD não sofre influência de pontos atípicos. Já a elipse construída utilizando-se a média e a matriz de covariância amostral como estimadores de locação e escala, respectivamente, é bastante influenciada por esses pontos.

De acordo com Filzmoser et al. (2005), o fato de classificar observações como *outliers* utilizando-se um quantil da distribuição qui-quadrado é algo subjetivo, pois deve-se escolher qual é o percentual dos dados (valor obtido pelo quantil da distribuição qui-quadrado) que não está de acordo com a distribuição do conjunto completo dos dados. Por isso, o gráfico qui-quadrado é um procedimento mais adequado para encontrar pontos atípicos do que simplesmente a escolha de um quantil da qui-quadrado. Através dele, é possível visualizar, nas caudas, o desvio da distribuição dos dados com relação à normalidade multivariada. Esse desvio é calculado utilizando-se a seguinte expressão:

$$p_n(\delta) = \sup_{u \geq \delta} (G(u) - G_n(u))^+, \quad (3.26)$$

em que $G(u)$ é a função de distribuição da χ_p^2 , $G_n(u)$ é a função de distribuição empírica do quadrado da distância de Mahalanobis robusta RD_i^2 e $p_n(\delta)$ é a diferença entre a distribuição empírica e a teórica

para um quantil δ da $\chi^2_{p;1-\alpha}$ para α pequeno. Tomam-se apenas as diferenças positivas, pois, nesses casos, a função de probabilidade acumulada da distribuição dos dados está “abaixo” da função de probabilidade acumulada da χ^2 , isto é, atinge o valor 1 em um quantil mais alto do que a distribuição χ^2 , o que significa que os dados contêm *outliers*. Assim, $p_n(\delta)$ mede a distância da distribuição empírica com relação à teórica apenas nas caudas, definidas pelo valor de δ . Essa comparação pode ser feita, pois $G_n(u)$ converge para $G(u)$ para amostras com distribuição Normal multivariada.

Compara-se, então, $p_n(\delta)$ obtido em (3.26) com um valor crítico $p_{crit}(\delta, n, p)$, definido como certo quantil $(1 - \alpha)$ de todos os valores de $p_n(\delta)$ (originados de amostras simuladas, de tamanho n , da distribuição Normal padrão p -variada) para um pequeno valor de α , por exemplo, $\alpha = 0,05$. Se o valor $p_n(\delta)$ for superior ao valor $p_{crit}(\delta, n, p)$, a observação é definida como *outlier* multivariado. Define-se, então, uma medida de *outliers* na amostra:

$$\alpha_n(\delta) = \begin{cases} 0 & \text{se } p_n(\alpha) \leq p_{crit}(\delta, n, p) \\ p_n(\delta) & \text{se } p_n(\alpha) > p_{crit}(\delta, n, p) \end{cases} \quad (3.27)$$

O valor limite é determinado como $c_n(\delta) = G_n^{-1}(1 - \alpha_n(\delta))$.

Para a implementação desse método, utiliza-se o programa “mvoutlier”, que está disponível no *software* estatístico R. O programa com o desenvolvimento desse método encontra-se no Apêndice A.

3.4 Método Baseado na MED (Gao et al., 2005)

O método proposto por Gao et al. (2005) para detectar *outliers*, chamado *Max-Eigen difference*, baseia-se nos autovalores e autovetores da matriz de covariância do conjunto de dados de interesse.

Considerando um conjunto de dados X de dimensão $p \times n$, pode-se construir $X^{(i)}$ como sendo o conjunto de dados no qual a i -ésima observação foi excluída. O procedimento proposto consiste na aplicação dos seguintes passos:

1. Calcular a média amostral $\bar{x}$, a matriz de covariância S de X , os autovalores $\lambda_1 \geq \dots \geq \lambda_p$ e os correspondentes autovetores $v_1, \dots, v_p$ de S .
2. Calcular a matriz de covariância amostral $S^{(i)}$ de $X^{(i)}$, os autovalores $\lambda_1^{(i)} \geq \dots \geq \lambda_p^{(i)}$ e os correspondentes autovetores $v_1^{(i)}, \dots, v_p^{(i)}$ de $S^{(i)}$.

3. Calcular $d_i = \left\| \lambda_1^{(i)} \mathbf{v}_1^{(i)} - \lambda_1 \mathbf{v}_1 \right\| \left(1 - \prod_{j=1}^p I_{\{y_{ij}^2 < \lambda_j\}} \right)$, em que $\|\cdot\|$ denota a norma euclidiana, $I_{\{\cdot\}}$ é uma função indicadora, $y_{ij} = (x_i - \bar{x})^\top \mathbf{v}_j$.
4. Calcular os d_i 's padronizados como $MED_i = \frac{d_i}{\sum_{j=1}^n d_j}$.

MED_i é chamado *Max-Eigen difference* para a i -ésima observação. O gráfico que mostra o valor de MED para cada observação permite encontrar *outliers* visualmente, como é feito no caso univariado. Isso ocorre, pois observações com valores grandes de MED ficarão distantes quando comparadas com as demais, neste gráfico, o que faz com que sejam classificadas como pontos atípicos ou *outliers*.

De acordo com Gao et al. (2005), a existência de *outliers* na direção próxima ao primeiro autovetor da matriz de covariância causará mudanças no primeiro autovalor e a existência de pontos atípicos nas outras direções pode mudar tanto o primeiro autovalor como seu correspondente autovetor e, portanto, em qualquer um dos casos, $\lambda_1 \mathbf{v}_1$ será uma medida profundamente influenciada por *outliers*.

Para contornar essa situação, o Teorema 2.1 (Gao et al (2005), p.1859) fornece dois casos em que a observação x_i possui valor alto de MED : o primeiro em que $|y_{i1}^2 - \lambda_1|$ é grande e o segundo em que y_{ik}^2 é grande para algum $k \geq 2$, especialmente para valores pequenos de k , já que valores pequenos de k correspondem a diferenças $\lambda_1 - \lambda_k$ pequenas.

A aplicação desse método pode ser feita utilizando-se o programa desenvolvido no *software* estatístico R, que produz um gráfico com o valor de MED_i para $i = 1, \dots, n$. A observação é considerada um *outlier* se o seu valor de MED for muito diferente dos outros pontos do conjunto de dados.

Capítulo 4

Aplicação em Dados Reais

Neste capítulo, todos os métodos apresentados neste trabalho são comparados através da aplicação de cada um deles a um exemplo que pode ser encontrado em Draper e Smith (1980 - p.404), cujos dados estão exibidos na Tabela 4.1. Nesse exemplo, são analisadas 20 observações com relação a 5 variáveis. Neste trabalho, os dados das observações 4, 6, 8 e 19 foram modificados, como feito por Rousseeuw e Leroy (1987, p.243), e podem ser encontrados na Tabela 4.2.

Cada observação corresponde a uma amostra de madeira talhada de pinho. As variáveis X_1 , X_2 , X_3 , X_4 , X_5 e y (que não será utilizada na aplicação dos métodos) são, respectivamente, o número de filamentos/ mm^2 em madeira primaveril, o número de filamentos/ mm^2 em madeira de verão, a porcentagem de madeira primaveril, a porcentagem de absorção de luz em madeira primaveril, a porcentagem de absorção de luz em madeira de verão e a gravidade específica da madeira. Nesses dados modificados, a variável X_1 está dividida por 1.000, X_2 por 10.000 e X_3 , X_4 e X_5 por 100.

4.1 Análise Descritiva

Como este é um conjunto de dados multivariados, a detecção de *outliers* por meio de gráficos se torna ineficiente já que é possível analisar apenas duas variáveis ao mesmo tempo. Isso pode ser comprovado observando-se os *boxplots* para cada variável e os gráficos bidimensionais encontrados a seguir para os dados modificados.

Tabela 4.1: Dados de fatores anatômicos e gravidade específica da madeira em pinheiro talhado

Objeto	X_1	X_2	X_3	X_4	X_5	y
1	573	1059	46,5	53,8	84,1	0,534
2	651	1356	52,7	54,5	88,7	0,535
3	606	1273	49,4	52,1	92,0	0,570
4	630	1151	48,9	50,3	87,9	0,528
5	547	1135	53,1	51,9	91,5	0,548
6	557	1236	54,9	55,2	91,4	0,555
7	489	1231	56,2	45,5	82,4	0,481
8	685	1564	56,6	44,3	91,3	0,516
9	536	1182	59,2	46,4	85,4	0,475
10	685	1564	63,1	56,4	91,4	0,486
11	664	1588	50,6	48,1	86,7	0,554
12	703	1335	51,9	48,4	81,2	0,519
13	653	1395	62,5	51,9	89,2	0,492
14	586	1114	50,5	56,5	88,9	0,517
15	534	1143	52,1	57,0	88,9	0,502
16	523	1320	50,5	61,2	91,9	0,508
17	580	1249	54,6	60,8	95,4	0,520
18	448	1028	52,2	53,4	91,8	0,506
19	476	1057	42,9	53,2	92,9	0,595
20	528	1057	42,4	56,6	90,0	0,568

Tabela 4.2: Dados modificados em gravidade específica da madeira

Objeto	X_1	X_2	X_3	X_4	X_5	y
1	0,5730	0,1059	0,4650	0,5380	0,8410	0,5340
2	0,6510	0,1356	0,5270	0,5450	0,8870	0,5350
3	0,6060	0,1273	0,4940	0,5210	0,9200	0,5700
4	0,4370	0,1591	0,4460	0,4230	0,9920	0,4500
5	0,5470	0,1135	0,5310	0,5190	0,9150	0,5480
6	0,4440	0,1628	0,4290	0,4110	0,9840	0,4310
7	0,4890	0,1231	0,5620	0,4550	0,8240	0,4810
8	0,4130	0,1673	0,4180	0,4300	0,9780	0,4230
9	0,5360	0,1182	0,5920	0,4640	0,8540	0,4750
10	0,6850	0,1564	0,6310	0,5640	0,9140	0,4860
11	0,6640	0,1588	0,5060	0,4810	0,8670	0,5540
12	0,7030	0,1335	0,5190	0,4840	0,8120	0,5190
13	0,6530	0,1395	0,6250	0,5190	0,8920	0,4920
14	0,5860	0,1114	0,5050	0,5650	0,8890	0,5170
15	0,5340	0,1143	0,5210	0,5700	0,8890	0,5020
16	0,5230	0,1320	0,5050	0,6120	0,9190	0,5080
17	0,5800	0,1249	0,5460	0,6080	0,9540	0,5200
18	0,4480	0,1028	0,5220	0,5340	0,9180	0,5060
19	0,4170	0,1687	0,4050	0,4150	0,9810	0,4010
20	0,5280	0,1057	0,4240	0,5660	0,9090	0,5680

Analisando-se os gráficos apresentados nas Figuras 4.2, 4.5, 4.10 e 4.11, os pontos 4, 6, 8 e 19, localizados na parte superior esquerda dos mesmos, podem ser considerados *outliers* bivariados para as variáveis X_1 e X_2 , X_1 e X_5 , X_3 e X_5 , X_4 e X_5 , respectivamente. Esses mesmos pontos estão em destaque com relação aos demais nos gráficos bivariados das Figuras 4.3, 4.4 e 4.9, na parte inferior esquerda, que correspondem às variáveis X_1 e X_3 , X_1 e X_4 , X_3 e X_4 , respectivamente. Para o gráfico bivariado X_2 e X_3 da Figura 4.6, esses pontos estão em destaque na parte inferior direita. Ainda com relação à Figura 4.2, os pontos 10 e 11 localizados na parte superior direita, estão distante da distribuição do restante dos pontos.

As observações 10 e 13 estão distantes da massa de dados nos gráficos bivariados das Figuras 4.3, 4.6 e 4.9, localizando-se na parte superior direita dos mesmos. Isso é um indício de que esses pontos sejam *outliers* bivariados.

Nos gráficos de dispersão bivariados para as variáveis X_2 e X_4 , X_2 e X_5 , Figuras 4.7 e 4.8, respectivamente, as observações 4, 6, 8, 10, 11 e 19 estão destacadas com relação às demais, podendo ser consideradas pontos atípicos deste conjunto de dados.

Na maior parte dos gráficos, as observações 4, 6, 8 e 19 mostraram-se distantes da massa de dados. Isso faz sentido já que foram essas as observações modificadas com relação ao conjunto de dados original.

Figura 4.1: Boxplot para a variável X_1 a X_5 dos dados modificados

Observando-se os *boxplots* da Figura 4.1 para cada uma das variáveis isoladamente, não é

possível detectar nenhum ponto atípico. Mesmo nos gráficos bidimensionais é possível ter uma idéia de quais observações são candidatas a *outliers* multivariados, mas não é possível ter certeza sobre essa suspeita. Isso porque não conseguimos analisar todas as variáveis de uma só vez.

Como já foi mencionado anteriormente, utilizar a distância de Mahalanobis calculada com a média e a matriz de covariância dos dados também não é muito eficiente para detectar *outliers*. Para os dados da Tabela 4.2, obtemos o gráfico da Figura 4.12, que mostra um resultado distorcido por conta da presença de *outliers* multivariados no conjunto de dados.

Dada a dificuldade de se encontrar pontos atípicos visualmente em dados com mais de duas variáveis, como é o caso deste exemplo, os quatro métodos estudados neste trabalho são aplicados e comparados.

Figura 4.2: Gráfico de dispersão X_1 e X_2

Figura 4.3: Gráfico de dispersão X_1 e X_3 Figura 4.4: Gráfico de dispersão X_1 e X_4

Figura 4.5: Gráfico de dispersão X_1 e X_5 Figura 4.6: Gráfico de dispersão X_2 e X_3

Figura 4.7: Gráfico de dispersão X_2 e X_4 Figura 4.8: Gráfico de dispersão X_2 e X_5

Figura 4.9: Gráfico de dispersão X_3 e X_4 Figura 4.10: Gráfico de dispersão X_3 e X_5

Figura 4.11: Gráfico de dispersão X_4 e X_5

Figura 4.12: Distância de Mahalanobis para os dados da Tabela 4.2, utilizando a média e a matriz de covariância dos dados

4.2 Resultado da Aplicação do Método Baseado no MVE

Nesta seção, o estimador MVE (obtido através da função “cov.mve” desenvolvida no *software* estatístico R) será utilizado para a obtenção de estimativas robustas para os parâmetros de locação e escala no exemplo apresentado na Tabela 4.2. Feito isso, a distância de Mahalanobis robusta para cada indivíduo é calculada substituindo-se a média e a matriz de covariância dos dados pelos estimadores MVE. A Tabela 4.3 mostra o quadrado da distância de Mahalanobis robusta.

Como já foi dito anteriormente, a distância de Mahalanobis robusta permite encontrar pontos atípicos em um conjunto de dados comparando-se o seu valor ao quadrado para cada uma das observações com um quantil da distribuição qui-quadrado escolhido. Neste caso, o quantil escolhido foi 95%, ou seja, $\chi^2_{5;0,95}$ cujo valor é 11,07. O ponto será classificado como *outlier* multivariado se a distância de Mahalanobis robusta ao quadrado para esse ponto for maior do que o quantil da distribuição qui-quadrado.

Seguindo a idéia apresentada em Rousseeuw e Van Zomeren (1990) e utilizando o programa em R disponibilizado no Apêndice A, para o conjunto de dados da Tabela 4.2, classificamos como *outliers* multivariados as unidades 3, 4, 6, 7, 8, 11, 16 e 19, que são as que possuem valores de RD_i^2 acima do quantil escolhido, como pode ser comprovado observando-se a Tabela 4.3.

Além dos pontos que já eram esperados ser detectados como atípicos, as observações 3, 7, 11 e 16 também foram classificadas como tal. É importante lembrar que, se outro quantil da distribuição qui-quadrado for utilizado, esse resultado pode ser um pouco diferente. Por exemplo, se utilizarmos o quantil 0,98 (13,39), as observações atípicas serão as de número 4, 6, 7, 8, 11, 16 e 19.

4.3 Resultado da Aplicação do Método *The Forward Search* (FS)

Como já foi mencionado anteriormente, a aplicação desse método para detecção de *outliers* multivariados será feita através do programa desenvolvido em R e que se encontra no Apêndice A.

Para o exemplo da Tabela 4.2, aplica-se a função “fwdmv.init”, sendo que o conjunto inicial de observações é escolhido através de *boxplots* robustos bivariados (Zani, Riani e Corbellini, 1998) e é composto pelas seguintes observações: 2, 5, 9, 11, 12 e 13.

Depois de escolhido o conjunto inicial, pode-se obter quais observações são incluídas a cada passo do processo iterativo. Isso pode ser visto no gráfico da Figura 4.13, que corresponde à função

Tabela 4.3: Distância de Mahalanobis robusta ao quadrado para os 20 indivíduos da Tabela 4.2

Objeto	RD_i^2
1	5,81
2	0,66
3	12,39
4	7080,87
5	2,00
6	9190,41
7	16,29
8	10649,90
9	4,23
10	8,34
11	193,05
12	11,07
13	3,35
14	2,23
15	1,35
16	16,52
17	3,25
18	4,21
19	13449,48
20	4,54

“fwdmvEntryPlot” do programa “Rfwdmv” desenvolvido no R. O comprimento do retângulo branco indica o momento em que a observação é incluída no subconjunto. No primeiro passo, por exemplo, a observação incluída é a 3, no segundo, a 10, no terceiro, a 14 e assim por diante, sendo que o tamanho do conjunto passa a ser 7, 8 e 9 com a entrada desses pontos, respectivamente. A Tabela 4.4 traz essa informação para cada passo do procedimento iterativo.

Figura 4.13: Ordem de entrada das observações

Além disso, também é possível verificar como se comporta a distância de Mahalanobis de cada observação em cada passo do método. A função a ser utilizada para a construção desse gráfico no *software* estatístico R é a “fwdmvDistancePlot” do programa já mencionado “Rfwdmv”. O resultado para o exemplo da Tabela 4.2 encontra-se na Figura 4.14. De acordo com esse gráfico, nota-se que os pontos 4, 6, 8 e 19, com linhas mais escuras, são os que possuem as maiores distâncias de Mahalanobis em quase todo o processo iterativo. Isso deixa de ser verdade a partir do passo em que o subconjunto atinge tamanho 17, no qual a observação 4 é incluída no conjunto. Com isso, concluímos que essas observações podem ser consideradas atípicas pelo método de Atkinson e Riani (2004), já que com a entrada de uma delas, a distância de Mahalanobis de todas têm um decréscimo evidente.

Os gráficos da Figura 4.15 foram obtidos através da função “fwdmvMinmaxPlot” do programa

Tabela 4.4: Observação incluída na amostra em cada passo do método de Atkinson e Riani (2004)

Passo	Observação Incluída
2	3
3	10
4	14
5	17
6	20
7	1
8	15
9	18
10	7
11	16
12	4
13	6
14	8
15	19

“Rfwdmv”. O primeiro mostra duas linhas, sendo uma delas a distância de Mahalanobis máxima entre as unidades contidas no subconjunto e a outra a m -ésima distância de Mahalanobis ordenada em toda a amostra. O segundo mostra também duas linhas, sendo uma delas a distância de Mahalanobis mínima entre as unidades fora do subconjunto e a outra a $(m + 1)$ -ésima distância de Mahalanobis ordenada em toda a amostra. Essas duas curvas são quase sempre coincidentes nos dois gráficos. Quando há um pico na curva da distância de Mahalanobis, significa que um *outlier* foi incluído na amostra. Observando-se esses gráficos, pode-se perceber que esse pico ocorre no momento em que a amostra é composta por 17 observações do conjunto de dados. Esse é o passo 12 no qual a observação 4 é incluída, ou seja, podemos classificá-la como um ponto atípico. Como não há outro pico, as próximas observações incluídas também podem ser consideradas *outliers* multivariados.

Figura 4.14: Distância de Mahalanobis

4.4 Resultado da Aplicação do Método Baseado no MCD

De acordo com Filzmoser et al. (2005), a aplicação deste método pode ser feita através de funções do programa “mvoutlier” desenvolvido no *software* estatístico R. Para exemplificar os seus resultados, utilizamos novamente os dados da Tabela 4.2.

A primeira função a ser utilizada é a “aq.plot”. A Figura 4.16 mostra quatro gráficos obtidos através desse comando no *software* estatístico R. Essa função também produz um vetor (“\$outliers”) que indica se a observação é *outlier* (caso em que o vetor nessa posição recebe “TRUE”) ou não (caso em que o vetor recebe “FALSE” na posição correspondente a esse ponto). Para o exemplo da Tabela 4.2, esse vetor indica as observações 4, 6, 8, 11, 16 e 19 como *outliers* multivariados.

Analisando agora a Figura 4.16, o primeiro gráfico é o diagrama de dispersão das duas primeiras componentes principais para o conjunto de dados da Tabela 4.2. O segundo traz a distância de Mahalanobis robusta ao quadrado ordenada das observações e a distribuição empírica da função de MD_i^2 . O terceiro mostra os *outliers* (pontos destacados em vermelho) detectados quando se compara o quadrado da distância de Mahalanobis robusta (obtida utilizando os estimadores MCD) com certo quantil da distribuição qui-quadrado, nesse caso, 95%. O último detecta pontos atípicos (destaca-

Figura 4.15: Distância de Mahalanobis mínima e máxima

dos em vermelho) observando-se as caudas da distribuição, a partir de certo quantil da qui-quadrado, como está descrito em Filzmoser et al. (2005).

Pelo que foi dito acima e observando o segundo gráfico da Figura 4.16, no qual comparamos a distância de Mahalanobis robusta ao quadrado com a distribuição empírica de MD_i^2 , vemos que os pontos 4, 6, 8 e 19 são atípicos. Quando comparamos o quadrado da distância de Mahalanobis robusta com o quantil 95% da $\chi^2_{5;0,95}$, gráfico “*Outliers* baseados no quantil 95%”, é possível notar que os pontos em destaque são 4, 6, 7, 8, 11, 16 e 19. As mesmas observações são detectadas no gráfico “*Outliers* baseados no quantil ajustado”.

Outra função deste programa que pode ser utilizada no diagnóstico de pontos atípicos é a “dd.plot”. O gráfico obtido com esse comando pode ser visto na Figura 4.17 e é o diagrama de dispersão entre as distâncias de Mahalanobis clássica e robusta. Esta função também fornece um vetor que indica se a observação é *outlier* ou não. De acordo com esse vetor e com o gráfico, as observações 4, 6, 8, 11 e 19 são consideradas *outliers*.

Figura 4.16: Gráficos obtidos pelo “aq.plot”

Figura 4.17: Gráfico obtido pelo “dd.plot”

4.5 Resultado da Aplicação do Método Baseado na MED

De acordo com o que já foi dito, o procedimento de Gao et al. (2005) é iterativo e é um método gráfico para a detecção de *outliers* multivariados. Para exemplificar a sua utilização, o programa com esse método, que pode ser encontrado no Apêndice A, foi aplicado aos dados da Tabela 4.2. O resultado disso é o gráfico da Figura 4.18, no qual é possível verificar que os pontos que mais se distanciam dos demais são o 8 e o 19. Além disso, as observações 4, 6, 10 e 12 também estão em destaque. Todos esses pontos podem ser classificados como atípicos por esse método.

Figura 4.18: MED para cada observação do conjunto de dados

4.6 Comparação dos Métodos

Como todos os métodos apresentados neste trabalho foram aplicados nos dados da Tabela 4.2, é possível fazer uma comparação entre eles para verificar quais observações foram classificadas como *outliers* multivariados e são de fato pontos atípicos, e quais observações foram classificadas como *outliers* e não o são, sabendo que os pontos atípicos desse conjunto são as observações 4, 6, 8 e 19 (como pode ser visto em Rousseeuw e Leroy, 1987, p.243).

A Tabela 4.5 é um resumo dos resultados da aplicação dos quatro métodos. Nessa tabela, as caselas preenchidas por 1 significam que a respectiva observação foi classificada como *outlier* multivariado.

Analisando-se a Tabela 4.5, é possível verificar que as observações classificadas como *outliers* multivariados por todos os métodos citados foram a 4, a 6, a 8 e a 19, como já era esperado, ou seja, nenhum método deixou de detectar esses pontos como atípicos. Por outro lado, algumas observações foram erroneamente classificadas como *outliers* multivariados quando, na verdade, elas não o são.

No método MVE de Rousseeuw e Van Zomeren (1990), os pontos 3, 7, 11 e 16 foram classificados como atípicos, mas não são. Já no método que utiliza o estimador robusto MCD (Filzmoser et al. 2005), as observações classificadas de forma incorreta foram: 7, 11 e 16. Por fim, o método de Gao et al. (2005) detectou os pontos 10 e 12 como atípicos, mas, de acordo com o conjunto de dados originais, eles não são. O método de Atkinson e Riani (2004) foi o único que não classificou como *outlier* outra observação além das quatro esperadas.

É importante salientar que é possível a existência de outros *outliers* no conjunto de dados, além dos quatro modificados. As observações 7, 11 e 16 poderiam ser *outliers* que não foram detectadas por todos os métodos.

Tabela 4.5: Resultado e comparação da aplicação dos métodos no conjunto de dados da Tabela 4.2

Observação	MVE	FS	MCD	MED
1	0	0	0	0
2	0	0	0	0
3	1	0	0	0
4	1	1	1	1
5	0	0	0	0
6	1	1	1	1
7	1	0	1	0
8	1	1	1	1
9	0	0	0	0
10	0	0	0	1
11	1	0	1	0
12	0	0	0	1
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	1	0	1	0
17	0	0	0	0
18	0	0	0	0
19	1	1	1	1
20	0	0	0	0

Capítulo 5

Aplicação em Dados Simulados

Neste capítulo, os métodos de detecção de *outliers* multivariados apresentados neste trabalho são aplicados a dados obtidos através de simulação. Depois de aplicados nos dados simulados, todos os métodos são comparados, seguindo a idéia do que foi apresentado no Capítulo 4.

Foram feitas duas simulações da distribuição Normal, na primeira delas mantendo a média constante e modificando a variância e na segunda, o inverso. Os dados simulados são compostos por 30 observações e 6 variáveis, sendo as 25 primeiras geradas da distribuição Normal com média zero e variância um, $N(0, 1)$, e as 5 últimas de outra distribuição Normal. Espera-se que as cinco últimas observações sejam detectadas pelos métodos como *outliers* multivariados, pois possuem distribuição diferente da maior parte dos dados.

5.1 Mudança na Variância dos Dados Simulados

Nesta seção, as 25 primeiras observações são simulações da distribuição Normal univariada com média igual a zero e variância igual a um, isto é, $N(0, 1)$ para cada uma das variáveis. As demais observações são simulações da distribuição Normal univariada com média igual a zero e variância igual a dez, isto é, $N(0, 10)$ para as 6 variáveis. Estes dados podem ser vistos na Tabela 5.1.

Tabela 5.1: Dados simulados com a variância modificada

Observação	X_1	X_2	X_3	X_4	X_5	X_6
1	-2,373	-0,254	1,220	-1,321	-0,634	0,710
2	0,177	1,607	0,182	0,718	-1,962	-0,423
3	1,129	-1,009	-0,758	0,204	-0,959	2,513
4	-0,221	-0,520	-0,516	1,233	-0,720	-0,903
5	0,071	0,373	-1,404	-1,574	-0,776	-2,257
6	1,852	1,472	-1,303	-1,354	-0,214	0,999
7	-0,470	0,057	1,102	1,413	1,203	-1,296
8	-0,587	0,250	-0,007	-0,476	0,042	0,590
9	-0,055	-0,758	-0,147	-1,087	1,679	-1,266
10	0,215	1,149	0,262	0,893	-0,949	-0,690
11	-0,796	0,142	-2,125	0,038	-0,967	-1,256
12	-0,524	-0,945	0,008	-1,151	-1,371	1,042
13	0,153	-1,581	-0,339	0,399	0,035	0,935
14	0,647	-1,519	0,782	-0,474	0,571	0,920
15	-0,273	1,230	0,124	-0,711	-0,514	-0,734
16	-1,136	-0,772	-1,399	-1,054	0,228	1,464
17	0,482	0,220	-0,844	-2,127	-0,718	-0,451
18	1,592	0,726	0,318	1,083	0,136	0,237
19	-1,169	0,638	1,370	1,960	-1,088	1,068
20	0,261	1,364	1,181	1,150	-0,199	-0,531
21	-1,171	0,826	0,947	0,839	-1,383	-0,605
22	-0,348	0,900	1,211	-1,526	0,451	-0,200
23	-0,599	0,135	1,217	1,127	0,021	0,478
24	-0,553	1,636	-0,606	-1,018	0,373	-0,609
25	2,021	0,894	-0,549	0,189	-2,069	1,323
26	4,468	12,094	-6,318	13,821	10,577	1,664
27	-1,425	-8,933	-11,609	1,636	-6,635	4,451
28	-4,210	9,057	-15,113	3,477	-9,174	4,841
29	1,255	16,990	1,830	-2,234	-13,796	5,658
30	-19,445	5,893	5,223	4,365	-2,690	10,350

5.1.1 Detalhes da Simulação Envolvendo a Distribuição $N(0, 10)$

Resultado da Aplicação do Método Baseado no MVE

Utilizando-se o quantil 0,95 da distribuição qui-quadrado, $\chi^2_{6;0,95}$ (12,59), as observações classificadas como *outliers* multivariados são: 1, 5, 6, 9, 11, 14, 16, 25, 26, 27, 28, 29 e 30. Se passamos a utilizar o quantil 0,98 da distribuição qui-quadrado, $\chi^2_{6;0,98}$ (15,03), algumas observações deixam de ser atípicas e as classificadas como *outliers* multivariados são as seguintes: 1, 5, 6, 11, 16, 26, 27, 28, 29 e 30. As distâncias de Mahalanobis robustas ao quadrado encontram-se na Tabela 5.2.

Para ambos os quantis utilizados, as observações que possuem distribuição diferente da maior parte dos dados foram classificadas como atípicas, mas observações que não eram para ser classificadas como *outliers* multivariados foram classificadas erroneamente.

Resultado da Aplicação do Método *The Forward Search*

Como este é um método iterativo, primeiramente escolhe-se o conjunto inicial de observações que, no caso dos dados da Tabela 5.1, possui tamanho 7 e é composto pelas seguintes observações: 8, 10, 15, 20, 22, 23 e 24. O conjunto inicial é escolhido através de *boxplots* bivariados (Zani, Riani e Corbellini, 1998), como já foi dito anteriormente.

Para esse exemplo de dados simulados, a ordem de entrada das observações em cada passo do processo iterativo desse método pode ser vista no gráfico da Figura 5.1. Só para exemplificar, as cinco primeiras observações a serem incluídas são as seguintes: 12, 16, 21, 18 e 1.

O gráfico da Figura 5.2 mostra como se comporta a distância de Mahalanobis para cada observação em cada passo do procedimento. As curvas em destaque mostram que as observações 26, 27, 28, 29 e 30 possuem a maior distância de Mahalanobis durante o processo e podem ser classificadas como atípicas ou *outliers* multivariados.

Analisando os gráficos da Figura 5.3, é possível perceber que, em ambos os casos, a distância de Mahalanobis apresenta um grande aumento a partir do momento em que o conjunto obtém tamanho acima de 25, ou seja, a partir do passo 19 do procedimento iterativo. É neste passo que a observação 27 é incluída no conjunto, como pode ser visto no gráfico da Figura 5.1, o que nos permite concluir que é uma observação atípica. Como a distância de Mahalanobis continua aumentando, os próximos pontos incluídos, 29, 28, 26 e 30, também podem ser classificados como *outliers* multivariados.

Tabela 5.2: Distância de Mahalanobis robusta ao quadrado para os dados da Tabela 5.1

Observação	RD_i^2
1	21,21
2	4,78
3	11,68
4	6,53
5	16,98
6	20,38
7	11,27
8	0,59
9	12,72
10	0,91
11	21,44
12	8,35
13	4,81
14	12,68
15	0,73
16	23,38
17	5,42
18	5,71
19	10,87
20	2,82
21	3,85
22	9,22
23	1,49
24	10,91
25	14,24
26	402641,50
27	27397,49
28	178319,20
29	74091,40
30	194686,30

Figura 5.1: Ordem de entrada das observações

Figura 5.2: Distância de Mahalanobis

Figura 5.3: Distância de Mahalanobis mínima e máxima

Resultado da Aplicação do Método Baseado no MCD

De acordo com o que foi dito no Capítulo 4 e observando-se o gráfico “Distância robusta ordenada ao quadrado” da Figura 5.4, que compara a distância de Mahalanobis robusta ao quadrado com a distribuição empírica de MD_i^2 , pode-se dizer que as observações atípicas deste conjunto de dados simulados são: 26, 27, 28, 29 e 30.

De acordo com os gráficos “*Outliers* baseados no quantil 95%”, que compara o quadrado da distância de Mahalanobis robusta com o quantil 95% da $\chi^2_{6;0,95}$, e “*Outliers* baseados no quantil ajustado”, as observações em destaque são as seguintes: 6, 11, 16, 24, 26, 27, 28, 29 e 30.

Para o exemplo em questão, como pode ser visto na Figura 5.5, as observações que são classificadas como *outliers* multivariados são 6, 11, 16, 24, 26, 27, 28, 29 e 30.

Resultado da Aplicação do Método Baseado na MED

Para os dados da Tabela 5.1, a aplicação deste método resulta no gráfico da Figura 5.6. Analisando o gráfico, é possível perceber que algumas observações estão distantes da massa de dados e, por isso, podem ser classificadas como *outliers* multivariados. São elas: 4, 10, 19, 26, 28, 29 e 30.

Figura 5.4: Gráficos obtidos pelo “aq.plot”

Figura 5.5: Gráfico obtido pelo “dd.plot”

Figura 5.6: MED para cada observação do conjunto de dados

5.1.2 Comparação dos Métodos para os Dados Simulados

Como uma mudança de 0 para 10 na variância é grande, simulamos mais dois conjuntos de dados: um no qual a variância da distribuição Normal é igual a 4 e outro com variância igual a 7 para as 5 últimas observações, ambas com média igual a zero e sendo as 25 primeiras observações as mesmas já apresentadas na seção anterior (Tabela 5.1). As 5 últimas observações para cada uma das variâncias encontram-se nas Tabelas 5.3 e 5.4.

Tabela 5.3: Dados gerados pela $N(0, 4)$

Observação	X_1	X_2	X_3	X_4	X_5	X_6
26	-0,084	-1,088	-0,231	8,544	-7,802	-2,691
27	-9,798	5,393	-0,395	0,852	0,908	-4,295
28	0,842	-0,835	0,153	0,897	-0,403	1,357
29	-4,522	5,348	-4,791	-1,505	-11,209	-3,827
30	4,855	1,480	8,387	-6,753	-1,206	0,697

Tabela 5.4: Dados gerados pela $N(0, 7)$

Observação	X_1	X_2	X_3	X_4	X_5	X_6
26	-6,701	-7,903	-0,146	-5,875	8,985	-4,740
27	4,191	-6,477	-4,717	2,142	-1,679	-5,452
28	4,425	7,486	-6,038	2,872	-1,349	-8,550
29	2,457	-4,216	-8,928	9,833	-0,590	-1,182
30	11,499	5,986	-0,581	6,381	-3,484	-3,466

Como já foi feita uma apresentação e uma comparação bem detalhada para o caso em que a variância muda de 1 para 10, os resultados das simulações utilizando as variâncias com valores iguais a 4 e 7 são apresentados de forma resumida na Tabela 5.5, que mostra quais observações foram classificadas como *outliers* multivariados.

5.2 Mudança na Média dos Dados Simulados

Nesta seção, todas as variáveis possuem distribuição Normal univariada com média igual a zero e variância igual a um ($N(0, 1)$) para as 25 primeiras observações e distribuição Normal univariada com média igual a vinte e variância igual a um ($N(20, 1)$) para as 5 observações restantes. Estes dados estão apresentados na Tabela 5.6. Os resultados das aplicações dos quatro métodos são apresentados a seguir.

5.2.1 Detalhes da Simulação Envolvendo a Distribuição $N(20, 1)$

Resultado da Aplicação do Método Baseado no MVE

Utilizando-se o quantil 0,95 da distribuição qui-quadrado, $\chi^2_{6;0,95}(12,59)$, as observações classificadas como atípicas são: 1, 4, 6, 7, 10, 11, 12, 18, 21, 23, 24, 26, 27, 28, 29 e 30. Se o quantil utilizado for o 0,98 da distribuição qui-quadrado, $\chi^2_{6;0,98}(15,03)$, a observação 6 deixa de ser atípica. A Tabela 5.7 mostra as distâncias de Mahalanobis robustas ao quadrado.

Tabela 5.5: Observações detectadas como *outliers* multivariados

Métodos	$\sigma^2 = 4$	$\sigma^2 = 7$	$\sigma^2 = 10$
MVE	1 16	1 25	1 25
	3 22	5 26	5 26
	5 24	6 27	6 27
	6 25	9 28	9 28
	8 26	11 29	11 29
	11 27	14 30	14 30
	12 29	16	16
FS	14 30		
	9	26	26
	26	27	27
	27	28	28
	29	29	29
MCD	30	30	30
	6 26	6 27	6 27
	11 27	11 28	11 28
	16 29	16 29	16 29
	24 30	24 30	24 30
MED		26	26
	26	2 27	4 28
	27	6 28	10 29
	29	17 29	19 30
	30	25 30	26
		26	

Tabela 5.6: Dados simulados com a média modificada

Observação	X_1	X_2	X_3	X_4	X_5	X_6
1	-0,461	1,780	0,517	-0,879	1,147	-2,259
2	1,332	0,133	-1,527	0,382	1,067	-0,319
3	-0,118	-0,744	0,831	-0,387	-0,363	-1,051
4	0,607	1,092	-0,351	-0,986	-0,258	-2,346
5	0,156	-0,725	0,024	-0,409	0,155	-0,802
6	-1,178	-0,227	0,402	-0,918	-1,390	0,767
7	-0,762	0,198	-0,523	-1,342	-1,631	1,592
8	0,341	-0,554	-1,248	1,037	-1,391	-1,960
9	0,904	-0,605	0,342	0,973	-0,134	0,404
10	0,164	-0,292	1,281	2,106	-1,611	-1,716
11	2,344	0,368	2,504	-0,946	-0,476	-1,318
12	0,130	1,198	2,300	-0,968	0,578	0,766
13	0,831	-0,332	0,013	1,426	-0,396	0,318
14	-0,147	-0,573	0,965	-0,889	-0,386	1,076
15	0,907	-0,614	-0,417	-0,487	0,310	-2,170
16	0,215	0,803	0,451	0,559	-0,204	0,881
17	-0,461	0,199	0,481	0,251	-1,365	1,112
18	1,453	0,276	-1,031	-0,379	-1,141	-1,166
19	-0,132	-0,701	-1,330	1,125	-0,377	0,023
20	0,499	1,102	-0,671	0,044	0,363	-0,232
21	-0,421	0,209	-0,681	-0,126	0,123	-1,068
22	0,457	0,575	-0,758	-0,714	0,682	0,306
23	-0,065	-0,617	0,730	0,586	1,946	2,034
24	-1,327	-0,423	0,034	-0,713	1,226	-1,944
25	-0,053	-1,691	0,269	-0,853	-0,132	-1,379
26	20,661	20,511	19,267	19,742	17,543	19,686
27	19,353	19,977	20,041	20,767	21,095	20,411
28	19,522	19,753	18,276	22,107	19,631	20,489
29	19,519	18,719	19,785	20,515	20,765	20,745
30	20,390	20,339	20,267	20,683	19,158	20,232

Tabela 5.7: Distância de Mahalanobis robusta ao quadrado para os dados da Tabela 5.6

Observação	RD_i^2
1	327,00
2	8,64
3	1,62
4	25,50
5	0,78
6	14,43
7	133,07
8	10,64
9	5,22
10	31,30
11	916,86
12	23,66
13	3,85
14	5,81
15	4,94
16	1,83
17	6,14
18	314,96
19	10,43
20	2,56
21	16,18
22	2,32
23	390,78
24	804,45
25	10,58
26	3095256,00
27	3981340,00
28	3598939,00
29	3716707,00
30	3625414,00

Para ambos os quantis utilizados, as observações que possuem distribuição diferente da maior parte dos dados foram classificadas como atípicas, mas observações que não eram para ser classificadas como *outliers* multivariados foram classificadas erroneamente.

Resultado da Aplicação do Método *The Forward Search*

No caso dos dados da Tabela 5.6, o conjunto inicial é composto pelas seguintes observações: 5, 9, 14, 15, 16, 20 e 21.

Conforme mostra o gráfico da Figura 5.7, a primeira observação a ser incluída nesse subconjunto é a 25, a segunda é a 3, a terceira é a 22 e assim por diante até a entrada da última observação que, nesse caso, é a 29.

Figura 5.7: Ordem de entrada das observações

Os gráficos das Figuras 5.8 e 5.9 mostram que as observações 26, 27, 28, 29 e 30 são as que possuem as maiores distâncias de Mahalanobis durante todo o processo. As distâncias de Mahalanobis destas observações diminuem no passo em que o subconjunto atinge tamanho 26, que é justamente o momento em que ocorre a entrada da observação 26. A ordem de entrada das outras quatro observações é 30, 28, 27 e 29.

Figura 5.8: Distância de Mahalanobis

Figura 5.9: Distância de Mahalanobis mínima e máxima

Resultado da Aplicação do Método Baseado no MCD

O gráfico “Distância robusta ordenada ao quadrado” da Figura 5.10 compara a distância de Mahalanobis robusta ao quadrado com a distribuição empírica de MD_i^2 e permite dizer que as observações 26, 27, 28, 29 e 30 são atípicas.

De acordo com os gráficos “*Outliers* baseados no quantil 95%”, que compara o quadrado da distância de Mahalanobis robusta com o quantil 95% da $\chi^2_{6;0,95}$, e “*Outliers* baseados no quantil ajustado”, as observações em destaque são as seguintes: 1, 11, 23, 24, 26, 27, 28, 29 e 30.

Figura 5.10: Gráficos obtidos pelo “aq.plot”

O gráfico da Figura 5.11 mostra que os *outliers* multivariados são as seguintes observações: 11, 24, 26, 27, 28, 29 e 30.

Resultado da Aplicação do Método Baseado na MED

Para os dados da Tabela 5.6, o gráfico resultante da aplicação deste método está apresentado na Figura 5.12. Analisando o gráfico, é possível perceber que as observações 8, 15, 18, 19 e 29 estão distantes da massa de dados e, por isso, podem ser classificados como *outliers* multivariados. As

Figura 5.11: Gráfico obtido pelo “dd.plot”

observações 26, 27, 28 e 30 não estão muito afastadas da massa de dados, mas têm um leve destaque com relação ao restante dos dados.

5.2.2 Comparação dos Métodos para os Dados Simulados

Como uma mudança de 0 para 20 na média também é grande, simulamos mais dois conjuntos de dados: um no qual a média da distribuição Normal é igual a 7 e outro com média igual a 14 para as 5 últimas observações, ambas com variância igual a um e sendo as 25 primeiras observações as mesmas já apresentadas na seção anterior (Tabela 5.6). As 5 últimas observações para cada uma das médias encontram-se nas Tabelas 5.8 e 5.9.

Como já foi feita uma apresentação e uma comparação bem detalhada para o caso em que a média muda de 0 para 20, os resultados das simulações utilizando as médias com valores iguais a 7 e 14 são apresentados de forma resumida, como foi feito no caso de mudança na variância, na Tabela 5.10, que mostra quais observações foram classificadas como *outliers* multivariados para cada um dos métodos quando a média foi alterada para 7, 14 e 20, respectivamente.

Figura 5.12: MED para cada observação do conjunto de dados

Tabela 5.8: Dados gerados pela $N(7, 1)$

Observação	X_1	X_2	X_3	X_4	X_5	X_6
26	6,158	5,700	6,794	5,516	7,455	6,194
27	7,504	7,645	7,084	6,367	8,864	8,298
28	7,594	7,919	6,243	6,992	7,608	7,558
29	7,333	6,036	7,050	5,524	6,745	6,517
30	7,894	9,804	5,997	7,584	6,061	7,156

Tabela 5.9: Dados gerados pela $N(14, 1)$

Observação	X_1	X_2	X_3	X_4	X_5	X_6
26	12,797	13,752	13,268	12,959	15,049	13,872
27	14,320	12,955	13,003	13,035	12,948	14,826
28	13,836	14,161	13,484	13,712	14,295	12,993
29	12,702	13,719	15,246	11,888	13,309	13,870
30	13,204	13,596	13,915	13,317	13,760	13,653

5.3 Resultados

Após aplicarmos todos os métodos de detecção de *outliers* multivariados em um mesmo conjunto de dados, é interessante criar uma medida que permita a comparação de cada um deles com relação ao desempenho na classificação das observações. Uma forma possível de se fazer isso é calcular qual foi o percentual de acerto na classificação das observações do conjunto de dados estudado. O modelo “erra” quando classifica uma observação como *outlier* quando ela não é e quando deixa de classificar uma observação como *outlier* sendo que ela é um *outlier*.

A Tabela 5.11 apresenta as porcentagens de acerto para os *outliers*, os não-*outliers* e total. Pode-se notar que, sob o critério adotado, o método que teve o melhor desempenho foi o FS e o pior foi o MVE. Os métodos MCD e MED apresentaram resultados similares e isso talvez possa ser explicado, pois um é baseado no determinante da matriz de covariância e o outro nos autovalores dessa matriz.

Com exceção do MED, o desempenho dos métodos melhora quando se aumenta a variância de 4 para 7, mas não se modifica quando a mudança é de 7 para 10.

Com relação à variação das médias, os métodos MVE, FS e MCD não apresentaram alterações em seus desempenhos, enquanto que o MED teve pior desempenho na variação da média de 7 para 14.

Tabela 5.10: Observações detectadas como *outliers* multivariados

Métodos	$\mu = 7$	$\mu = 14$	$\mu = 20$
MVE	1 21	1 21	1 21
	4 23	4 23	4 23
	6 24	6 24	6 24
	7 26	7 26	7 26
	10 27	10 27	10 27
	11 28	11 28	11 28
	12 29	12 29	12 29
	18 30	18 30	18 30
FS	26	26	26
	27	27	27
	28	28	28
	29	29	29
	30	30	30
MCD	1 27	1 27	1 27
	11 28	11 28	11 28
	23 29	23 29	23 29
	24 30	24 30	24 30
	26	26	26
MED	7 28	2 26	8 27
	23 29	14 27	15 28
	26 30	22 28	18 29
	27	23 29	19 30
		24 30	26

Tabela 5.11: Porcentagem de acertos

Distribuição	Grupo	Método			
		MVE	FS	MCD	MED
$N(0,4)$	<i>Outlier</i>	80	80	80	80
	<i>Não Outlier</i>	52	96	84	100
	Total	57	93	83	97
$N(0,7)$	<i>Outlier</i>	100	100	100	100
	<i>Não Outlier</i>	68	100	84	84
	Total	73	100	87	87
$N(0,10)$	<i>Outlier</i>	100	100	100	80
	<i>Não Outlier</i>	68	100	84	88
	Total	73	100	87	87
$N(7,1)$	<i>Outlier</i>	100	100	100	100
	<i>Não Outlier</i>	56	100	84	92
	Total	63	100	87	93
$N(14,1)$	<i>Outlier</i>	100	100	100	100
	<i>Não Outlier</i>	56	100	84	80
	Total	63	100	87	83
$N(20,1)$	<i>Outlier</i>	100	100	100	100
	<i>Não Outlier</i>	56	100	84	84
	Total	63	100	87	87

Capítulo 6

Conclusões

Neste trabalho, foram apresentados alguns métodos robustos de detecção de *outliers* multivariados. Há muitos outros na literatura, mas, por uma questão de limitação de tempo e espaço, foram apresentados apenas quatro.

Considerando os erros e acertos na classificação das observações dos conjuntos de dados analisados, pode-se concluir que o método que teve melhor desempenho foi o de Atinkson e Riani (2004), *The Forward Search*, pois ele classificou, com exceção de um único conjunto de dados, todas as observações atípicas corretamente e não errou ao classificar observações como *outliers* multivariados quando não eram.

Todos os outros métodos erraram ao classificar algumas observações como atípicas quando isso não era verdade, mas o pior desempenho, principalmente nos dados simulados, foi o do método MVE de Rousseeuw e Van Zomeren (1990).

Além do desempenho dos métodos, pode-se levar em conta os recursos que cada método apresenta para a detecção de pontos atípicos. Os métodos FS, MCD e MED possuem a vantagem de produzir gráficos como resultados, o que facilita a visualização do processo.

Todos os métodos apresentados neste trabalho possuem uma rotina desenvolvida no *software* R, o que não deixa de ser uma vantagem já que o R é gratuito.

Como indicação para continuidade deste trabalho, outros métodos robustos de detecção de *outliers* multivariados podem ser estudados e comparados com os quatro aqui apresentados.

As simulações apresentadas foram muito restritas e podem ser estendidas para um número bem maior de amostras.

Apêndice A

Programas

A.1 Método Baseado no MVE

```
library (MASS)
# LEITURA DOS DADOS #
data <- scan("C:\\...\\arquivo.dat", what=list
(var1=0,var2=0,var3=0,var4=0,var5=0))
attach(data)

# DIMENSÃO DOS DADOS #
n <- 20 # NÚMERO DE OBSERVAÇÕES #
p <- 5 # NÚMERO DE VARIÁVEIS #

# TRANSFORMA OS DADOS EM MATRIZ - ATUALIZAR O NÚMERO DE VARIÁVEIS #
matriz.data <- matrix(data,n,p)
matriz.data <- cbind(var1,var2,var3,var4,var5)

# CÁLCULO DAS ESTIMATIVAS MVE ROBUSTAS PARA OS PARÂMETROS DE LOCAÇÃO E ESCALA #
estrob <- cov.mve(matriz.data, cor = TRUE,
  quantile.used = floor((n + p + 1)/2),
  method = "mve", nsamp = "exact")
```

```
# O PARÂMETRO DE LOCAÇÃO É O CENTRO DA ELIPSE DETERMINADA PELO
ESTIMADOR MVE #
Tx <- estrob$center

# O PARÂMETRO DE ESCALA É PROPORCIONAL À ESTRUTURA DE COVARIÂNCIA
OBTIDA PELO ESTIMADOR MVE #
Cx <- estrob$cov

# CÁLCULO DA DISTÂNCIA DE MAHALANOBIS ROBUSTA PARA CADA UMA DAS OBSERVAÇÕES
UTILIZANDO O ESTIMADOR ROBUSTO MVE #
invCx <- solve(Cx)
RD <- matrix(0,n,1)
quadrRD <- matrix(0,n,1)

for (i in 1:n) {
RD[i] <- ((t(matriz.data[i,]-Tx))%*%invCx%*%(matriz.data[i,]-Tx))^1/2
quadrRD[i] <- (RD[i])^2
}

# QUANTIL DA DISTRIBUIÇÃO QUI-QUADRADO ESCOLHIDO PARA DETECTAR
OUTLIERS MULTIVARIADOS #
confianca <- 0.95
quantil <- qchisq(confianca, p)

# CRIAÇÃO DA MATRIZ QUE INDICA SE A OBSERVAÇÃO É OUTLIER #
outlier <- matrix(0,n,1)

for (i in 1:n) {
if (quadrRD[i] > quantil){
```

```
outlier[i] <- 1
}
if (quadrRD[i] <= quantil){
outlier[i] <- 0
}
}
outlier
```

A.2 Método *The Forward Search*

```
library (MASS)
library (Rfwdmv)

# LEITURA DOS DADOS #
data <- scan("C:\\...\\arquivo.dat", what=list
(var1=0,var2=0,var3=0,var4=0,var5=0))
attach(data)

# DIMENSÃO DOS DADOS#
n = 20 # NÚMERO DE OBSERVAÇÕES #
p = 5 # NÚMERO DE VARIÁVEIS #

# TRANSFORMA DADOS EM MATRIZ - ATUALIZAR O NÚMERO DE VARIÁVEIS #
X <- matrix(data,n,p)
X <- cbind(var1,var2,var3,var4,var5)

# MÉTODO FORWARD PARA O CONJUNTO DE DADOS X, UTILIZANDO BOXPLOTS
BIVARIADOS PARA DEFINIR O CONJUNTO INICIAL #
saida <- fwdmv.init(X, bsb = bb.subset, scaled = TRUE, monitor = "all")

# CONJUNTO INICIAL OBTIDO ATRAVÉS DOS BOXPLOTS ROBUSTOS BIVARIADOS #
conj.inicial <- matrix(0,saida$m,1)
aux <- 1
for(i in 1:n){
  if(saida$Unit[i,1]>0){
    conj.inicial[aux,1] <- i
    aux <- aux + 1}
}
```

```
# GRÁFICO QUE MOSTRA QUAL OBSERVAÇÃO FOI INCLUÍDA EM CADA PASSO DO PROCEDIMENTO #  
graf1 <- fwdmvEntryPlot(saida, entry.order = "final", subset.size = -1,  
  psfrag.labels = FALSE)
```

```
# GRÁFICO DA DISTÂNCIA DE MAHALANOBIS PARA CADA INDIVÍDUO  
EM CADA PASSO DO PROCEDIMENTO #  
graf2 <- fwdmvDistancePlot(saida, group = NULL, id = TRUE,  
  psfrag.labels = FALSE)
```

```
# GRÁFICO DA DISTÂNCIA DE MAHALANOBIS MÍNIMA E MÁXIMA #  
graf3 <- fwdmvMinmaxPlot(saida, psfrag.labels = FALSE)
```

A.3 Método Baseado no MCD

```
library (MASS)
library (mvoutlier)

#LEITURA DOS DADOS #
data <- scan("C:\\...\\arquivo.dat", what=list
(var1=0,var2=0,var3=0,var4=0,var5=0))
attach(data)

# DIMENSÃO DOS DADOS #
n = 20 # NÚMERO DE OBSERVAÇÕES #
p = 5 # NÚMERO DE VARIÁVEIS #

#TRANSFORMA DADOS EM MATRIZ - ATUALIZAR O NÚMERO DE VARIÁVEIS #
x <- matrix(data,n,p)
x <- cbind(var1,var2,var3,var4,var5)

graf1 <- aq.plot(x, delta=qchisq(0.95, df=ncol(x)), quan=1/2, alpha=0.05)
graf2 <- dd.plot(x, quan=1/2, alpha=0.025)
```

A.4 Método Baseado na MED

```
library (MASS)

#LEITURA DOS DADOS #
data <- scan("C:\\...\\arquivo.dat", what=list
(var1=0,var2=0,var3=0,var4=0,var5=0))
attach(data)

# DIMENSÃO DOS DADOS #
n = 20 # NÚMERO DE OBSERVAÇÕES #
p = 5 # NÚMERO DE VARIÁVEIS #

#TRANSFORMA DADOS EM MATRIZ - ATUALIZAR O NÚMERO DE VARIÁVEIS #
matriz.dados <- matrix(data,n,p)
matriz.dados <- cbind(var1,var2,var3,var4,var5)

# INICIALIZA MATRIZ DE AUTOVALORES E AUTOVETORES DOS DADOS #
autovalores <- matrix(0,n+1,1)
autovetores <- matrix(0,p,n+1)

# CÁLCULO DA MATRIZ DE MÉDIA DOS DADOS #
media.dados <- colMeans(matriz.dados)

# CÁLCULO DA MATRIZ DE VARIÂNCIA E COVARIÂNCIA DOS DADOS #
var.dados <- var(matriz.dados)

# CÁLCULO DOS AUTOVALORES E AUTOVETORES DA MATRIZ DE VARIÂNCIA
E COVARIÂNCIA DOS DADOS #
avv.dados <- eigen(var.dados)
autoval.dados <- eigen(var.dados, only.values=TRUE)$values
```

```

autovet.dados <- eigen(var.dados, only.values=FALSE)$vectors

# CÁLCULO DO MED PARA CADA OBSERVAÇÃO DO CONJUNTO DE DADOS #
autovalores[n+1] <- autoval.dados[1]
autovetores[,n+1] <- autovet.dados[,1]
for (i in 1:n) {
  matriz.dados1 <- matriz.dados[-i,]
  media.dados1 <- colMeans(matriz.dados1)
  var.dados1 <- var(matriz.dados1)
  avv.dados1 <- eigen(var.dados1)
  autoval.dados1 <- eigen(var.dados1, only.values=TRUE)$values
  autovet.dados1 <- eigen(var.dados1, only.values=FALSE)$vectors
  autovalores[i] <- autoval.dados1[1]
  autovetores[,i] <- autovet.dados1[,1]
}
produtos <- matrix(0,p,n+1)
produtos[,n+1] <- autovalores[n+1]*autovetores[,n+1]
for (i in 1:n) {
  produtos[,i] <- autovalores[i]*autovetores[,i]
}
diferenca <- matrix(0,p,n)
for (i in 1:n) {
  diferenca[,i] <- produtos[,i]-produtos[,n+1]
}
norma <- matrix(0,n,1)
for (i in 1:n) {
  norma[i] <- sqrt(diferenca[1,i]^2+diferenca[2,i]^2+diferenca[3,i]^2+
diferenca[4,i]^2+diferenca[5,i]^2)
}
y <- matrix(0,n,p)

```



```

for (i in 1:n){
for (j in 1:p) {
y[i,j] <- (t(matriz.dados[i,]-media.dados)%*%autovet.dados[,j])^2
}
}
indicador <- matrix(0,n,p)
for (i in 1:n) {
for (j in 1:p) {
if (y[i,j] < autoval.dados[j]){
indicador[i,j] <- 1
}
if (y[i,j] >= autoval.dados[j]){
indicador[i,j] <- 0
}
}
}
produtorio <- matrix(0,n,1)
d <- matrix(0,n,1)
for (i in 1:n) {
produtorio[i] <- indicador[i,1]*indicador[i,2]*indicador[i,3]*
indicador[i,4]*indicador[i,5]
d[i] <- norma[i]*(1-produtorio[i])
}
somad <- 0
for (i in 1:n) {
somad <- somad + d[i]
}
MED <- matrix(0,n,1)
for (i in 1:n) {
MED[i] <- d[i]/somad
}

```

```
}
```

```
MED
```

```
# GRÁFICO DO VALOR DO MED PARA CADA OBSERVAÇÃO DO CONJUNTO DE DADOS #
```

```
plot(MED)
```

Referências Bibliográficas

- [1] Atkinson, A. C. e Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag. New York.
- [2] Atkinson, A. C. e Riani, M. (2002). Tests in the Fan Plot for Robust, diagnostic Transformations in Regression. *Chemometrics and Intelligent Laboratory Systems*, 60, 87-100.
- [3] Atkinson, A. C. e Riani, M. (2004). The Forward Search and Data Visualisation. *Computational Statistics*, 19, 29-54.
- [4] Atkinson, A., Cerioli, A. e Riani, M. (2005). *Rfwdmv: Forward Search for Multivariate Data*. R package version 0.72-2. URL <http://www.riani.it>.
- [5] Becker, C. e Gather, U. (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, 94, 947-955.
- [6] Bussab, W. O. e Morettin, P. A. (2006). *Estatística Básica*. Atual Editora.
- [7] Butler, R. W., Davies, P. L. e Jhun, M. (1993). Asymptotics for the Minimum Covariance Determinant Estimator. *The Annals of Statistics*, 21, N°3, 1385-1400.
- [8] Davies, L. (1992). The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator. *The Annals of Statistics*, 20, N°4, 1828-1843.
- [9] Draper, N. R. e Smith, H. (1980). *Applied Regression Analysis*. 2.ed., John Wiley & Sons.
- [10] Filzmoser, P. (2005). Identification of Multivariate Outliers: A Performance Study. *Austrian Journal of Statistics*, 34, N°2, 127-138.

- [11] Filzmoser, P., Garrett, R. G. e Reimann, C. (2005). Multivariate Outlier Detection in Exploration Geochemistry. *Computers & Geosciences*, 31, 579-587.
- [12] Gao, S., Li, G. e Wang, D. (2005). A New Approach for Detecting Multivariate Outliers. *Communications in Statistics - Theory and Methods*, 34, 1857-1865.
- [13] Gasko, M. e Donoho, D. (1982). Influential Observation in Data Analysis. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 104-109.
- [14] Gervini, D. (2003). A Robust and Efficient Adaptative Reweighted Estimator of Multivariate Location and Scatter. *Journal of Multivariate Analysis*, 84, 116-144.
- [15] Gschwandtner, M. e Filzmoser, P. (2004). *mvoutlier: Multivariate outlier detection based on robust methods*. R package version 1.1. URL <http://www.statistik.tuwien.ac.at/public/filz/>.
- [16] Lopuhaä, H. P. e Rousseeuw, P. J. (1991). Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *The Annals of Statistics*, 19, 229-248.
- [17] R Development Core Team. (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>.
- [18] Riani, M. e Atkinson, A. C. (2001). A Unified Approach to Outliers, Influence and Transformations in Discriminant Analysis. *Journal of Computational and Graphical Statistics*, 10, 513-544.
- [19] Rousseeuw, P. J. (1985). Multivariate Estimation With High Breakdown Point. *Mathematical Statistics and Applications*, Vol. B, 283-297.
- [20] Rousseeuw, P. J. e Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons.
- [21] Rousseeuw, P. J. e Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, N°3, 212-223.
- [22] Rousseeuw, P. J. e Van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, 85, 633-651.
- [23] Zani, S., Riani, M e Corbellini, A. (1998). Robust Bivariate Boxplots and Multiple Outlier Detection. *Computational Statistics and Data Analysis*, 28, 257-270.