

**Alana Tavares Viana**

**Análise discriminante aplicada à classificação  
de livros infantojuvenis antigos e atuais**

Niterói - RJ, Brasil

13 de dezembro de 2018

**Alana Tavares Viana**

**Análise discriminante aplicada à  
classificação de livros infantojuvenis  
antigos e atuais**

**Trabalho de Conclusão de Curso**

Monografia apresentada para obtenção do grau de Bacharel em  
Estatística pela Universidade Federal Fluminense.

Orientador: Dr. Hugo Henrique Kegler dos Santos

Niterói - RJ, Brasil

13 de dezembro de 2018

**Alana Tavares Viana**

**Análise discriminante aplicada à classificação  
de livros infantojuvenis antigos e atuais**

Monografia de Projeto Final de Graduação sob o título  
*“Análise discriminante aplicada à classificação de livros in-  
fantojuvenis antigos e atuais”*, defendida por Alana Tavares  
Viana e aprovada em 13 de dezembro de 2018, na cidade de  
Niterói, no Estado do Rio de Janeiro, pela banca examinadora  
constituída pelos professores:

---

**Prof. Dr. Hugo Henrique Kegler dos Santos**  
Departamento de Estatística – UFF

---

**Prof. Dr. Douglas Rodrigues Pinto**  
Departamento de Estatística – UFF

---

**Prof. Me. Victor Eduardo Leite de Almeida Duca**  
Departamento de Estatística – UFF

Niterói, 13 de dezembro de 2018

V614a Viana, Alana Tavares  
Análise discriminante aplicada à classificação de livros  
infantojuvenis antigos e atuais / Alana Tavares Viana ; Hugo  
Henrique Kegler dos Santos, orientador. Niterói, 2018.  
66 p. : il.

Trabalho de Conclusão de Curso (Graduação em  
Estatística)-Universidade Federal Fluminense, Instituto de  
Matemática e Estatística, Niterói, 2018.

1. Literatura infantojuvenil. 2. Produção intelectual.I.  
Henrique Kegler dos Santos, Hugo, orientador. II. Universidade  
Federal Fluminense. Instituto de Matemática e Estatística.  
III. Título.

CDD -

# Resumo

Contextos históricos apresentam a inexistência do conceito de infância na antiguidade; consequentemente, não existia a criação de livros especificamente para o público jovem ou infantil. Livros infantojuvenis sofreram mudanças significativas entre os séculos XIX e XXI; a análise discriminante é utilizada para criar uma regra de classificação que discrimine esses dois tipos de livros através das variáveis quantidade de palavras por lauda e tamanho médio de sentenças por lauda, posteriormente um livro do século XX é classificado de acordo com a regra criada com o objetivo de verificar se a mudança dos livros ocorreu de forma gradual. Simulações foram realizadas com o objetivo de visualizar os cenários possíveis na análise discriminante.

Palavras-chaves: análise. discriminante. classificação. livros. infantojuvenis. antigos. atuais.

# Dedicatória

Dedico essa nova conquista à minha vó Acy (*in memorian*), que infelizmente não pôde estar presente neste momento tão importante da minha vida.

# Agradecimentos

Agradeço ao meu professor orientador Hugo pela motivação e grande ajuda na realização deste trabalho.

Agradeço à minha mãe Maritza, pessoa que mais admiro, pelo incentivo durante toda minha vida e pelas palavras de apoio, sem você eu não teria conseguido.

Agradeço ao meu namorado Marcelo pelo carinho e pela incrível paciência. Obrigada por me apoiar mesmo com tantas crises de estresse e ansiedade.

Agradeço aos meus tios Aurea e Edmon, e ao meu irmão Allan por sempre me apoiarem e motivarem.

Agradeço aos meus amigos Leonardo, Raphael, Vitória, Gabriela e Lívia que, com muito carinho e apoio, me ajudaram a concluir esta etapa da minha vida.

# Sumário

<b>1</b>	<b>Introdução</b>	p. 9
<b>2</b>	<b>Objetivos</b>	p. 10
<b>3</b>	<b>Métodos</b>	p. 11
3.1	Classificação em Duas Populações . . . . .	p. 12
3.1.1	Regras de classificação . . . . .	p. 14
3.1.2	Caso de Duas Populações Normais Multivariadas . . . . .	p. 18
3.1.2.1	Caso em que $\Sigma_1 = \Sigma_2 = \Sigma$ . . . . .	p. 19
3.1.2.2	Caso em que $\Sigma_1 \neq \Sigma_2$ . . . . .	p. 22
3.1.3	Avaliação das Regras de Classificação . . . . .	p. 24
3.1.3.1	Métodos de Estimação da <i>PTCI</i> . . . . .	p. 28
3.2	Classificação em mais de duas populações normais multivariadas . . . . .	p. 33
3.3	Transformações para aproximação da normalidade . . . . .	p. 35
<b>4</b>	<b>Simulações</b>	p. 37
4.1	Matrizes de covariância iguais ( $\Sigma_1 = \Sigma_2 = \Sigma$ ) . . . . .	p. 38
4.1.1	Sem correlação entre as variáveis . . . . .	p. 38
4.1.2	Com correlação entre as variáveis . . . . .	p. 40
4.2	Matrizes de covariância diferentes ( $\Sigma_1 \neq \Sigma_2$ ) . . . . .	p. 45
4.2.1	Sem correlação entre as variáveis . . . . .	p. 45
4.2.2	Com correlação entre as variáveis . . . . .	p. 50



<b>5</b>	<b>Resultados</b>	p. 55
5.1	Verificação de normalidade multivariada . . . . .	p. 55
5.1.1	Como treinar seu dragão . . . . .	p. 55
5.1.2	A ilha do tesouro . . . . .	p. 58
5.2	Teste de igualdade de matrizes de covariância . . . . .	p. 60
5.3	Regra de Classificação . . . . .	p. 61
5.4	Avaliação da regra de classificação . . . . .	p. 64
5.4.1	Ressubstituição (R) . . . . .	p. 64
5.4.2	Ressubstituição com Divisão Amostral (RDA) . . . . .	p. 65
5.4.3	Pseudo- <i>jackknife</i> (JACK) . . . . .	p. 66
5.5	Classificação do livro “O escaravelho do diabo” . . . . .	p. 67
<b>6</b>	<b>Considerações finais e trabalhos futuros</b>	p. 68
	<b>Referências</b>	p. 70

# 1 Introdução

Na sociedade antiga, não existia a “infância” [1]; a criança era vista como um adulto em miniatura, não possuía um espaço separado do “mundo adulto” e não lhe era dedicada atenção especial, elas viviam entre os adultos e eram tratadas como tal [2].

Os primeiros livros a serem destinados ao público infantil foram os contos de fadas, que eram histórias que foram colhidas junto ao povo. Essas histórias eram contadas nas rodas de conversa, em que conviviam todas as idades, pois, como foi afirmado anteriormente, na antiguidade não se tinha construída a noção de infância tal como se tem hoje [2].

Na Idade Média, inicia-se uma construção da concepção de infância e, com isso, a criança começa ser vista e tratada de maneira diferente do adulto [2]. No século XIX, a criança começa a deixar de ser considerada social e literalmente como um adulto em miniatura [3], fato que propiciou a ascensão do gênero literário dirigido ao jovem [1].

Atualmente, há uma maior preocupação na produção de livros direcionados a crianças e jovens; quanto menor a criança, mais se requer ilustrações, textos curtos e vocabulário simples [4]. Considerando que no século XIX essa preocupação ainda não era tão grande, é razoável supor que os livros infantojuvenis deste século possuíam características diferentes dos livros infantojuvenis atuais.

Para o presente estudo, supõe-se que exista uma diferença significativa na quantidade de palavras diferentes por lauda e no tamanho médio das frases por lauda entre os livros infantojuvenis antigos e atuais. A lauda é o conjunto de configurações que garantem o padrão de paginação e há diversas formas de obter-se tal padrão; neste trabalho, considera-se que uma lauda é formada por 1200 caracteres.

No capítulo 2, determina-se os objetivos deste trabalho. No capítulo posterior, é explicado o método utilizado na análise dos dados. Já no capítulo 4 são apresentadas simulações realizadas com o objetivo de facilitar a visualização de possíveis cenários. Os resultados obtidos são observados no capítulo 5 e no capítulo 6 é feito um resumo dos passos realizados e obtém-se as conclusões obtidas a partir dos resultados.

## 2 Objetivos

Este trabalho tem como objetivo estudar as diferenças entre livros infantojuvenis antigos e atuais; para isto, são selecionados um livro infantojuvenil antigo e um atual, e, utilizando o software R [5], são calculados a quantidade de palavras diferentes por lauda e o tamanho médio de sentenças por lauda.

Graças à contextualização histórica, é esperado que os livros infantojuvenis antigos apresentem uma quantidade maior de palavras diferentes por lauda e um tamanho médio de frases por lauda também maior; utiliza-se a Análise Discriminante para analisar tais suposições. Para tal, é novamente utilizado o software R [5] com o intuito de programar o método utilizado.

### 3 Métodos

A análise discriminante é uma técnica multivariada utilizada quando a variável dependente é qualitativa e as variáveis independentes são quantitativas. Está direcionada à separação de conjuntos distintos de objetos através das variáveis que melhor os diferenciam. Posteriormente, essas variáveis são usadas na criação de uma “função discriminante” capaz de deixar claras as diferenças entre os grupos, de forma que novos indivíduos possam ser alocados a um dos grupos predefinidos [6].

Proposta por Fisher no início do século XX para classificar novas espécies vegetais de acordo com suas características biométricas, rapidamente começou a ser utilizada também na Taxonomia e na Sistemática Vegetal, possuindo também extensa aplicação em ciências sociais e humanas [6]. No final da década de 1950, a análise discriminante começou a ser utilizada em Marketing, sendo facilmente aceita graças ao interesse dos acadêmicos em estudar a relação de pertencimento a grupos [7].

O propósito da análise é determinar o perfil de características de grupos que são mais dominantes em termo de discriminação, identificar quais são as principais funções discriminantes que os diferenciam, e desenvolver modelos preditivos para classificar novos indivíduos [8]. A primeira etapa é chamada de discriminação, e é quando são criadas as regras de separação dos grupos. Já a segunda etapa é chamada de classificação, consistindo em aplicar as regras definidas na primeira, e é, portanto, menos exploratória. Tais processos dependem essencialmente da estrutura do problema [9].

A classificação de novos indivíduos em grupos pré-definidos pode ser sujeito a incerteza devido a algumas condições: conhecimento incompleto de performance futura, informação “perfeita” requer que o objeto seja destruído, ou informações indisponíveis ou caras. Um exemplo de conhecimento incompleto seria a coordenação de um curso de Estatística desejar saber se é provável que um estudante conseguirá colocação profissional como estatístico com base nas notas e em outros registros da faculdade. Entretanto, a verdadeira classificação somente será conhecida ao desenvolver da carreira do profissional. Um caso

em que a informação “perfeita” requer que o objeto seja destruído é o tempo de vida da bateria de um calculadora ser determinado através do uso dela até ela parar de funcionar; porém, os produtos que serão vendidos não podem ser testados desta forma. Deseja-se, então, classificar os produtos como bons ou ruins com base em determinadas medidas preliminares. Por último, como exemplo de informações indisponíveis ou caras é o caso de muitas doenças somente poderem ser identificadas com certeza através de operações caras, tornando-se necessário diagnosticá-las a partir de sintomas externos facilmente observáveis, apesar de potencialmente falíveis, diminuindo o custo final.

Quanto à quantidade de grupos de classificação, tem-se que a variável dependente pode consistir de dois grupos, como por exemplo, baixo versus alto, mas também pode envolver mais de dois grupos, como classificações de baixo, médio e alto. A análise discriminante pode ser utilizada em ambos os casos. Quando o interesse for estudar somente dois grupos de variáveis dependentes, a técnica é chamada de Análise Discriminante Simples ou Entre Duas Populações, e se houver mais de dois grupos, é denominada Análise Discriminante Múltipla [10].

Neste trabalho, é apresentada a Análise Discriminante Simples, detalhando suas etapas, que envolvem a determinação de regras de classificação que minimizem os erros, considerando seus respectivos custos. Tais regras são definidas para o caso de duas populações normais multivariadas. Posteriormente, apresenta-se alguns métodos para avaliação das regras de classificação definidas. Comenta-se também o caso de mais de duas populações normais multivariadas.

## 3.1 Classificação em Duas Populações

A análise discriminante é baseada na variável discriminante, que é uma combinação linear de duas (ou mais) variáveis independentes que classificarão novos objetos em um dos grupos predefinidos. Na Tabela 1 são listados exemplos onde há interesse em separar duas classes de objetos ou atribuir novos objetos a uma dessas classes (ou ambas). As classes são chamadas de  $\pi_1$  e  $\pi_2$ . Normalmente, os objetos são classificados com base na mensuração de, por exemplo,  $p$  variáveis aleatórias associadas  $\mathbf{X}^T = [X_1, X_2, \dots, X_p]$ ,  $p \geq 1$ . Os valores observados de  $\mathbf{X}$ ,  $\mathbf{x}$ , diferem em cada uma das classes, ou seja, diferem para objetos nas populações  $\pi_1$  e  $\pi_2$  [9].

Tabela 1: Exemplos

Populações $\pi_1$ e $\pi_2$	Variável mensurada $\mathbf{X}$
Companhias de seguro solventes e inadimplentes	Ativo total (patrimônios), custo dos estoques e títulos, valor de mercado de estoques e títulos, despesas com perdas, excedente, prêmio seguros.
Pessoas com e sem perturbação gástrica	Mensuração de ansiedade, dependência, culpa e perfeccionismo.
Artigos de “O Federalista” escritos por James Madison e os escritos por Alexander Hamilton	Frequências de palavras diferentes e tamanho de sentenças.
Pessoas fáceis e difíceis de serem convencidas a comprar um novo produto	Educação, renda, tamanho da família, quantidade de mudanças anteriores de marca comprada.
Sucesso ou fracasso de estudantes universitários em se graduar	Nota na prova de ingresso na universidade, média das notas no ensino médio, número de atividades no ensino médio.
Risco de crédito bom e ruim	Renda, idade, quantidade de cartões de crédito, tamanho da família.
Alcoólico e não alcoólico	Atividade da enzima monoamina oxidase, atividade da enzima adenilato ciclase.

No quarto exemplo, os consumidores são separados em dois grupos (pessoas fáceis e difíceis de serem convencidas a comprar um novo produto) baseados nas variáveis que se presume que sejam relevantes (educação, renda, tamanho da família, quantidade de mudanças anteriores de marca comprada). Tem-se interesse em identificar uma observação da forma  $\mathbf{x}^\top = [x_1(\text{educação}), x_2(\text{renda}), x_3(\text{tamanho da família}), x_4(\text{quantidade de mudanças anteriores de marca comprada})]$  como populações  $\pi_1$  de pessoas fáceis de serem convencidas ou  $\pi_2$  de pessoas difíceis de serem convencidas [9].

### Representação geométrica da classificação em duas populações

A Figura 1 ilustra o gráfico de dispersão de um exemplo de análise discriminante envolvendo duas variáveis,  $X_1$  e  $X_2$ , no caso de duas populações,  $\pi_1$  e  $\pi_2$ , que possuem distribuição normal multivariada com médias  $[0 \ 0]$  e  $[0,5 \ -2]$ , respectivamente, e a mesma matriz de covariância  $\begin{bmatrix} 1 & 0,8 \\ 0,8 & 1 \end{bmatrix}$ . Foram geradas duas amostras normais multivariadas com tais parâmetros com o Software R [5], os círculos representam as observações obtidas na população  $\pi_1$  e os triângulos as obtidas na população  $\pi_2$ ; ao redor das observações é possível observar as elipses que cobrem 95% dos valores de cada amostra gerada e a linha pontilhada representa a regra de classificação para este exemplo.

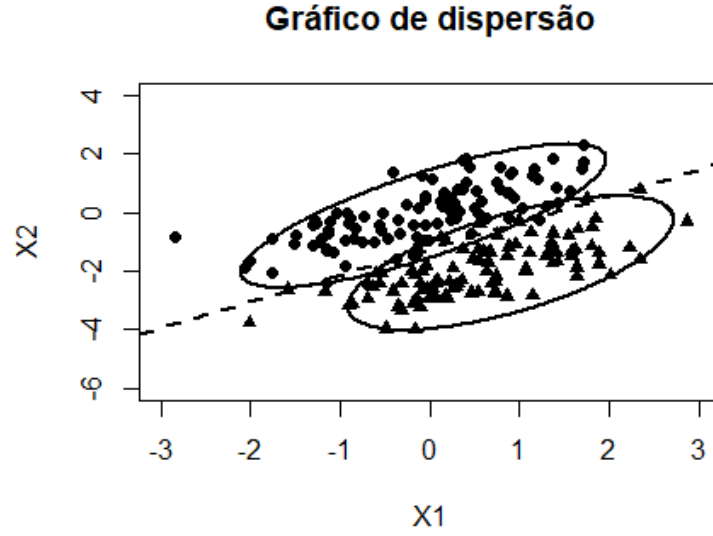


Figura 1: Ilustração gráfica da análise discriminante com duas populações

### 3.1.1 Regras de classificação

Na análise discriminante, para que as variáveis úteis na separação dos grupos sejam identificadas, deve-se considerar a diferença entre os grupos [10]. Em um bom processo de classificação, a probabilidade de que um elemento seja classificado incorretamente é pequena.

Sejam  $f_1(\mathbf{x})$  e  $f_2(\mathbf{x})$  as funções densidade de probabilidade associadas à variável aleatória  $\mathbf{X}$  para as populações  $\pi_1$  e  $\pi_2$ , respectivamente. Um objeto deve ser atribuído para  $\pi_1$  ou  $\pi_2$ . Sejam  $\Omega$  o espaço amostral, ou seja, o conjunto de todas as possíveis observações  $\mathbf{x}$ . Seja  $R_1$  o conjunto de valores de  $\mathbf{x}$  para os quais os objetos são classificados como pertencentes a  $\pi_1$  e  $R_2 = \Omega - R_1$  os demais valores de  $\mathbf{x}$  para os quais os objetos são classificados como pertencentes a  $\pi_2$ . Como todo objeto deve ser atribuído a uma única população, os conjuntos  $R_1$  e  $R_2$  são mutuamente exclusivos e exaustivos <sup>1</sup> [9].

A probabilidade de que um objeto seja classificado como pertencente a  $\pi_1$  quando ele pertence a  $\pi_2$  é dada por

$$P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}. \quad (3.1)$$

<sup>1</sup>Eventos são exaustivos quando a união desses eventos equivale ao espaço amostral.

Analogamente, a probabilidade de que um objeto seja classificado como pertencente a  $\pi_2$  quando na verdade ele pertence a  $\pi_1$  é dada por

$$P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \int_{R_2 = \Omega - R_1} f_1(\mathbf{x}) d\mathbf{x}. \quad (3.2)$$

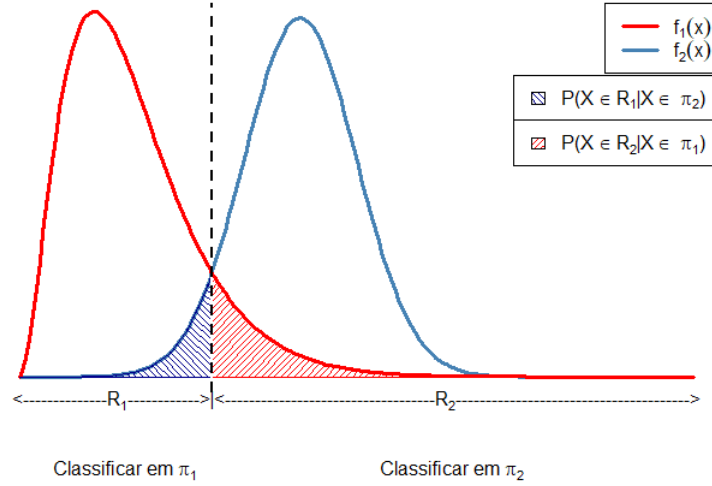


Figura 2: Probabilidades de classificações incorretas

Denotando as probabilidades *a priori*, isto é, probabilidades de um elemento realmente pertencer a uma população,  $p_1 = P(\mathbf{X} \in \pi_1)$  e  $p_2 = P(\mathbf{X} \in \pi_2)$  e observando que  $p_1 + p_2 = 1$ , então, as probabilidades de classificações corretas e incorretas dos objetos são dadas por:

- 1) Probabilidade da observação ser corretamente classificada em  $\pi_1$ :

$$P(\{\mathbf{X} \in R_1\} \cap \{\mathbf{X} \in \pi_1\}) = P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_1) \cdot p_1;$$

- 2) Probabilidade da observação ter vindo de  $\pi_2$  e ser incorretamente classificada em  $\pi_1$ :

$$P(\{\mathbf{X} \in R_1\} \cap \{\mathbf{X} \in \pi_2\}) = P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) \cdot p_2;$$

- 3) Probabilidade da observação ter vindo de  $\pi_2$  e ser corretamente classificada em  $\pi_2$ :

$$P(\{\mathbf{X} \in R_2\} \cap \{\mathbf{X} \in \pi_2\}) = P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_2) \cdot p_2;$$

- 4) Probabilidade da observação ter vindo de  $\pi_1$  e ser incorretamente classificada em  $\pi_2$ :

$$P(\{\mathbf{X} \in R_2\} \cap \{\mathbf{X} \in \pi_1\}) = P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) \cdot p_1.$$



Outro aspecto que deve ser considerado é o custo. Caso aconteça de a classificação de um objeto da população  $\pi_1$  como pertencente à população  $\pi_2$  representar um erro mais sério do que classificar um objeto da população  $\pi_2$  como pertencente à  $\pi_1$ , deve-se ter um determinado cuidado na classificação. Por exemplo, deixar de diagnosticar uma doença potencialmente fatal é consideravelmente mais “custoso” do que concluir que a doença está presente quando na verdade não está. É importante que o processo de classificação considere os custos associados aos erros de classificação sempre que possível [9].

Quando o objeto é classificado corretamente, o custo de erro de classificação é igual a 0, se uma observação de  $\pi_1$  é classificada como pertencente a  $\pi_2$ , seu custo é denotado por  $c(2|1)$ , e denota-se  $c(1|2)$  quando uma observação de  $\pi_2$  é classificada como pertencente a  $\pi_1$ . Os custos costumam ser apresentados em uma matriz, como na Tabela 2.

Tabela 2: Matriz dos custos de classificação

		Classificado como:	
		$\pi_1$	$\pi_2$
Pertence a:	$\pi_1$	0	$c(2 1)$
	$\pi_2$	$c(1 2)$	0

O custo médio de classificação incorreta, denotado por  $CMCI$ , é dado por

$$\begin{aligned}
 CMCI &= c(2|1) P(\{\mathbf{X} \in R_2\} \cap \{\mathbf{X} \in \pi_1\}) + c(1|2) P(\{\mathbf{X} \in R_1\} \cap \{\mathbf{X} \in \pi_2\}) \\
 &= c(2|1) P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) p_1 + c(1|2) P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) p_2.
 \end{aligned} \tag{3.3}$$

Os esquemas de classificação são frequentemente estimados em termos dos erros de classificação, ignorando o custo destes erros, o que pode causar problemas. Uma probabilidade aparentemente pequena pode ser significativa se o custo da classificação incorreta for muito alto.

No caso do conjunto de observações de cada população ser independente e a distribuição de probabilidades das características medidas dos elementos amostrais de cada população ser conhecida, pode-se utilizar o princípio da máxima verossimilhança para construir uma regra de classificação que minimize a chance de se classificar um elemento amostral incorretamente. A razão entre as duas distribuições de probabilidades, chamada de razão de verossimilhança entre as duas populações e denotada por  $\lambda(\mathbf{x})$ , é definida como

$$\lambda(\mathbf{x}) = \frac{\text{função densidade de } \mathbf{x} \text{ na população } \pi_1}{\text{função densidade de } \mathbf{x} \text{ na população } \pi_2} = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}.$$

Uma boa regra de classificação deve ter o menor  $CMCI$  possível. Neste sentido, substituindo 3.1 e 3.2 em 3.3, obtém-se

$$CMCI = c(2|1)p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + c(1|2)p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x}.$$

Como  $\Omega = R_1 \cup R_2$ , então a probabilidade total é

$$1 = \int_{\Omega} f_1(\mathbf{x})d\mathbf{x} = \int_{R_1} f_1(\mathbf{x})d\mathbf{x} + \int_{R_2} f_1(\mathbf{x})d\mathbf{x},$$

logo,

$$\int_{R_2} f_1(\mathbf{x})d\mathbf{x} = 1 - \int_{R_1} f_1(\mathbf{x})d\mathbf{x}.$$

Assim, pode-se escrever

$$\begin{aligned} CMCI &= c(2|1)p_1 \left[ 1 - \int_{R_1} f_1(\mathbf{x})d\mathbf{x} \right] + c(1|2)p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \\ &= c(2|1)p_1 + \int_{R_1} [c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})] d\mathbf{x}. \end{aligned}$$

Sabe-se que  $p_1$ ,  $p_2$ ,  $c(1|2)$  e  $c(2|1)$  são não negativos e, além disso,  $f_1(\mathbf{x})$  e  $f_2(\mathbf{x})$  são não negativos para todo  $\mathbf{x}$  e são as únicas quantidades no  $CMCI$  que dependem de  $\mathbf{x}$ . Portanto, o  $CMCI$  é minimizado se  $R_1$  incluir somente os valores de  $\mathbf{x}$  para os quais o integrando  $[c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})]$  é menor ou igual a 0. Assim, os valores de  $\mathbf{x}$  que satisfazem essas condições são encontrados desenvolvendo a inequação abaixo:

$$\begin{aligned} c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x}) &\leq 0 \\ \Leftrightarrow c(1|2)p_2 f_2(\mathbf{x}) &\leq c(2|1)p_1 f_1(\mathbf{x}) \\ \Leftrightarrow \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} &\leq \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}. \end{aligned}$$

Assim, as regiões  $R_1$  e  $R_2$  que minimizam o  $CMCI$  são definidas como

$$\begin{aligned} R_1 &= \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right\}, \\ R_2 &= \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right\}. \end{aligned} \tag{3.4}$$

### Casos especiais de regiões do *CMCI* mínimo

a) Probabilidades a priori iguais :

Quando as probabilidades a priori são desconhecidas, tais probabilidades são tomadas como iguais, isto é,  $\frac{p_2}{p_1} = 1$ , e a regra do *CMCI* mínimo envolve comparar as razões das densidades com a razão dos custos dos erros de classificação. Assim, as regiões de classificação são dadas por

$$R_1 = \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left( \frac{c(1|2)}{c(2|1)} \right) \right\}, \quad R_2 = \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left( \frac{c(1|2)}{c(2|1)} \right) \right\}.$$

b) Custos dos erros de classificação iguais:

Quando a razão dos custos dos erros de classificação é indeterminado, tal razão é tomada como igual a 1, isto é,  $\frac{c(1|2)}{c(2|1)} = 1$  e a razão de densidade é comparada com a razão das probabilidades a priori. Desta forma, tem-se que as regiões de classificação são:

$$R_1 = \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left( \frac{p_2}{p_1} \right) \right\}, \quad R_2 = \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left( \frac{p_2}{p_1} \right) \right\}. \quad (3.5)$$

c) Probabilidades a priori iguais e custos dos erros de classificação iguais:

Quando ambas as razões entre as probabilidades a priori e entre os custos dos erros de classificação são iguais a 1, isto é,  $\frac{p_2}{p_1} = \frac{c(1|2)}{c(2|1)} = 1$ , ou a razão entre as probabilidades a priori são o inverso da razão dos custos, ou seja,  $\frac{p_2}{p_1} = \frac{c(2|1)}{c(1|2)}$ , as regiões de classificação são determinadas através da comparação dos valores das funções de densidade. Logo,

$$R_1 = \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \right\} \quad R_2 = \left\{ \mathbf{x} \in \Omega : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1 \right\}.$$

### 3.1.2 Caso de Duas Populações Normais Multivariadas

A densidade da normal multivariada é uma generalização da densidade da normal univariada para  $p \geq 2$  dimensões. Para a distribuição normal univariada com média  $\mu$  e variância  $\sigma^2$ , a função de densidade de probabilidade é dada por

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\}, \quad -\infty < x < +\infty. \quad (3.6)$$

O termo

$$\left(\frac{x - \mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu) \quad (3.7)$$

no expoente da função de densidade normal univariada mede a distância quadrada de  $x$  em relação a  $\mu$  em unidade de desvio padrão. Esta distância pode ser generalizada para o caso multivariado, com um vetor  $\mathbf{x}$  de observações como

$$(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), \quad (3.8)$$

sendo o vetor  $\boldsymbol{\mu}$  o valor esperado do vetor de variáveis  $\mathbf{X}$ , e  $\boldsymbol{\Sigma}$  a matriz de covariância de  $\mathbf{X}$ . Assumindo-se que a matriz simétrica  $\boldsymbol{\Sigma}$  é positiva definida, então a expressão 3.8 representa a distância generalizada de  $\mathbf{x}$  para  $\boldsymbol{\mu}$ .

A densidade da normal multivariada é obtida através da substituição da distância univariada 3.7 pela distância generalizada 3.8 na função de densidade 3.6. No caso multivariado, as probabilidades são representadas pelos volumes sob as superfícies nas regiões definidas pelos intervalos dos valores  $x_i$ , assim, a constante univariada de normalização  $(2\pi)^{-1/2}(\sigma^2)^{-1/2}$  deve ser trocada por  $(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$  de modo a fazer com que o volume sob a superfície da função de densidade multivariada obtida seja igual a 1 para qualquer  $p$ . Logo, a densidade da normal  $p$ -dimensional para o vetor  $\mathbf{X}^\top = [X_1, X_2, \dots, X_p]$  é dada por:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right],$$

sendo  $-\infty < x_i < +\infty; i = 1, 2, \dots, p$ . Denota-se a densidade dessa normal  $p$ -dimensional por  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

Considerando que as populações  $\pi_1$  e  $\pi_2$  seguem, respectivamente, distribuições  $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  e  $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ , as regras de classificação dependem da relação entre as matrizes de covariância  $\boldsymbol{\Sigma}_1$  e  $\boldsymbol{\Sigma}_2$ .

### 3.1.2.1 Caso em que $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

Supondo que as densidades conjuntas de  $\mathbf{X}^\top = [X_1, X_2, \dots, X_p]$  para as populações  $\pi_1$  e  $\pi_2$  são dada por:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \right\} \quad \text{para } i = 1, 2. \quad (3.9)$$

Sejam os parâmetros  $\boldsymbol{\mu}_1$ ,  $\boldsymbol{\mu}_2$  e  $\boldsymbol{\Sigma}$  conhecidos, a função de verossimilhança é:

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \\ &= \frac{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right]}{\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right]} \\ &= \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right].\end{aligned}$$

Assim, o *CMCI* é mínimo se  $\mathbf{x}$  pertencer a  $R_1$  quando

$$\exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right) \left(\frac{p_2}{p_1}\right)$$

e pertencer a  $R_2$  caso contrário.

Pode-se simplificar o expoente da expressão à esquerda da desigualdade:

$$\begin{aligned}& -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \\ &= -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}[\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}[(\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1})(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).\end{aligned}$$

Logo,

$$\exp\left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right\} \geq \left(\frac{c(1|2)}{c(2|1)}\right) \left(\frac{p_2}{p_1}\right).$$

Como ambos os termos da desigualdade são positivos, pode-se tomar o logaritmo preservando a ordem da desigualdade. Assim, uma nova observação  $\mathbf{x}$  pertencerá a  $R_1$  quando

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right) \left(\frac{p_2}{p_1}\right)\right] \quad (3.10)$$

e pertencerá a  $R_2$  caso contrário.

Em situações reais, os parâmetros  $\boldsymbol{\mu}_1$ ,  $\boldsymbol{\mu}_2$  e  $\boldsymbol{\Sigma}$  não são conhecidos, sendo necessário modificar a regra 3.10 substituindo os parâmetros populacionais por suas respectivas estimativas de forma que tenha-se uma boa estimativa para a regra de classificação [11].

Supondo que de uma variável aleatória multivariada  $\mathbf{X}^\top = [X_1, X_2, \dots, X_p]$  é retirada uma amostra de  $n_1$  observações de  $\pi_1$  e  $n_2$  observações de  $\pi_2$ , sendo  $n_1 + n_2 - 2 \geq p$ . As matrizes de dados são:

$$\mathbf{X}_{1(n_1 \times p)} = \begin{bmatrix} \mathbf{x}_{11}^\top \\ \mathbf{x}_{12}^\top \\ \vdots \\ \mathbf{x}_{1n_1}^\top \end{bmatrix}, \quad \mathbf{X}_{2(n_2 \times p)} = \begin{bmatrix} \mathbf{x}_{21}^\top \\ \mathbf{x}_{22}^\top \\ \vdots \\ \mathbf{x}_{2n_2}^\top \end{bmatrix}.$$

A partir dessas matrizes de dados, os vetores de médias amostrais,  $\bar{\mathbf{x}}_{i(p \times 1)}$ , e as matrizes de covariâncias amostrais,  $\mathbf{S}_{i(p \times p)}$ ,  $i = 1, 2$ , são determinados por:

$$\begin{aligned} \bar{\mathbf{x}}_{1(p \times 1)} &= \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbf{x}_{1j}, & \mathbf{S}_{1(p \times p)} &= \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)^\top; \\ \bar{\mathbf{x}}_{2(p \times 1)} &= \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{x}_{2j}, & \mathbf{S}_{2(p \times p)} &= \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)^\top. \end{aligned}$$

No caso em que assume-se que as populações possuem a mesma matriz de covariância  $\boldsymbol{\Sigma}$ , as matrizes de covariância amostral  $\mathbf{S}_1$  e  $\mathbf{S}_2$  são combinadas para que se obtenha uma estimativa para  $\boldsymbol{\Sigma}$ , denotada por  $\mathbf{S}_p$ . A média ponderada

$$\begin{aligned} \mathbf{S}_p &= \left[ \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} \right] \mathbf{S}_1 + \left[ \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} \right] \mathbf{S}_2 \\ &= \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2} \end{aligned}$$

é o melhor estimador não-viesado para  $\boldsymbol{\Sigma}$  se as matrizes de dados  $\mathbf{X}_1$  e  $\mathbf{X}_2$  contiverem amostras aleatórias das populações  $\pi_1$  e  $\pi_2$ , respectivamente.

Substituindo  $\boldsymbol{\mu}_i$  por  $\bar{\mathbf{x}}_i$  e  $\boldsymbol{\Sigma}$  por  $\mathbf{S}_p$  na expressão 3.10, obtém-se a regra de classificação estimada que minimiza o custo médio de classificação incorreta. Portanto,  $\mathbf{x}$  deve ser alocado na população  $\pi_1$  se

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln \left[ \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right] \quad (3.11)$$

e na população  $\pi_2$ , caso contrário.

Se os custos e as probabilidades a priori forem iguais nas duas populações, ou seja, se

$$\left(\frac{c(1|2)}{c(2|1)}\right) \left(\frac{p_2}{p_1}\right) = 1,$$

a expressão 3.11 pode ser simplificada. Denominando o vetor  $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1}$  por  $\hat{\mathbf{a}}^\top$ , a região  $R_1$  fica definida por

$$R_1 : \left\{ \mathbf{x} \in \Omega : \hat{\mathbf{a}}^\top \mathbf{x} \geq \frac{1}{2} (\hat{\mathbf{a}}^\top \bar{\mathbf{x}}_1 + \hat{\mathbf{a}}^\top \bar{\mathbf{x}}_2) \right\}$$

e  $R_2$  como a região complementar a  $R_1$ .

Não há garantia de que o  $CMCI$  seja mínimo no caso em que os parâmetros populacionais estejam sendo substituídos pelos estimadores por serem desconhecidos. No entanto, em amostras grandes, espera-se um desempenho adequado do  $CMCI$ .

### 3.1.2.2 Caso em que $\Sigma_1 \neq \Sigma_2$

Verificou-se que sob a suposição de homogeneidade das matrizes de covariância, as regras de classificação são simples. Considerando-se a densidade da normal multivariada em 3.9 e substituindo-se  $\Sigma$  por  $\Sigma_i, i = 1, 2$ , as regras de classificação são mais complexas. Dessa forma, tem-se que  $f_1(\mathbf{x})$  e  $f_2(\mathbf{x})$  possuem distribuição normal multivariada com parâmetros  $\boldsymbol{\mu}_i$  e  $\Sigma_i, i = 1, 2$ , sendo  $\Sigma_1 \neq \Sigma_2$ . Conforme visto, a regra de classificação que minimiza o  $CMCI$  utiliza a razão de densidades  $f_1(\mathbf{x})/f_2(\mathbf{x})$  ou o logaritmo desta razão para ser definida.

Para o caso de matrizes de covariância heterogêneas, os termos  $|\Sigma_1|^{-1/2}$  e  $|\Sigma_2|^{-1/2}$  não se cancelam e, portanto, não podem ser feitas as simplificações que foram feitas no caso homocedástico. Assim, substituindo as densidades das normais multivariadas com matriz de covariância heterogêneas em 3.4, tomando o algoritmo natural e simplificando,  $\mathbf{x}$  deve ser classificado em  $\pi_1$  se

$$-\frac{1}{2} \mathbf{x}^\top (\Sigma_1^{-1} - \Sigma_2^{-1}) \mathbf{x} + (\boldsymbol{\mu}_1^\top \Sigma_1^{-1} - \boldsymbol{\mu}_2^\top \Sigma_2^{-1}) \mathbf{x} - \delta \geq \ln \left[ \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right],$$

sendo

$$\delta = \frac{1}{2} \ln \left( \frac{|\Sigma_1|}{|\Sigma_2|} \right) + \frac{1}{2} (\boldsymbol{\mu}_1^\top \Sigma_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \Sigma_2^{-1} \boldsymbol{\mu}_2),$$

e em  $\pi_2$  caso contrário.

Para a obtenção de uma regra de classificação estimada, substitui-se os parâmetros  $\mu_1$ ,  $\mu_2$ ,  $\Sigma_1$  e  $\Sigma_2$  pelas estimativas  $\bar{x}_1$ ,  $\bar{x}_2$ ,  $S_1$  e  $S_2$ , respectivamente. Assim,  $\mathbf{x}$  deve ser alocado em  $\pi_1$  se

$$-\frac{1}{2}\mathbf{x}^\top(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x} + (\bar{\mathbf{x}}_1^\top \mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2^\top \mathbf{S}_2^{-1})\mathbf{x} - \hat{\delta} \geq \ln \left[ \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right], \quad (3.12)$$

sendo

$$\hat{\delta} = \frac{1}{2} \ln \left( \frac{|\mathbf{S}_1|}{|\mathbf{S}_2|} \right) + \frac{1}{2} (\bar{\mathbf{x}}_1^\top \mathbf{S}_1^{-1} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top \mathbf{S}_2^{-1} \bar{\mathbf{x}}_2),$$

e em  $\pi_2$  caso contrário.

Pode-se observar na primeira imagem da Figura 3 que no caso de custos iguais e prioris iguais a regra baseada em duas distribuições normais com variâncias diferentes gera uma região  $R_2$  que consiste em dois conjuntos disjuntos.

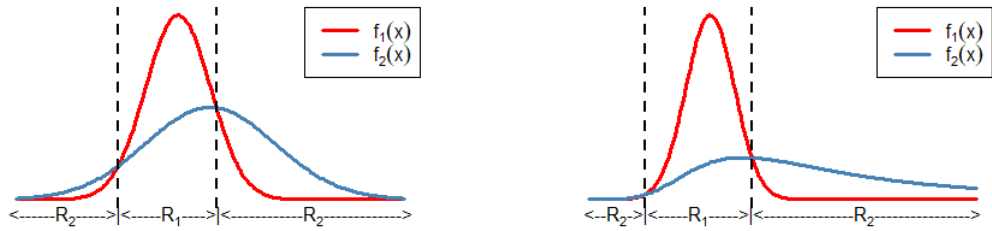


Figura 3: Regra de classificação para duas normais com variâncias diferentes e para duas distribuições com uma delas não possuindo distribuição normal

Técnicas discriminantes lineares ou quadráticas são muito sensíveis à violação de normalidade multivariada e podem gerar resultados insatisfatórios neste caso [9]. Em muitas aplicações, a cauda inferior da distribuição de  $\pi_2$  é menor do que aquela determinada por uma distribuição normal. Então, conforme ilustra a segunda imagem da Figura 3, o intervalo inferior de  $R_2$  não coincide com as distribuições populacionais e pode gerar grandes erros.

No caso em que os dados não possuam distribuição normal multivariada, deve-se transformar os dados para alcançar normalidade e testar a igualdade das matrizes de covariância para que sejam aplicadas as regras de classificação linear 3.11 ou quadrática 3.12 ao conjuntos de dados modificados. Os testes comuns de homogeneidade das matrizes de covariância são muitos sensíveis à suposição de normalidade; dessa forma, a transformação dos dados deve ser feita antes de testar-se as covariâncias [9].



### 3.1.3 Avaliação das Regras de Classificação

É possível mensurar o desempenho de uma regra de classificação através das taxas de erros ou probabilidades de classificação incorretas, que podem ser calculadas de forma relativamente simples quando a forma e os parâmetros das densidades populacionais são conhecidos. Como isto raramente acontece na prática, as probabilidades de classificação incorretas devem ser estimadas a partir da regra de classificação estimada obtida com a amostra de treinamento. É interessante medir o desempenho da regra construída.

Considerando duas populações  $p$ -variadas, a probabilidade total de classificação incorreta,  $PTCI$ , é dada por:

$$\begin{aligned}
 PTCI &= P(\text{Classificar incorretamente uma observação de } \pi_1 \\
 &\quad \text{ou classificar incorretamente uma observação de } \pi_2) \\
 &= P(\text{Observação de } \pi_1 \text{ ser classificada em } R_2) \\
 &\quad + P(\text{Observação de } \pi_2 \text{ ser classificada em } R_1) \\
 &= P(\{\mathbf{X} \in R_2\} \cap \{\mathbf{X} \in \pi_1\}) + P(\{\mathbf{X} \in R_1\} \cap \{\mathbf{X} \in \pi_2\}) \\
 &= p_1 P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) + p_2 P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) \\
 &= p_1 \int_{R_2} f_1(\mathbf{x}) d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}.
 \end{aligned} \tag{3.13}$$

Deve ser determinada uma regra de classificação que minimize a  $PTCI$ , o seu valor mínimo é chamado de taxa de erro ótima. Quando os custos de classificação incorretas são iguais, a minimização da  $PTCI$  resulta em uma regra de  $CMCI$  mínimo.

Considerando custos de classificação incorretas iguais, é possível determinar as taxas de erro ótimas para o caso particular de duas densidades normais multivariadas conhecidas e homocedásticas, com  $p_1 = p_2 = 1/2$ . A partir da expressão 3.10, a região  $R_1$  pode ser definida por:

$$R_1 : \left\{ \mathbf{x} \in \Omega | \mathbf{a}^\top \mathbf{x} \geq \frac{1}{2}(\mathbf{a}^\top \boldsymbol{\mu}_1 + \mathbf{a}^\top \boldsymbol{\mu}_2) \right\},$$

sendo  $\mathbf{a}^\top = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}$ . Neste caso,  $R_2$  é definida pela região complementar.

Definindo  $Y = \mathbf{a}^\top \mathbf{X}$  e  $y$  sua observação,  $\mathbf{x}$  é classificado em  $\pi_1$  se:

$$y \geq \frac{1}{2}(\mathbf{a}^\top \boldsymbol{\mu}_1 + \mathbf{a}^\top \boldsymbol{\mu}_2)$$

e em  $\pi_2$  caso contrário.

Nota-se que se  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , então qualquer combinação linear de  $\mathbf{X}$  dada por  $\mathbf{a}^\top \mathbf{X} = a_1 X_1 + a_2 X_2 + \dots + a_p X_p$  possui distribuição  $N(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a})$  [9]. Desta forma, se  $\mathbf{X}$  pertence a  $\pi_1$ , então  $Y \sim N(\mathbf{a}^\top \boldsymbol{\mu}_1, \Delta^2)$ , e se  $\mathbf{X}$  pertence a  $\pi_2$ , então  $Y \sim N(\mathbf{a}^\top \boldsymbol{\mu}_2, \Delta^2)$ , sendo

$$\begin{aligned}\mu_{1Y} &= \mathbf{a}^\top \boldsymbol{\mu}_1 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 \\ \mu_{2Y} &= \mathbf{a}^\top \boldsymbol{\mu}_2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \\ \sigma_Y^2 &= \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Delta^2.\end{aligned}$$

Assim, considerando  $p_1 = p_2 = 1/2$ , e denotando as distribuições de  $Y$  como  $g_1$  e  $g_2$  para caso  $\mathbf{X}$  pertencer as regiões  $\pi_1$  e  $\pi_2$ , respectivamente, e utilizando 3.13 para a classificação da realização  $y$ , obtém-se:

$$\begin{aligned}PTCI &= p_1 P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) + p_2 P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) \\ &= p_1 P\left[Y < \frac{1}{2} \mathbf{a}^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) | \mathbf{X} \in \pi_1\right] + p_2 P\left[Y \geq \frac{1}{2} \mathbf{a}^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) | \mathbf{X} \in \pi_2\right] \\ &= \frac{1}{2} \int_{-\infty}^{\frac{1}{2} \mathbf{a}^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)} g_1(y) dy + \frac{1}{2} \int_{\frac{1}{2} \mathbf{a}^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}^{\infty} g_2(y) dy.\end{aligned}$$

Pode-se verificar que  $\mu_{1Y} > \mu_{2Y}$ , pois

$$\begin{aligned}\mu_{1Y} &> \mu_{2Y} \\ \Leftrightarrow \mathbf{a}^\top \boldsymbol{\mu}_1 &> \mathbf{a}^\top \boldsymbol{\mu}_2 \\ \Leftrightarrow (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 &> (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \\ \Leftrightarrow (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 &> 0 \\ \Leftrightarrow (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) &> 0 \\ \Leftrightarrow \Delta^2 &> 0.\end{aligned}$$

Como as densidades  $g_1(y)$  e  $g_2(y)$  são normais univariadas, então a  $PTCI$  pode ser representada como na Figura 4.

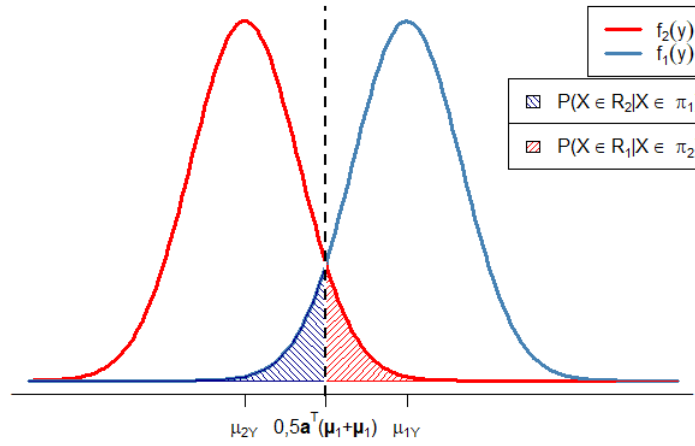


Figura 4: Probabilidades de classificação incorreta de uma observação  $\mathbf{x}$  em duas populações normais multivariadas, considerando  $p_1 = p_2 = 1/2$ , com base no ponto de corte  $\frac{1}{2}\mathbf{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$  e na transformação linear de  $\mathbf{X}$  para uma variável univariada  $Y$  com distribuição normal.

É possível calcular a *PTCI* padronizando o limite de integração  $\frac{1}{2}\mathbf{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ , com o objetivo de utilizar a normal-padrão para a obtenção das probabilidades. Para isso, são consideradas as densidades da variável aleatória normal  $Y$  de forma que os limites possam ser padronizados por  $[\frac{1}{2}\mathbf{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \mu_{iY}]/\sigma_Y$ . Considerando a população  $\pi_1$ , obtém-se:

$$\begin{aligned}
 z_1 &= \frac{\frac{1}{2} [\mathbf{a}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] - \mu_{1Y}}{\sigma_Y} \\
 &= \frac{1}{2} \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{\Delta} \\
 &= \frac{1}{2} \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 - 2\boldsymbol{\mu}_1)}{\Delta} \\
 &= \frac{1}{2} \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)}{\Delta} \\
 &= -\frac{1}{2} \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta} \\
 &= -\frac{1}{2} \frac{\Delta^2}{\Delta} \\
 &= -\frac{1}{2} \Delta.
 \end{aligned}$$

De forma análoga,

$$\begin{aligned}
 z_2 &= \frac{\frac{1}{2} [\mathbf{a}^\top (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)] - \mu_{2Y}}{\sigma_Y} \\
 &= \frac{\frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2}{\Delta} \\
 &= \frac{\frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 - 2\boldsymbol{\mu}_2)}{\Delta} \\
 &= \frac{\frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta} \\
 &= \frac{1}{2} \frac{\Delta^2}{\Delta} \\
 &= \frac{1}{2} \Delta.
 \end{aligned}$$

Assim, a *PTCI* pode ser simplificada em:

$$\begin{aligned}
 PTCI &= \frac{1}{2} \int_{-\infty}^{-\frac{1}{2}\Delta} \phi(z) dz + \frac{1}{2} \int_{\frac{1}{2}\Delta}^{\infty} \phi(z) dz \\
 &= \frac{1}{2} \Phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{2} \left[1 - \Phi\left(\frac{1}{2}\Delta\right)\right] \\
 &= \Phi\left(-\frac{1}{2}\Delta\right),
 \end{aligned} \tag{3.14}$$

sendo  $\phi(z) = (2\pi)^{-1/2} \exp\{-z^2/2\}$  e  $\Phi(z) = \int_{-\infty}^z (2\pi)^{-1/2} \exp\{-t^2/2\} dt$  as funções densidade e de distribuição da normal-padrão, respectivamente.

A probabilidade condicional de uma observação  $\mathbf{x}$  ser classificada incorretamente na população  $\pi_2$  dado que ela é originada de  $\pi_1$  é dada por

$$P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \Phi\left(-\frac{1}{2}\Delta\right) \tag{3.15}$$

e a probabilidade condicional de classificação incorreta na população  $\pi_1$  dado que a observação  $\mathbf{x}$  é originada de  $\pi_2$  é dada por

$$P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \Phi\left(-\frac{1}{2}\Delta\right). \tag{3.16}$$

Em ambos os casos, os custos e as probabilidades a priori estão sendo considerados idênticos. Se  $\frac{c(1|2)p_2}{c(2|1)p_1} \neq 1$ , tal fato deve ser contemplado pela regra de classificação. Definindo  $\Psi = \ln \left[ \frac{c(1|2)p_2}{c(2|1)p_1} \right]$ , observa-se que seu valor é positivo se o argumento da função logaritmo for maior que 1 e negativo, caso contrário. Assim, a partir de 3.10 obtém-se

que  $\mathbf{x}$  é classificado em  $\pi_1$  se

$$\mathbf{a}^\top \mathbf{x} \geq \frac{1}{2} \mathbf{a}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \Psi$$

e em  $\pi_2$ , caso contrário, sendo  $\mathbf{a}^\top = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}$ . Se o valor de  $\Psi$  for positivo, o valor limitante das regiões de classificação se deslocará para a direita, aumentando o valor de  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e diminuindo o valor de  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$ . Se o valor de  $\Psi$  for negativo, o deslocamento será para a esquerda, diminuindo o valor de  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e aumentando o valor de  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$ .

Padronizando  $\frac{1}{2} \mathbf{a}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \Psi$ , obtém-se os valores  $z_1 = -\frac{\Delta}{2} + \frac{\Psi}{\Delta}$  e  $z_2 = \frac{\Delta}{2} + \frac{\Psi}{\Delta}$ . Desta forma, a *PTCI* é dada por

$$\begin{aligned} PTCI &= p_1 \int_{-\infty}^{-\frac{\Delta}{2} + \frac{\Psi}{\Delta}} \phi(z) dz + p_2 \int_{\frac{\Delta}{2} + \frac{\Psi}{\Delta}}^{\infty} \phi(z) dz \\ &= p_1 \Phi \left( -\frac{\Delta}{2} + \frac{\Psi}{\Delta} \right) + p_2 \left[ 1 - \Phi \left( \frac{\Delta}{2} + \frac{\Psi}{\Delta} \right) \right] \end{aligned} \quad (3.17)$$

Da expressão 3.17 é possível verificar que

$$P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \Phi \left( -\frac{\Delta}{2} + \frac{\Psi}{\Delta} \right) \quad \text{e} \quad P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = 1 - \Phi \left( \frac{\Delta}{2} + \frac{\Psi}{\Delta} \right),$$

sendo que  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  se  $\Phi = 0$ ,  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) > P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  se  $\Phi > 0$  e  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) < P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  se  $\Phi < 0$ .

### 3.1.3.1 Métodos de Estimação da *PTCI*

Até então, a *PTCI* foi obtida quando as densidades populacionais são totalmente conhecidas. Usualmente, isto não ocorre, portanto os parâmetros dessas densidades e as regras devem ser estimados a partir da amostra de treinamento. A seguir, são apresentados alguns dos principais métodos utilizados para estimar a *PTCI*, as regiões de classificação e a função discriminante quando os parâmetros das densidades populacionais não são conhecidos, sendo utilizados estimadores obtidos a partir de amostras de treinamento. A taxa de erro aparente (*TEA*) é um estimador para a *PTCI* que pode ser utilizada para medir a qualidade da regra, tal medida independe da forma das densidades das populações estudadas.

**Ressubstituição:**

Na Ressubstituição, são utilizadas as amostras aleatórias das populações assumidas como normais multivariadas, ou seja,  $n_1$  observações p-variadas da população  $\pi_1$ ,  $\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1n_1}$  e  $n_2$  da população  $\pi_2$ ,  $\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2n_2}$ , com  $n_1 + n_2 - 2 \geq p$ , para estimar os parâmetros e a função discriminante. A partir da regra de classificação estimada, cada observação amostral é classificada em uma das populações e as classificações podem estar corretas ou incorretas. A Tabela 3 apresentada a seguir resume os resultados em uma matriz denominada *matriz de confusão* ou tabela de contingência.

Tabela 3: Matriz de confusão			
População real	População classificada		Total
	$\pi_1$	$\pi_2$	
$\pi_1$	$n_{11}$	$n_{12}$	$n_1$
$\pi_2$	$n_{21}$	$n_{22}$	$n_2$
			$n = n_1 + n_2$

Nessa matriz,  $n_i$  representa o tamanho da amostra obtida na  $i$ -ésima população,  $n_{ij}$   $i \neq j$  é o número de observações da  $i$ -ésima população classificadas incorretamente na  $j$ -ésima população e  $n_{ii}$  é o número de observações da  $i$ -ésima população classificadas corretamente, sendo  $i, j = 1, 2$ .

A proporção de observações classificadas incorretamente representa a probabilidade total estimada de classificação incorreta. O estimador da *PTCI*, denominado de *TEA*, é dado por:

$$TEA = \frac{n_{12} + n_{21}}{n_1 + n_2} = \frac{n_{12} + n_{21}}{n}.$$

As probabilidades  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  também podem ser estimadas por

$$\hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \frac{n_{12}}{n_1} \quad \text{e} \quad \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \frac{n_{21}}{n_2}.$$

As estimativas das taxas de erros de classificação fornecidas são muito otimistas, ou seja, há tendência de subestimar-se a *PTCI*,  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$ , já que são utilizadas as mesmas amostras tanto para estimar-se a função discriminante quanto para estimar-se as taxas de erro. O método somente possui bom desempenho com amostras grandes, pois o viés da estimação das taxas de erro se aproxima de zero.

### Ressubstituição com Divisão Amostral:

É uma variação do método anterior, porém obtendo-se estimativas das taxas de erros mais precisas. Divide-se a amostra em duas partes, a amostra de treinamento e amostra de validação. A primeira é utilizada na estimação da função discriminante e especificação da regra estimada de classificação, enquanto na segunda as observações são classificadas de acordo com a regra estimada na amostra de treinamento e a partir do resultado final são estimadas as taxas de erro.

Apesar do problema do viés ter sido superado, os resultados somente são considerados adequados em amostras grandes e há perda de informações valiosas na especificação da regra de classificação, já que apenas parte dos dados são utilizados.

### Pseudo-jackknife:

Este é o primeiro método apresentado por Lachenbruch e Mickey [12]. Omite-se uma observação por vez das  $n_1 + n_2$  observações originais na amostra de treinamento e utiliza-se as  $n_1 + n_2 - 1$  observações remanescentes para estimar as regras de classificação. Cada observação omitida é classificada em uma das populações e é possível determinar se tal observação foi classificada correta ou incorretamente, já que sabe-se de qual população ela foi originalmente amostrada. É possível notar que a  $j$ -ésima observação da  $i$ -ésima população omitida,  $\mathbf{x}_{i,j}$ , não entra na estimação da regra de classificação, mas é utilizada para avaliar o desempenho da regra recém-construída. Assim, pode-se determinar a quantidade de observações originadas de  $\pi_1$  que foram classificadas de forma incorreta em  $\pi_2$ , denotada por  $n_{12}^{(j)}$ , e a quantidade de observações de  $\pi_2$  que foram classificadas incorretamente em  $\pi_1$ , denotada por  $n_{21}^{(j)}$  e as probabilidades  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  podem ser estimadas por

$$\hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \frac{n_{12}^{(j)}}{n_1} \quad \text{e} \quad \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \frac{n_{21}^{(j)}}{n_2},$$

em que o superescrito  $(j)$  denota o procedimento *Jackknife*.

A TEA estimada pelo método é

$$TEA^{(j)} = \frac{n_{12}^{(j)} + n_{21}^{(j)}}{n_1 + n_2}$$

Esse método não é muito sensível à violação de normalidade, já que mesmo com tamanhos moderados de amostras fornece estimativas aproximadamente não-viesadas das taxas de erros reais.

**Probabilidade de Classificação Incorreta Estimada:**

No caso em que há suposição de normalidade multivariada, se os parâmetros populacionais são conhecidos e os custos de classificação incorreta e probabilidades a priori são iguais, a *PTCI* é determinada a partir da expressão 3.14, e as probabilidades de classificação incorreta  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  são determinadas a partir das expressões 3.15 e 3.16, respectivamente. É possível estimar as probabilidades substituindo  $\Delta$  pelo seu valor estimado dado por

$$\hat{\Delta} = \sqrt{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}. \quad (3.18)$$

Sob normalidade, tem-se que

$$\frac{n_1 n_2}{n_1 + n_2} \hat{\Delta}^2 \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - 1 - p} F(p, n_1 + n_2 - 1 - p, \delta^2),$$

sendo  $\delta^2 = \frac{n_1 n_2}{n_1 + n_2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$  o parâmetro de não centralidade da distribuição  $F$ . Então,

$$\begin{aligned} \frac{n_1 n_2}{n_1 + n_2} E(\hat{\Delta}^2) &= \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - 1 - p} E[F(p, n_1 + n_2 - 1 - p, \delta^2)] \\ &= \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - 1 - p} \frac{n_1 + n_2 - 1 - p}{p} \left[ \frac{p + \frac{n_1 n_2}{n_1 + n_2} \Delta^2}{n_1 + n_2 - 3 - p} \right] \\ \Leftrightarrow E(\hat{\Delta}^2) &= \frac{n_1 + n_2 - 2}{n_1 + n_2 - 3 - p} \frac{n_1 + n_2}{n_1 n_2} \left( p + \frac{n_1 n_2}{n_1 + n_2} \Delta^2 \right) \\ &= \frac{n_1 + n_2 - 2}{n_1 + n_2 - 3 - p} \left[ \Delta^2 + \frac{p(n_1 + n_2)}{n_1 n_2} \right] \end{aligned}$$

Assim, verifica-se que a estimativa de  $\Delta$  apresentada na equação 3.18 é viesada e, portanto, não deve ser utilizada. Um estimador não-viesado de  $\Delta$  é dado por

$$\tilde{\Delta}^2 = \frac{n_1 + n_2 - 3 - p}{n_1 + n_2 - 2} \hat{\Delta}^2 - \frac{p(n_1 + n_2)}{n_1 n_2}.$$

As probabilidades  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  são estimadas por

$$\hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \Phi \left( -\frac{\tilde{\Delta}}{2} \right)$$

e a *TEA* por

$$TEA = \Phi \left( -\frac{\tilde{\Delta}}{2} \right). \quad (3.19)$$



No caso em que  $\Psi$  seja não nulo, ou seja, custos e probabilidades a priori não necessariamente iguais, a  $TEA$  é dada por

$$TEA^\Psi = p_1 \Phi \left( -\frac{\tilde{\Delta}}{2} + \frac{\Psi}{\tilde{\Delta}} \right) + p_2 \left[ 1 - \Phi \left( \frac{\tilde{\Delta}}{2} + \frac{\Psi}{\tilde{\Delta}} \right) \right]. \quad (3.20)$$

Da expressão 3.20 verifica-se que

$$\hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \Phi \left( -\frac{\tilde{\Delta}}{2} + \frac{\Psi}{\tilde{\Delta}} \right) \quad \text{e} \quad \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \left[ 1 - \Phi \left( \frac{\tilde{\Delta}}{2} + \frac{\Psi}{\tilde{\Delta}} \right) \right].$$

### Método Dois de Lachenbruch e Mickey:

Este método é baseado em um procedimento que combina a técnica *jackknife* e o método das probabilidades de classificações incorretas estimadas. Das  $n_1 + n_2$  observações, deve ser omitida a realização  $\mathbf{x}_{ij}$  da  $i$ -ésima população referente à  $j$ -ésima unidade amostral, sendo  $i = 1, 2$  e  $j = 1, 2, \dots, n_i$ . As médias amostrais das populações 1 e 2 são representadas por  $\bar{\mathbf{X}}_1^{-(ij)}$  e  $\bar{\mathbf{X}}_2^{-(ij)}$ , respectivamente, e a matriz de covariâncias amostrais comum é representada por  $\mathbf{S}_p^*$ , tais medidas são calculadas excluindo a observação  $\mathbf{x}_{ij}$ , para a qual o valor  $y_{ij}$  é dado por

$$y_{ij} = \left( \bar{\mathbf{X}}_1^{-(ij)} - \bar{\mathbf{X}}_2^{-(ij)} \right)^\top \mathbf{S}_p^{*-1} \mathbf{x}_{ij} - \frac{1}{2} \left( \bar{\mathbf{X}}_1^{-(ij)} - \bar{\mathbf{X}}_2^{-(ij)} \right)^\top \mathbf{S}_p^{*-1} \left( \bar{\mathbf{X}}_1^{-(ij)} + \bar{\mathbf{X}}_2^{-(ij)} \right).$$

Determinando  $y_{ij}$  para todos os valores de  $i$  e  $j$ , omitindo somente a observação  $\mathbf{x}_{ij}$  em cada etapa, obtém-se uma amostra  $y_{11}, y_{12}, \dots, y_{1n_1}$  da população  $\pi_1$  e outra  $y_{21}, y_{22}, \dots, y_{2n_2}$  da população  $\pi_2$ . As médias e variâncias de cada amostra são dadas por

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad \text{e} \quad S_i^2 = \frac{1}{n_i - 1} \left[ \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left( \sum_{j=1}^{n_i} y_{ij} \right)^2}{n_i} \right],$$

para  $i = 1, 2$ .

As probabilidades  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  são estimadas por

$$\hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \Phi \left( -\frac{\bar{y}_1}{S_1} \right) \quad \text{e} \quad \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \Phi \left( \frac{\bar{y}_2}{S_2} \right)$$

e a taxa de erro aparente por

$$TEA = \frac{1}{2}\Phi\left(-\frac{\bar{y}_{1.}}{S_1}\right) + \frac{1}{2}\Phi\left(\frac{\bar{y}_{2.}}{S_2}\right).$$

No caso de custos e probabilidades a priori potencialmente diferentes, utiliza-se a expressão

$$y_{ij}^* = \left(\bar{\mathbf{X}}_1^{-(ij)} - \bar{\mathbf{X}}_2^{-(ij)}\right)^\top \mathbf{S}_p^{*-1} \mathbf{x}_{ij} - \frac{1}{2} \left(\bar{\mathbf{X}}_1^{-(ij)} - \bar{\mathbf{X}}_2^{-(ij)}\right)^\top \mathbf{S}_p^{*-1} \left(\bar{\mathbf{X}}_1^{-(ij)} + \bar{\mathbf{X}}_2^{-(ij)}\right) - \Psi.$$

Após a obtenção das observações  $y_{ij}^*$ , todo o processo é repetido.

## 3.2 Classificação em mais de duas populações normais multivariadas

Para a situação em que uma variável p-variada  $\mathbf{x}$  deve ser alocada em uma entre  $k$  populações normais, é utilizada a regra de minimização da *PTCI*, ou seja, supõe-se que os custos de classificação incorreta  $c(i|j)$  são todos iguais, qualquer que seja  $j \neq i = 1, 2, \dots, k$ . Seja a função densidade normal multivariada dada por

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\}$$

para  $i = 1, 2, \dots, k$ , sendo  $\boldsymbol{\mu}_i$  o vetor de médias e  $\boldsymbol{\Sigma}_i$  a matriz de covariâncias da  $i$ -ésima população. Se esses parâmetros forem conhecidos, então, utilizando a regra de classificação de mínima *PTCI*,  $\mathbf{x}$  é classificado na população  $\pi_i$  se

$$\begin{aligned} \ln[p_i f_i(\mathbf{x})] &= \ln(p_i) - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln(|\boldsymbol{\Sigma}_i|) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \\ &= \max_j \ln[p_j f_j(\mathbf{x})]. \end{aligned} \quad (3.21)$$

Ou seja, a observação  $\mathbf{x}$  é alocada à população que maximiza  $\ln[p_j f_j(\mathbf{x})]$  em relação a todos os valores de  $j$ ,  $j = 1, 2, \dots, k$ . Pode-se observar que a variância generalizada  $|\boldsymbol{\Sigma}_i|$ , a probabilidade *a priori*  $p_i$  e a distância quadrática entre  $\mathbf{x}$  e a média populacional  $\boldsymbol{\mu}_i$  dependem do subscrito  $i$ , referente à  $i$ -ésima população e portanto são diferentes para cada uma delas, contribuindo para o critério, enquanto o termo  $\frac{p}{2} \ln(2\pi)$  em 3.21 é constante para todas as  $k$  populações e pode ser ignorado. Assim, pode ser determinado o escore quadrático de discriminação e para a  $i$ -ésima população é dado por

$$d_i^Q(\mathbf{x}) = -\frac{1}{2} \ln(|\boldsymbol{\Sigma}_i|) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln(p_i). \quad (3.22)$$

Com isso, a regra de classificação pode ser simplificada da seguinte forma: a observação  $\mathbf{x}$  é classificada em  $\pi_i$  se

$$d_i^Q(\mathbf{x}) = \max_j \left[ d_j^Q(\mathbf{x}) \right]$$

para  $j = 1, 2, \dots, k$ .

Em situações reais os parâmetros populacionais são desconhecidos, então são usados os estimadores correspondentes obtidos na amostra de treinamento. Considerando uma amostra de  $n_i$  observações p-variadas  $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{ij}, \dots, \mathbf{x}_{in_i}$  da população  $\pi_i$ , com  $i = 1, 2, \dots, k$  e  $j = 1, 2, \dots, n_i$ , então

$$\bar{\mathbf{x}}_i = \frac{\sum_{j=1}^{n_i} \mathbf{x}_{ij}}{n_i} \quad \text{e} \quad \mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^\top,$$

são os estimadores da média e da matriz de covariância da população  $i$ -ésima população.

Assim, o estimador da função quadrática  $d_i^Q(\mathbf{x})$ , representado por  $Q_i(\mathbf{x})$ , é dado por

$$Q_i(\mathbf{x}) = -\frac{1}{2} \ln(|\mathbf{S}_i|) - \frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}}_i)^\top \mathbf{S}_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i) + \ln(p_i),$$

para  $i = 1, 2, \dots, k$ .

Pela regra estimada de mínima *PTCI*,  $\mathbf{x}$  é classificado em  $\pi_i$  se

$$Q_i(\mathbf{x}) = \max_j [Q_j(\mathbf{x})],$$

para  $j = 1, 2, \dots, k$ .

No caso em que as matrizes de covariância são iguais, ou seja,  $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$ , o escore 3.22 pode ser simplificado por

$$\begin{aligned} d_i^Q(\mathbf{x}) &= -\frac{1}{2} \ln(|\Sigma|) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln(p_i) \\ &= -\frac{1}{2} \ln(|\Sigma|) - \frac{1}{2} \mathbf{x}^\top \Sigma^{-1} \mathbf{x} + \boldsymbol{\mu}_i^\top \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i^\top \Sigma^{-1} \boldsymbol{\mu}_i + \ln(p_i), \end{aligned}$$

para  $i = 1, 2, \dots, k$ .

Como os termos  $-\frac{1}{2} \ln(|\Sigma|)$  e  $-\frac{1}{2} \mathbf{x}^\top \Sigma^{-1} \mathbf{x}$  são constantes em todas as  $k$  populações, podem ser ignorados, obtendo-se o escore discriminante linear  $d_i(\mathbf{x})$ , dado por

$$d_i(\mathbf{x}) = \boldsymbol{\mu}_i^\top \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i^\top \Sigma^{-1} \boldsymbol{\mu}_i + \ln(p_i), \quad (3.23)$$

para  $i = 1, 2, \dots, k$ .

Portanto,  $\mathbf{x}$  é classificado em  $\pi_i$  se

$$d_i(\mathbf{x}) = \max_j [d_j(\mathbf{x})]$$

para  $j = 1, 2, \dots, k$ .

Obtém-se uma estimativa para o escore 3.23 substituindo os parâmetros populacionais por suas respectivas estimativas. Como as matrizes de covariâncias são iguais entre si, então o melhor estimador não-viesado para  $\Sigma$  é dado por

$$\begin{aligned} \mathbf{S}_p &= \frac{\sum_{i=1}^k (n_i - 1) \mathbf{S}_i}{\sum_{i=1}^k (n_i - 1)} \\ &= \frac{(n_1 - 1) \mathbf{S}_1 + (n_2 - 1) \mathbf{S}_2 + \dots + (n_k - 1) \mathbf{S}_k}{n_1 + n_2 + \dots + n_k - k} \end{aligned}$$

Assim, o estimador do escore discriminante linear é dado por

$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i^\top \mathbf{S}_p^{-1} \mathbf{x} - \frac{1}{2} \bar{\mathbf{x}}_i^\top \mathbf{S}_p^{-1} \bar{\mathbf{x}}_i + \ln(p_i),$$

para  $i = 1, 2, \dots, k$ .

De acordo com a regra estimada baseada no escore discriminante linear,  $\mathbf{x}$  é classificado em  $\pi_i$  se

$$\hat{d}_i(\mathbf{x}) = \max_j [\hat{d}_j(\mathbf{x})] \quad (3.24)$$

para  $j = 1, 2, \dots, k$ .

### 3.3 Transformações para aproximação da normalidade

Se a normalidade não é uma suposição viável, existem duas alternativas para o próximo passo. Uma alternativa é ignorar os resultados dos testes de normalidade e prosseguir como se os dados fossem normalmente distribuídos. Esta prática não é recomendada, já que, em muitos casos, pode levar a conclusões erradas. Uma segunda alternativa é aplicar transformações para que populações que não possuam distribuição normal se aproximem da normalidade.

Transformações são reexpressões dos dados em unidades diferentes. Por exemplo, quando um histograma de observações positivas apresentar uma cauda direita longa, transformar as observações usando logaritmo ou raiz quadrada pode ajudar na sime-

tria ao redor da média e na aproximação para a normalidade. Muitas vezes, a escolha da transformação que deve ser realizada para aproximar os dados da normalidade não é óbvia, sendo conveniente deixar os dados sugerirem uma transformação. Uma família de transformações úteis para este propósito é a família de *transformações potência*, que são definidas apenas para variáveis positivas. Entretanto, esta restrição pode ser contornada adicionando-se uma única constante a cada observação no conjunto de dados se alguns dos valores forem negativos.

Seja  $\mathbf{x}$  um vetor de observação de uma variável  $x$ . Para selecionar uma transformação potência, deve-se observar o histograma e decidir se a quantidade de valores grandes devem ser diminuída ou aumentada para aumentar a simetria ao redor da média. As transformações potência são feitas a partir de uma constante  $\lambda \in \mathbb{R}$ . Por exemplo, considerar  $x^\lambda$  e escolher  $\lambda = -1$  corresponde à transformação da inversa; para  $\lambda = 0$ , é definido  $x^0 = \ln(x)$ . Para aumentar a quantidade de valores grandes de  $x$ , utiliza-se valores para  $\lambda$  menores que 1, como por exemplo,  $x^{-1}$ ,  $\ln(x)$ ,  $x^{1/4}$ ,  $x^{1/2}$ . Já para diminuir a quantidade de valores grandes de  $x$ , utiliza-se valores para  $\lambda$  maiores que 1, como por exemplo,  $x^2$ ,  $x^3$ . A escolha final deve ser examinada a partir de um Q-Q plot ou outras verificações de normalidade. Posteriormente, as transformações anteriormente citadas devem ser realizadas para que seja selecionada a melhor.

Para que seja realizada a análise discriminante discutida, é necessário que os dados possuam distribuição normal. Assim, caso eles não possuam, são realizadas transformações nos dados com o intuito de aproximação da normalidade. Parte dos dados é utilizada para que seja definida uma regra de classificação e outra parte é classificada de acordo com a regra definida.

No caso em que sejam considerados dois grupos e não ocorra a normalidade, devem ser aplicadas transformações aos conjuntos de dados das variáveis de cada grupo. Antes de uma nova observação ser classificada de acordo com a regra, deve também ser aplicada nesta a mesma transformação utilizada nos dados que geraram a regra, para que a escala seja mantida. Desta forma, não podem ser utilizadas duas transformações diferentes em uma variável de cada grupo, sendo necessário buscar uma que aproxime da normalidade a variável nos dois grupos.

## 4 Simulações

A realização de simulações facilita a visualização e comparação de cenários possíveis. Foram consideradas as seguintes situações:

- Matrizes de covariância iguais ( $\Sigma_1 = \Sigma_2 = \Sigma$ )
  - Existência de correlação entre as variáveis (Com correlação)
  - Não existência de correlação entre as variáveis (Sem correlação)
- Matrizes de covariância diferentes ( $\Sigma_1 \neq \Sigma_2$ )
  - Existência de correlação entre as variáveis (Com correlação)
  - Não existência de correlação entre as variáveis (Sem correlação)

Dentro de cada situação, foram variadas a média e a matriz de covariância e geradas amostras de duas normais multivariadas de tamanho 100. A média da população  $\pi_1$  foi fixada como  $\mu_1 = [0 \ 0]^\top$  para que a média da população  $\pi_2$ ,  $\mu_2$ , representasse a distância entre as médias.

Com o intuito de estimar a Probabilidade Total de Classificação Incorreta, foi calculada, para cada cenário possível, a Taxa de Erro Aparentes (TEA) através dos métodos de avaliação das regras de classificação explicitados anteriormente, considerando custos iguais e prioris iguais. No caso em que as matrizes de covariância são iguais, a TEA foi calculada através dos métodos de Ressubstituição (R), Ressubstituição com Divisão Amostral (RDA), Pseudo-*jackknife* (JACK), Probabilidade de Classificação Incorreta Estimada (PCIE) e Método Dois de Lachenbruch e Mickey (LACH). Já no caso em que as matrizes de covariância são diferentes, a TEA foi calculada somente através dos métodos de Ressubstituição (R), Ressubstituição com Divisão Amostral (RDA) e Pseudo-*jackknife* (JACK), já que os outros dois métodos consideram uma variância comum, o que pode levar a uma estimativa ruim.

## 4.1 Matrizes de covariância iguais ( $\Sigma_1 = \Sigma_2 = \Sigma$ )

No caso em que as matrizes de covariância são iguais, houve interesse em observar o comportamento dos dados no caso em que há existência de correlação entre as variáveis e no caso em que não há. Em cada caso, a média da população  $\pi_2$  foi variada para que fosse possível comparar as regras de classificação obtidas nas diferentes distâncias entre as médias  $\mu_1$  e  $\mu_2$ . A matriz de covariância  $\Sigma$  também foi variada com o propósito de comparar as regras de classificação obtidas em dados com diferentes variabilidades.

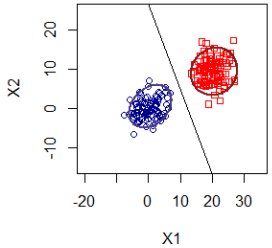
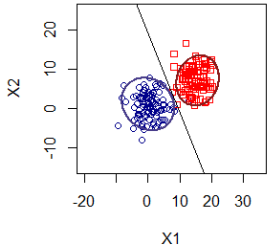
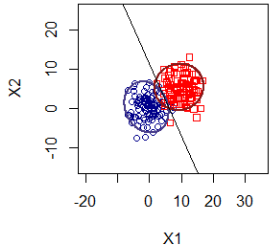
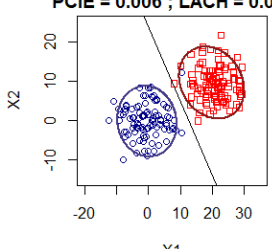
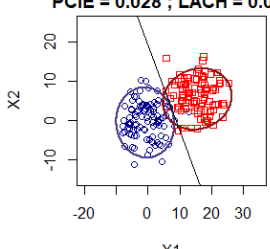
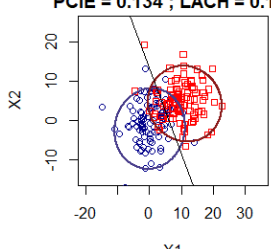
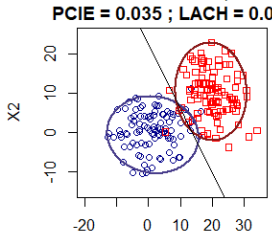
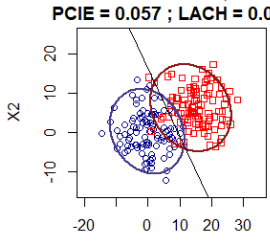
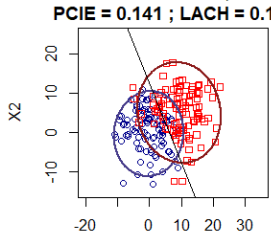
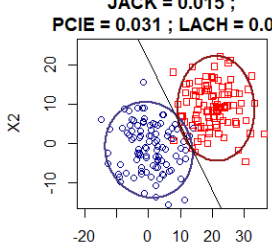
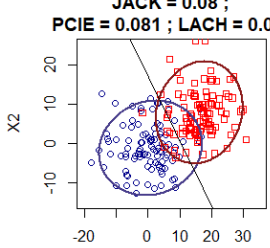
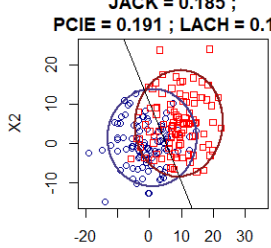
Para cada cenário possível, a amostra obtida da população  $\pi_1$  foi graficada em azul, já a amostra obtida da população  $\pi_2$  foi graficada em vermelho. Além disso, em volta de cada amostra foram graficadas, em azul escuro e em vermelho escuro, respectivamente, as elipses que cobrem 95% dos valores de cada amostra. Como as matrizes de covariância são iguais, a regra de classificação é representada por uma reta, que pode ser observada em preto. Os estimadores da PTCI calculados a partir dos métodos R, RDA, JACK, PCIE e LACH foram informados acima dos gráficos correspondentes.

### 4.1.1 Sem correlação entre as variáveis

No caso em que não há correlação entre as variáveis, a matriz de covariância  $\Sigma$  é representada por  $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ , sendo  $\sigma_1^2$  e  $\sigma_2^2$  as variâncias das populações  $\pi_1$  e  $\pi_2$ , respectivamente. Neste caso, a variabilidade dos dados depende exclusivamente destes valores.

Na Tabela 4, é possível observar que, considerando as mesmas médias, quanto maior a variabilidade dos dados, maior a TEA. Além disso, ao fixar as matrizes de covariância, observa-se que as TEA aumentam à medida que as médias se aproximam.

Tabela 4: Cenários possíveis no caso em que as matrizes de covariância são iguais e não há existência de correlação entre as variáveis

	$\mu_1 = [0 \ 0]^T$ , $\mu_2 = [20 \ 10]^T$	$\mu_1 = [0 \ 0]^T$ , $\mu_2 = [15 \ 7]^T$	$\mu_1 = [0 \ 0]^T$ , $\mu_2 = [10 \ 5]^T$
$\Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$	<b>R = 0 ; RDA = 0 ;</b> <b>JACK = 0 ;</b> <b>PCIE = 0 ; LACH = 0</b> 	<b>R = 0.005 ; RDA = 0.01 ;</b> <b>JACK = 0.01 ;</b> <b>PCIE = 0.009 ; LACH = 0.01</b> 	<b>R = 0.035 ; RDA = 0.05 ;</b> <b>JACK = 0.035 ;</b> <b>PCIE = 0.032 ; LACH = 0.033</b> 
$\Sigma = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}$	<b>R = 0.005 ; RDA = 0.01 ;</b> <b>JACK = 0.005 ;</b> <b>PCIE = 0.006 ; LACH = 0.006</b> 	<b>R = 0.015 ; RDA = 0.02 ;</b> <b>JACK = 0.015 ;</b> <b>PCIE = 0.028 ; LACH = 0.029</b> 	<b>R = 0.12 ; RDA = 0.12 ;</b> <b>JACK = 0.125 ;</b> <b>PCIE = 0.134 ; LACH = 0.137</b> 
$\Sigma = \begin{bmatrix} 30 & 0 \\ 0 & 30 \end{bmatrix}$	<b>R = 0.04 ; RDA = 0.04 ;</b> <b>JACK = 0.04 ;</b> <b>PCIE = 0.035 ; LACH = 0.036</b> 	<b>R = 0.05 ; RDA = 0.05 ;</b> <b>JACK = 0.06 ;</b> <b>PCIE = 0.057 ; LACH = 0.059</b> 	<b>R = 0.15 ; RDA = 0.14 ;</b> <b>JACK = 0.15 ;</b> <b>PCIE = 0.141 ; LACH = 0.143</b> 
$\Sigma = \begin{bmatrix} 40 & 0 \\ 0 & 40 \end{bmatrix}$	<b>R = 0.015 ; RDA = 0.03 ;</b> <b>JACK = 0.015 ;</b> <b>PCIE = 0.031 ; LACH = 0.032</b> 	<b>R = 0.08 ; RDA = 0.05 ;</b> <b>JACK = 0.08 ;</b> <b>PCIE = 0.081 ; LACH = 0.083</b> 	<b>R = 0.185 ; RDA = 0.17 ;</b> <b>JACK = 0.185 ;</b> <b>PCIE = 0.191 ; LACH = 0.193</b> 



### 4.1.2 Com correlação entre as variáveis

No caso em que há correlação entre as variáveis, a variabilidade dos dados depende não só das variâncias das variáveis, mas também da correlação entre elas. A Tabela 5 sugere que correlações positivas geram elipses “apontando” para as direções nordeste e sudoeste e correlações negativas, para o noroeste e sudeste. Isso ocorre graças aos autovetores da matriz de covariância e quanto maior a correlação em módulo, mais achatada é a elipse.

É possível observar também que no caso em que  $\Sigma = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$ , apesar da distância entre  $\mu_1 = [0 \ 0]^\top$  e  $\mu_2 = [5 \ 5]^\top$  ser menor do que a distância entre  $\mu_1 = [0 \ 0]^\top$  e  $\mu_2 = [5 \ 10]^\top$ , observa-se que no segundo caso as estimativas da PTCI são maiores. Isto ocorre pois para os valores de média  $\mu_1 = [0 \ 0]^\top$  e  $\mu_2 = [5 \ 10]^\top$  as populações se sobrepõem, o que dificulta a classificação correta dos dados.

Já na Tabela 6, observa-se TEA relativamente alta na ocorrência de médias próximas e variabilidades altas, e também quando as populações coincidem em uma área grande, como é o caso de quando  $\mu_1 = [0 \ 0]^\top$ ,  $\mu_2 = [10 \ 5]^\top$  e  $\Sigma = \begin{bmatrix} 20 & 14 \\ 14 & 20 \end{bmatrix}$ .

Conforme a variabilidade dos dados aumenta, as simulações são feitas considerando médias mais distantes para que seja possível fazer uma comparação dos estimadores da PTCI que tenha sentido. No caso em que as médias fossem muito distantes e a variabilidade dos dados muito pequena, obteria-se TEA nula. Já no caso em que as médias sejam muito próximas e a variabilidade dos dados muito grande, obteria-se TEA muito grande, significando que a regra de classificação obtida não é boa para classificar os dados considerados.

Tabela 5: Cenários possíveis no caso em que as matrizes de covariância são iguais e há existência de correlação entre as variáveis, fixando  $\sigma_1^2 = \sigma_2^2 = 10$ 

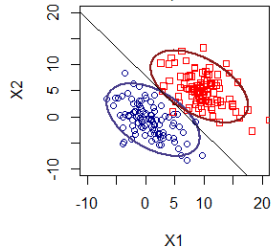
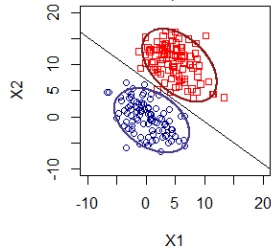
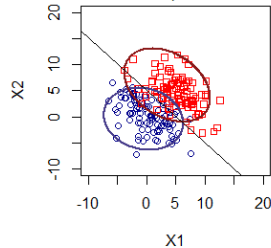
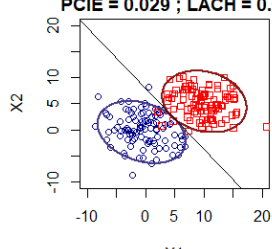
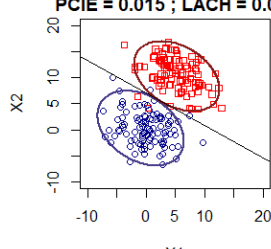
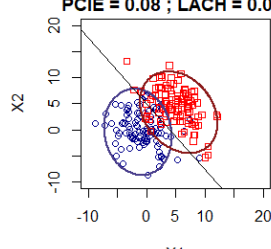
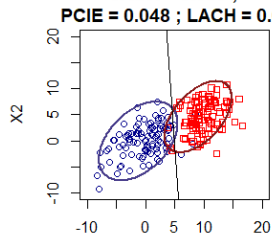
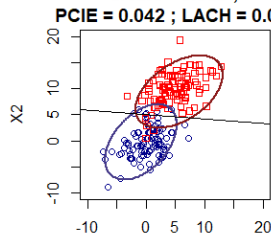
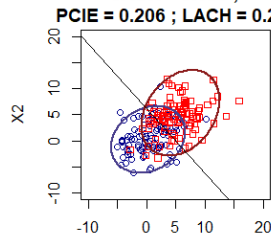
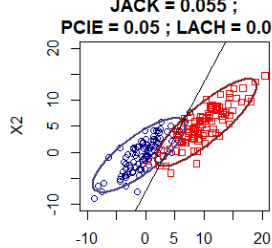
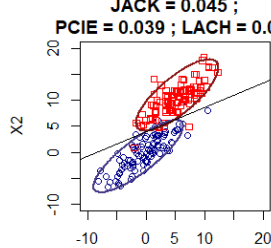
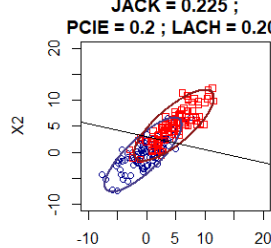
	$\mu_1 = [0 \ 0]^\top$ , $\mu_2 = [10 \ 5]^\top$	$\mu_1 = [0 \ 0]^\top$ , $\mu_2 = [5 \ 10]^\top$	$\mu_1 = [0 \ 0]^\top$ , $\mu_2 = [5 \ 5]^\top$
$\Sigma = \begin{bmatrix} 10 & -5 \\ -5 & 10 \end{bmatrix}$	<b>R = 0.01 ; RDA = 0.01 ;</b> <b>JACK = 0.01 ;</b> <b>PCIE = 0.011 ; LACH = 0.012</b> 	<b>R = 0.005 ; RDA = 0.01 ;</b> <b>JACK = 0.005 ;</b> <b>PCIE = 0.008 ; LACH = 0.009</b> 	<b>R = 0.07 ; RDA = 0.09 ;</b> <b>JACK = 0.08 ;</b> <b>PCIE = 0.073 ; LACH = 0.075</b> 
$\Sigma = \begin{bmatrix} 10 & -3 \\ -3 & 10 \end{bmatrix}$	<b>R = 0.02 ; RDA = 0.02 ;</b> <b>JACK = 0.025 ;</b> <b>PCIE = 0.029 ; LACH = 0.03</b> 	<b>R = 0.025 ; RDA = 0.01 ;</b> <b>JACK = 0.025 ;</b> <b>PCIE = 0.015 ; LACH = 0.016</b> 	<b>R = 0.095 ; RDA = 0.11 ;</b> <b>JACK = 0.1 ;</b> <b>PCIE = 0.08 ; LACH = 0.082</b> 
$\Sigma = \begin{bmatrix} 10 & 3 \\ 3 & 10 \end{bmatrix}$	<b>R = 0.035 ; RDA = 0.01 ;</b> <b>JACK = 0.035 ;</b> <b>PCIE = 0.048 ; LACH = 0.05</b> 	<b>R = 0.055 ; RDA = 0.04 ;</b> <b>JACK = 0.055 ;</b> <b>PCIE = 0.042 ; LACH = 0.044</b> 	<b>R = 0.195 ; RDA = 0.16 ;</b> <b>JACK = 0.195 ;</b> <b>PCIE = 0.206 ; LACH = 0.209</b> 
$\Sigma = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$	<b>R = 0.05 ; RDA = 0.04 ;</b> <b>JACK = 0.055 ;</b> <b>PCIE = 0.05 ; LACH = 0.051</b> 	<b>R = 0.045 ; RDA = 0.05 ;</b> <b>JACK = 0.045 ;</b> <b>PCIE = 0.039 ; LACH = 0.041</b> 	<b>R = 0.225 ; RDA = 0.2 ;</b> <b>JACK = 0.225 ;</b> <b>PCIE = 0.2 ; LACH = 0.201</b> 

Tabela 6: Cenários possíveis no caso em que as matrizes de covariância são iguais e há existência de correlação entre as variáveis, fixando  $\sigma_1^2 = \sigma_2^2 = 20$ 

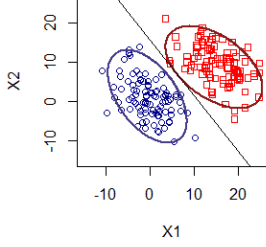
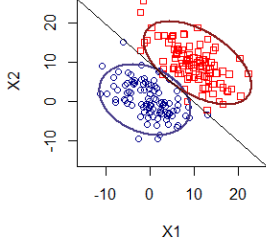
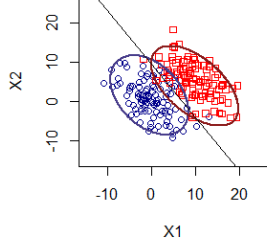
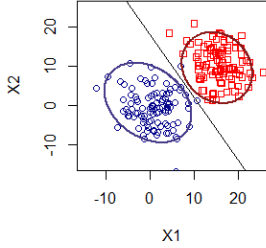
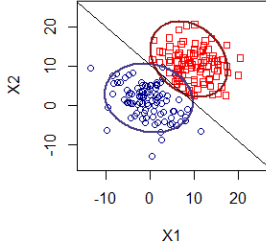
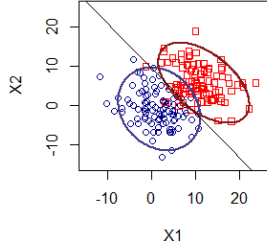
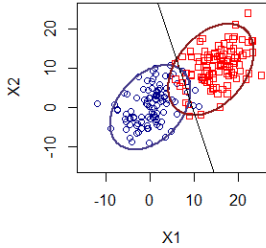
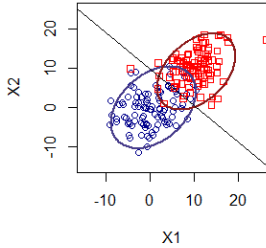
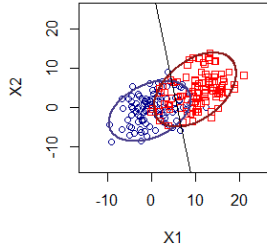
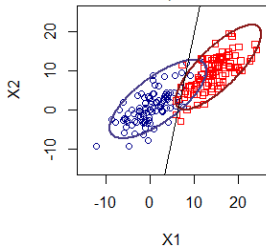
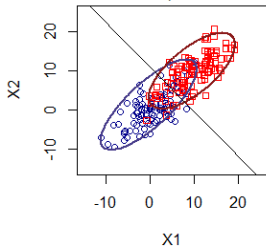
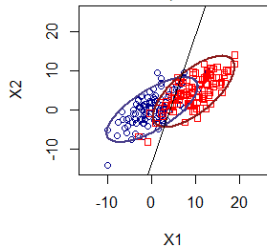
	$\mu_1 = [0 \ 0]^T$ , $\mu_2 = [15 \ 10]^T$	$\mu_1 = [0 \ 0]^T$ , $\mu_2 = [10 \ 10]^T$	$\mu_1 = [0 \ 0]^T$ , $\mu_2 = [10 \ 5]^T$
$\Sigma = \begin{bmatrix} 20 & -10 \\ -10 & 20 \end{bmatrix}$	<b>R = 0 ; RDA = 0 ;</b> <b>JACK = 0 ;</b> <b>PCIE = 0.005 ; LACH = 0.005</b> 	<b>R = 0.015 ; RDA = 0.03 ;</b> <b>JACK = 0.03 ;</b> <b>PCIE = 0.018 ; LACH = 0.019</b> 	<b>R = 0.06 ; RDA = 0.05 ;</b> <b>JACK = 0.06 ;</b> <b>PCIE = 0.04 ; LACH = 0.041</b> 
$\Sigma = \begin{bmatrix} 20 & -5 \\ -5 & 20 \end{bmatrix}$	<b>R = 0.01 ; RDA = 0.01 ;</b> <b>JACK = 0.01 ;</b> <b>PCIE = 0.009 ; LACH = 0.009</b> 	<b>R = 0.01 ; RDA = 0.01 ;</b> <b>JACK = 0.01 ;</b> <b>PCIE = 0.024 ; LACH = 0.025</b> 	<b>R = 0.06 ; RDA = 0.05 ;</b> <b>JACK = 0.06 ;</b> <b>PCIE = 0.06 ; LACH = 0.062</b> 
$\Sigma = \begin{bmatrix} 20 & 8 \\ 8 & 20 \end{bmatrix}$	<b>R = 0.075 ; RDA = 0.07 ;</b> <b>JACK = 0.075 ;</b> <b>PCIE = 0.044 ; LACH = 0.045</b> 	<b>R = 0.065 ; RDA = 0.08 ;</b> <b>JACK = 0.065 ;</b> <b>PCIE = 0.074 ; LACH = 0.076</b> 	<b>R = 0.14 ; RDA = 0.11 ;</b> <b>JACK = 0.14 ;</b> <b>PCIE = 0.125 ; LACH = 0.127</b> 
$\Sigma = \begin{bmatrix} 20 & 14 \\ 14 & 20 \end{bmatrix}$	<b>R = 0.05 ; RDA = 0.05 ;</b> <b>JACK = 0.05 ;</b> <b>PCIE = 0.046 ; LACH = 0.047</b> 	<b>R = 0.145 ; RDA = 0.13 ;</b> <b>JACK = 0.145 ;</b> <b>PCIE = 0.134 ; LACH = 0.136</b> 	<b>R = 0.14 ; RDA = 0.16 ;</b> <b>JACK = 0.14 ;</b> <b>PCIE = 0.131 ; LACH = 0.134</b> 

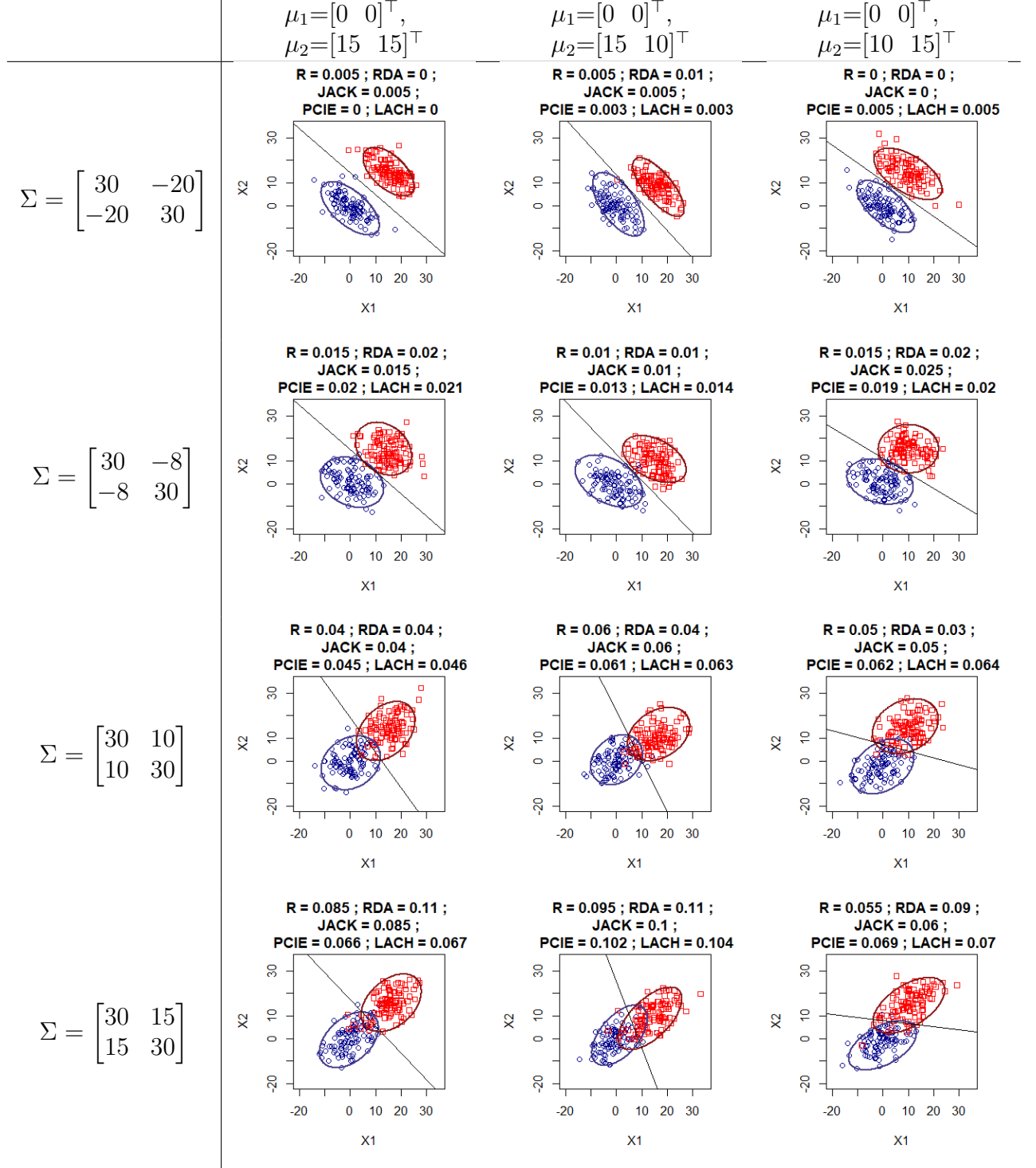
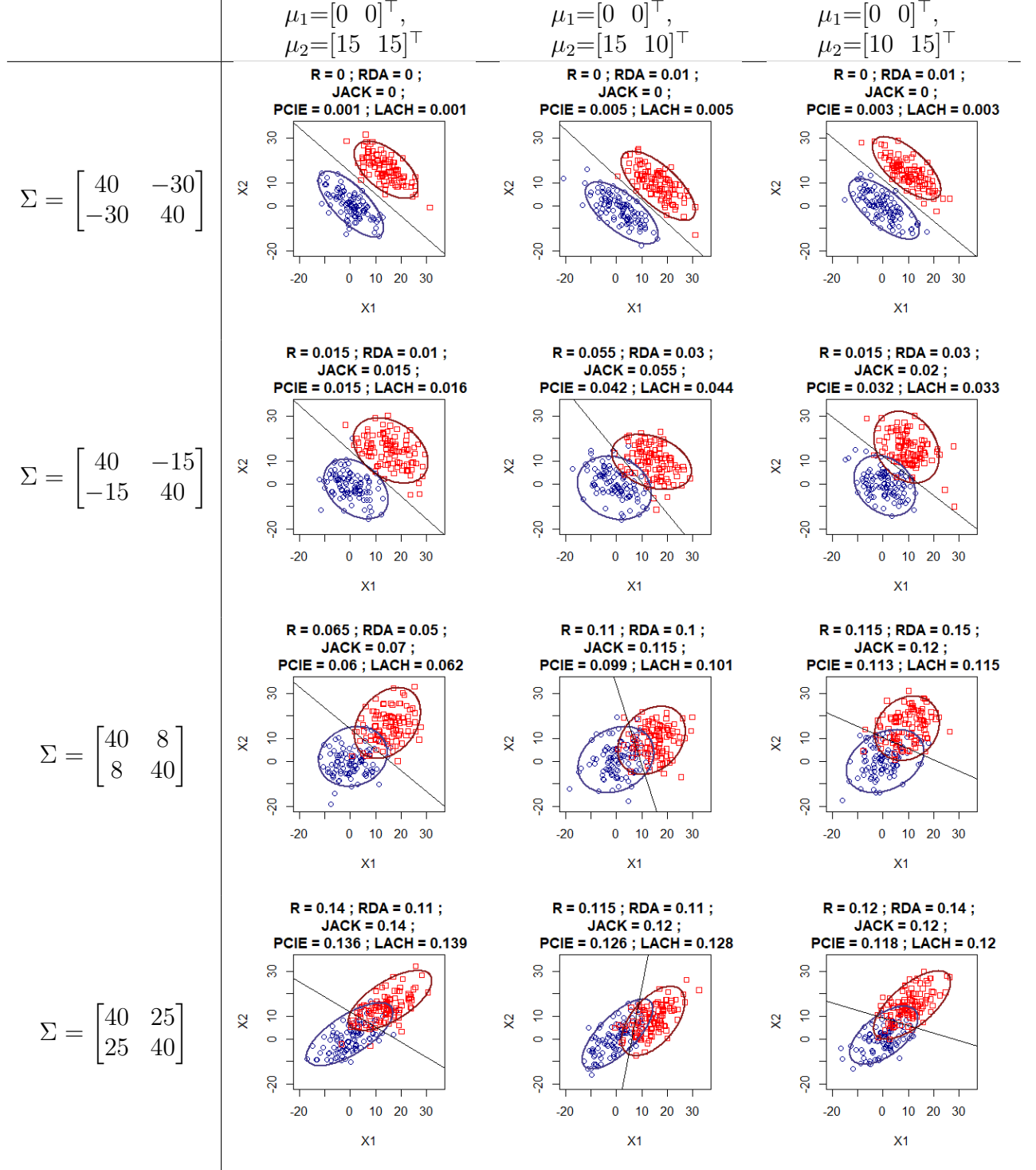
Tabela 7: Cenários possíveis no caso em que as matrizes de covariância são iguais e há existência de correlação entre as variáveis, fixando  $\sigma_1^2 = \sigma_2^2 = 30$ 

Tabela 8: Cenários possíveis no caso em que as matrizes de covariância são iguais e há existência de correlação entre as variáveis, fixando  $\sigma_1^2 = \sigma_2^2 = 40$ 

## 4.2 Matrizes de covariância diferentes ( $\Sigma_1 \neq \Sigma_2$ )

Da mesma forma, no caso em que as matrizes de covariância são iguais, foram considerados o caso em que há existência de correlação entre as variáveis e no caso em que não há. Novamente, em cada caso a média da população  $\pi_2$  foi variada para que fosse possível comparar as regras de classificação obtidas nas diferentes distâncias entre as médias  $\mu_1$  e  $\mu_2$ . As matrizes de covariância foram variadas de forma que, após fixar  $\Sigma_1$ , variou-se  $\Sigma_2$  com a finalidade de comparar as regras de classificação obtidas em diferentes combinações de variabilidades.

Mais uma vez, para cada cenário possível, a amostra obtida da população  $\pi_1$  foi graficada em azul, já a amostra obtida da população  $\pi_2$  foi graficada em vermelho. Em volta de cada amostra foram graficadas, em azul escuro e em vermelho escuro, respectivamente, as elipses que cobrem 95% dos valores de cada amostra. Como as matrizes de covariância são diferentes, a regra de classificação é representada por uma cônica (elipse, hipérbole ou parábola), que pode ser observada em verde claro. Os estimadores da PTCI calculados a partir dos métodos R, RDA e JACK foram informados acima dos gráficos correspondentes.

### 4.2.1 Sem correlação entre as variáveis

No caso em que não há correlação entre as variáveis, a variabilidade dos dados depende exclusivamente das variâncias de cada variável nas duas populações.

De forma similar, conforme a variabilidade dos dados aumenta, as simulações são feitas considerando médias mais distantes para que seja possível fazer uma comparação dos estimadores da PTCI que tenha sentido. No caso em que as médias fossem muito distantes e a variabilidade dos dados muito pequena, obter-se-ia TEA nula. Já no caso em que as médias sejam muito próximas e a variabilidade dos dados muito grande, obter-se-ia TEA muito grande, significando que a regra de classificação obtida não é boa para classificar os dados considerados.

Tabela 9: Cenários possíveis no caso em que as matrizes de covariância são diferentes e não há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$ .

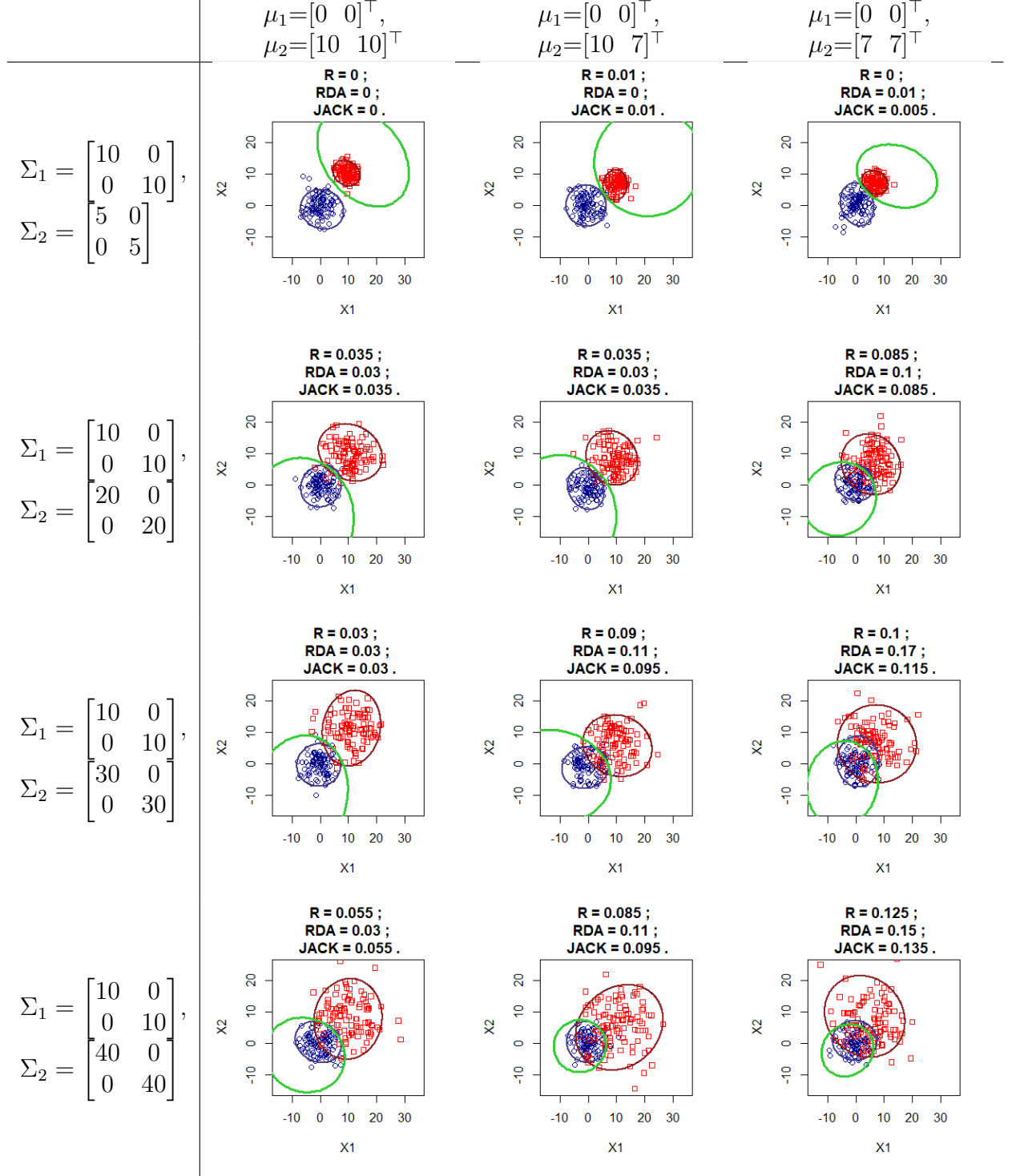


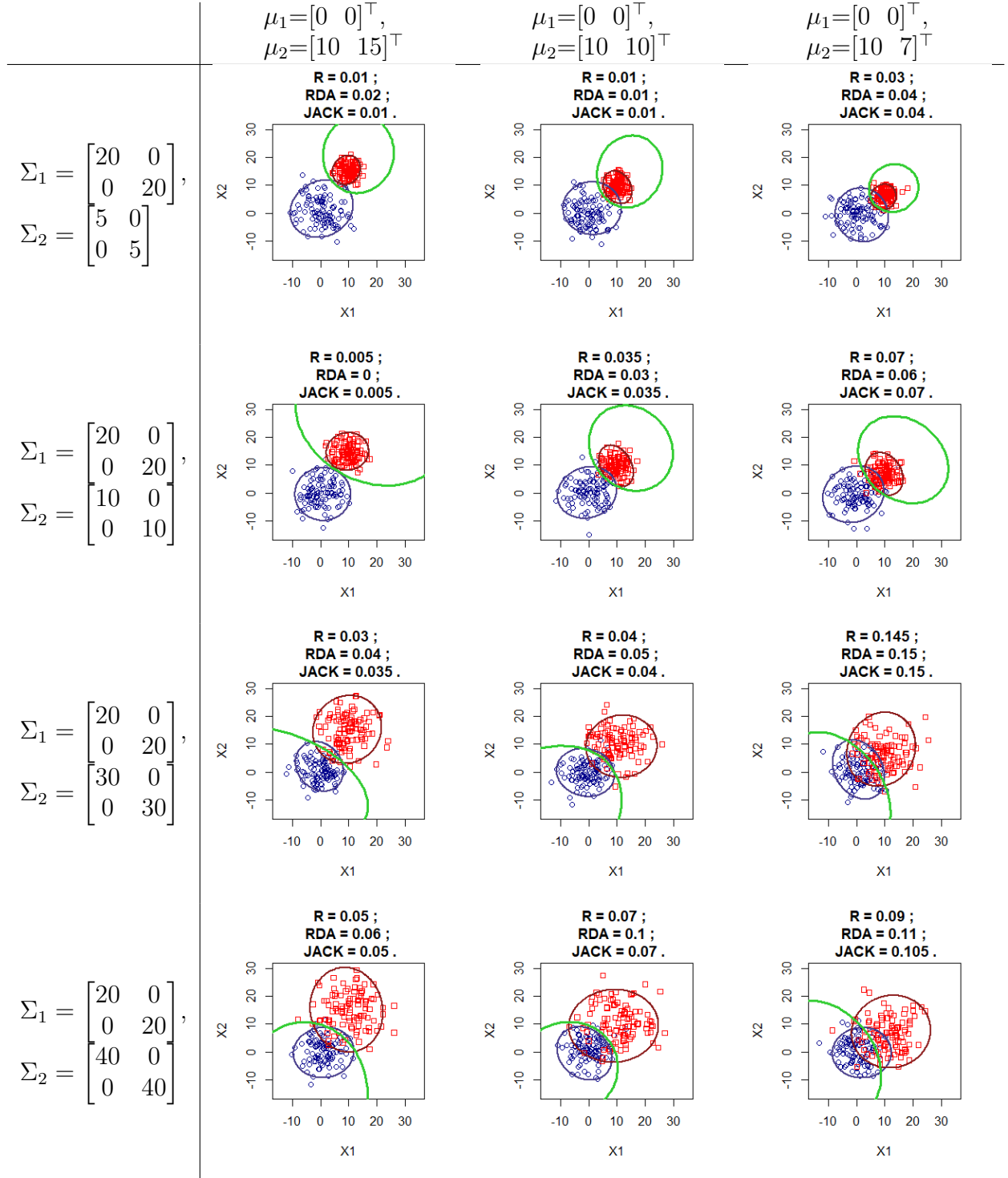
Tabela 10: Cenários possíveis no caso em que as matrizes de covariância são diferentes e não há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}$ .



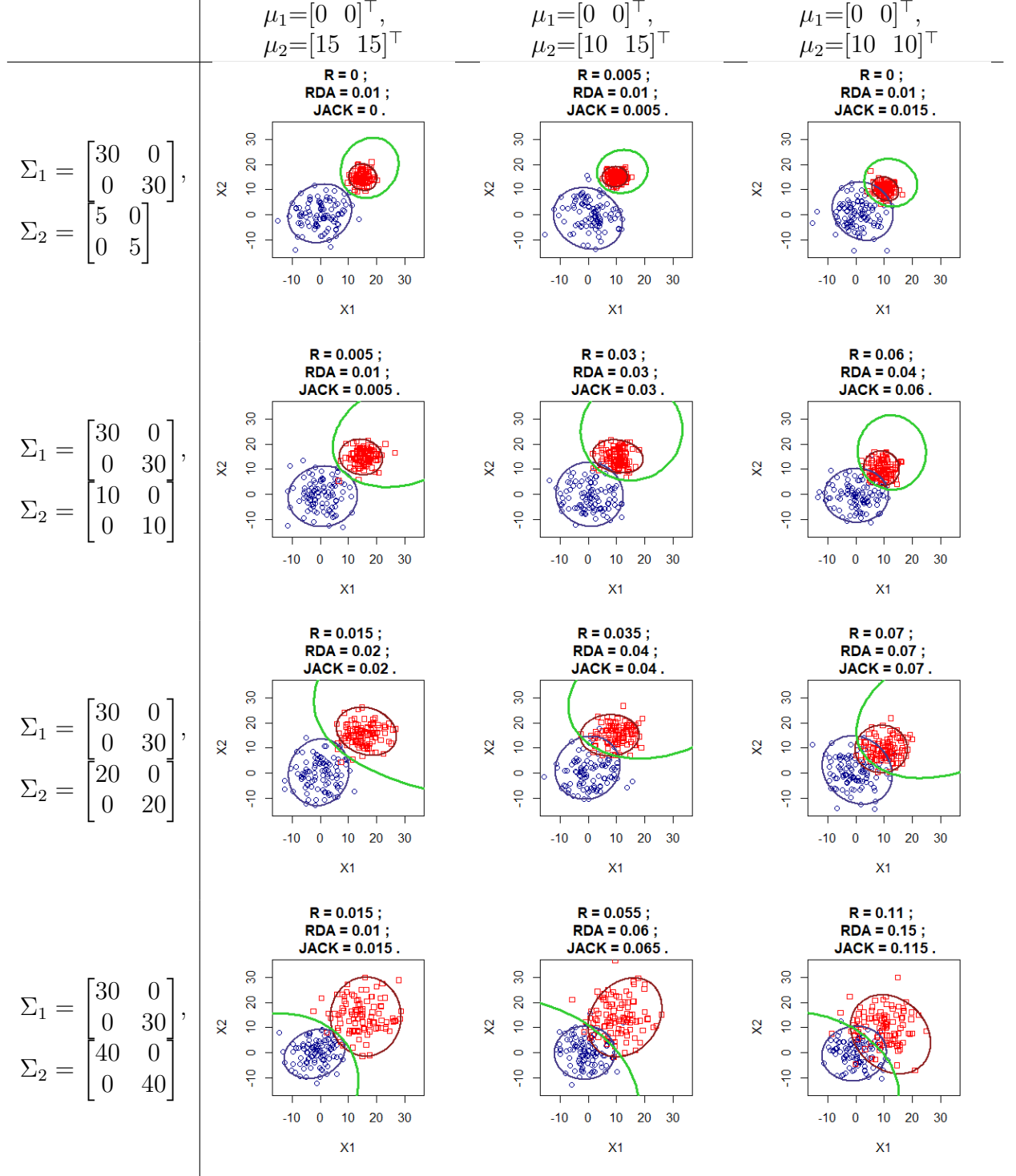
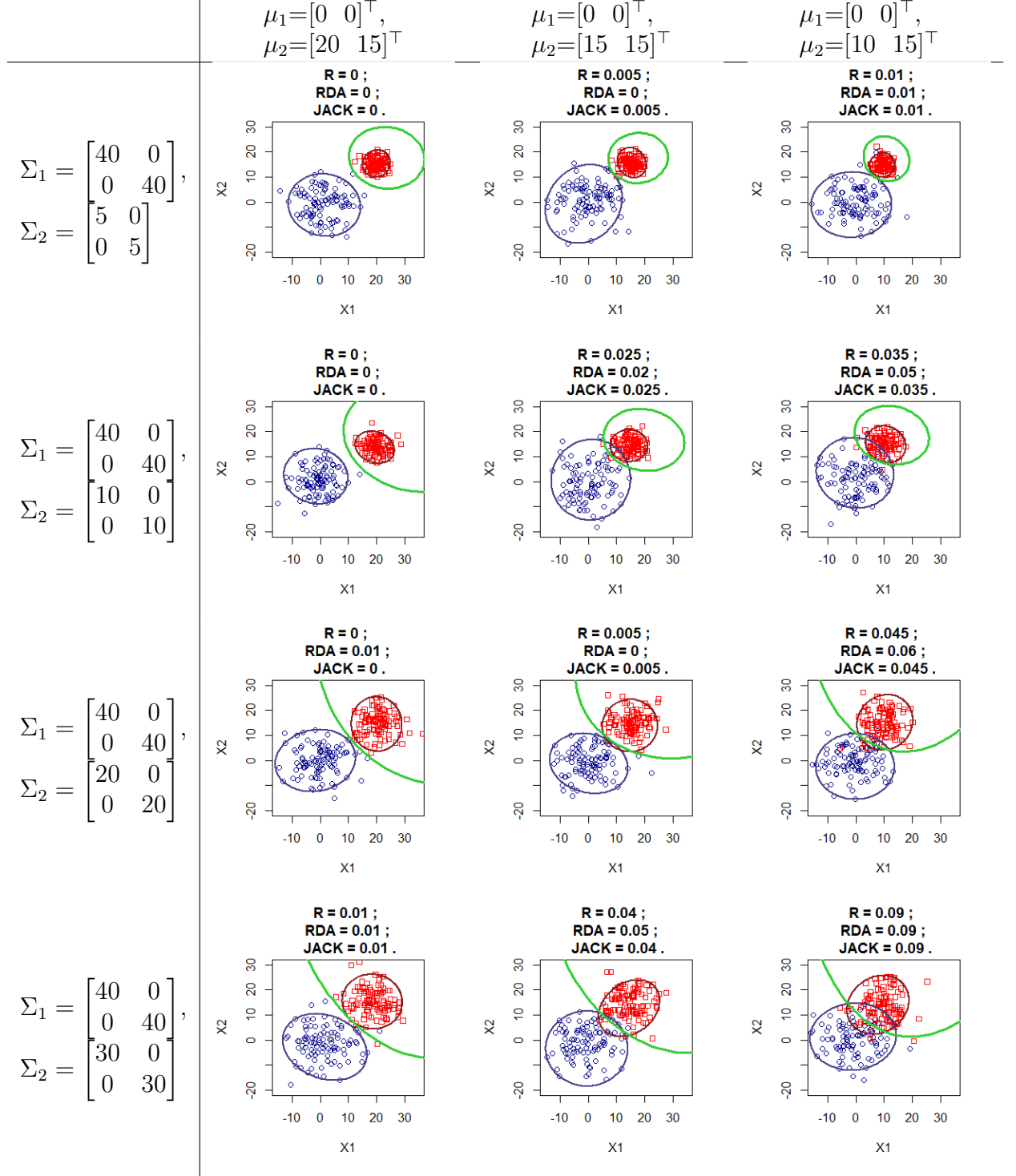
Tabela 11: Cenários possíveis no caso em que as matrizes de covariância são diferentes e não há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 30 & 0 \\ 0 & 30 \end{bmatrix}$ .

Tabela 12: Cenários possíveis no caso em que as matrizes de covariância são diferentes e não há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 40 & 0 \\ 0 & 40 \end{bmatrix}$ .



### 4.2.2 Com correlação entre as variáveis

Novamente, a variabilidade dos dados depende não só das variâncias das variáveis, mas também da correlação entre elas. Graças aos autovetores da matriz de covariância, correlações altas em módulo entre as variáveis fazem com que a elipse seja mais achatada, de forma que correlações positivas geram elipses “apontando” para as direções nordeste e sudoeste e correlações negativas, para o noroeste e sudeste.

Na Tabela 12, observa-se que, fixando as médias de dois dados cujas variáveis possuam correlações grandes em módulo e com sinais opostos (uma com correlação positiva e outra negativa), tendem a se confundir mais, já que “apontam” em sentidos opostos, se apresentando paralelos, enquanto no caso de correlações de mesmo sinal os dados se dispõem de forma paralela, havendo menor confusão.

Já no caso em que as mesmas matrizes de covariâncias são fixadas, quando as correlações forem grande em módulo e possuírem sinais oposto a distância entre as médias deve ser grande para que não haja muita interseção e a regra de classificação separe bem as populações. Caso os sinais das correlações sejam iguais, para uma boa classificação a diferença entre as médias deve ser grande caso estejam alinhadas.

Tabela 13: Cenários possíveis no caso em que as matrizes de covariância são diferentes e há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 10 & -5 \\ -5 & 10 \end{bmatrix}$ .

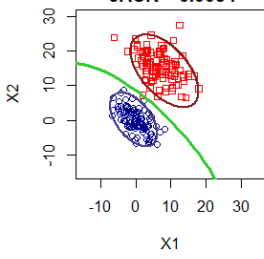
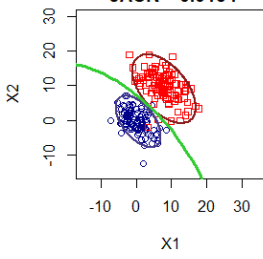
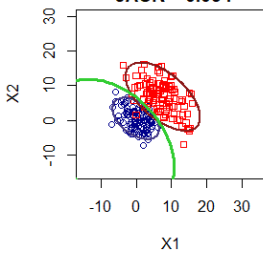
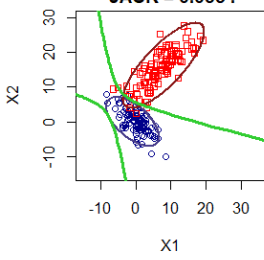
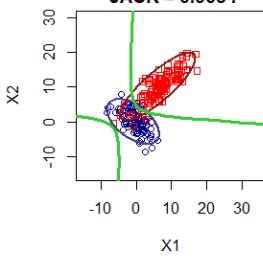
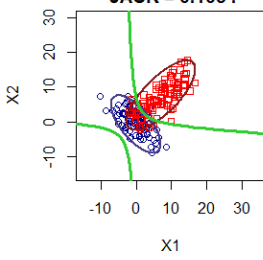
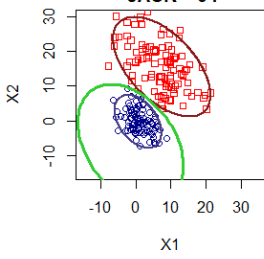
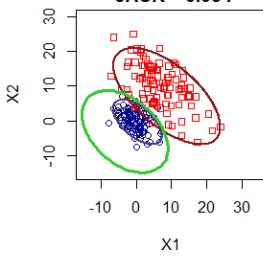
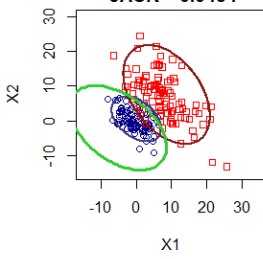
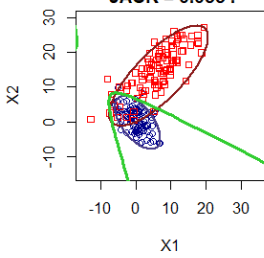
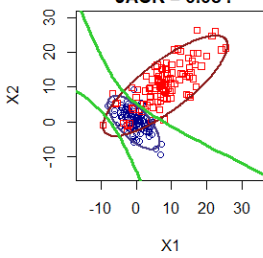
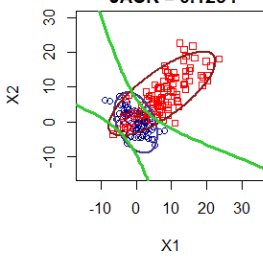
	$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ , $\mu_2 = \begin{bmatrix} 7 & 15 \end{bmatrix}^\top$	$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ , $\mu_2 = \begin{bmatrix} 7 & 10 \end{bmatrix}^\top$	$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ , $\mu_2 = \begin{bmatrix} 7 & 7 \end{bmatrix}^\top$
$\Sigma_1 = \begin{bmatrix} 10 & -5 \\ -5 & 10 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 20 & -10 \\ -10 & 20 \end{bmatrix}$	<b>R = 0 ;</b> <b>RDA = 0 ;</b> <b>JACK = 0.005 .</b> 	<b>R = 0.015 ;</b> <b>RDA = 0.03 ;</b> <b>JACK = 0.015 .</b> 	<b>R = 0.02 ;</b> <b>RDA = 0.01 ;</b> <b>JACK = 0.03 .</b> 
$\Sigma_1 = \begin{bmatrix} 10 & -5 \\ -5 & 10 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 20 & 15 \\ 15 & 20 \end{bmatrix}$	<b>R = 0.035 ;</b> <b>RDA = 0.02 ;</b> <b>JACK = 0.035 .</b> 	<b>R = 0.06 ;</b> <b>RDA = 0.04 ;</b> <b>JACK = 0.065 .</b> 	<b>R = 0.105 ;</b> <b>RDA = 0.13 ;</b> <b>JACK = 0.105 .</b> 
$\Sigma_1 = \begin{bmatrix} 10 & -5 \\ -5 & 10 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 40 & -20 \\ -20 & 40 \end{bmatrix}$	<b>R = 0 ;</b> <b>RDA = 0 ;</b> <b>JACK = 0 .</b> 	<b>R = 0.03 ;</b> <b>RDA = 0.03 ;</b> <b>JACK = 0.03 .</b> 	<b>R = 0.045 ;</b> <b>RDA = 0 ;</b> <b>JACK = 0.045 .</b> 
$\Sigma_1 = \begin{bmatrix} 10 & -5 \\ -5 & 10 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 40 & 30 \\ 30 & 40 \end{bmatrix}$	<b>R = 0.045 ;</b> <b>RDA = 0.1 ;</b> <b>JACK = 0.055 .</b> 	<b>R = 0.08 ;</b> <b>RDA = 0.11 ;</b> <b>JACK = 0.08 .</b> 	<b>R = 0.12 ;</b> <b>RDA = 0.14 ;</b> <b>JACK = 0.125 .</b> 

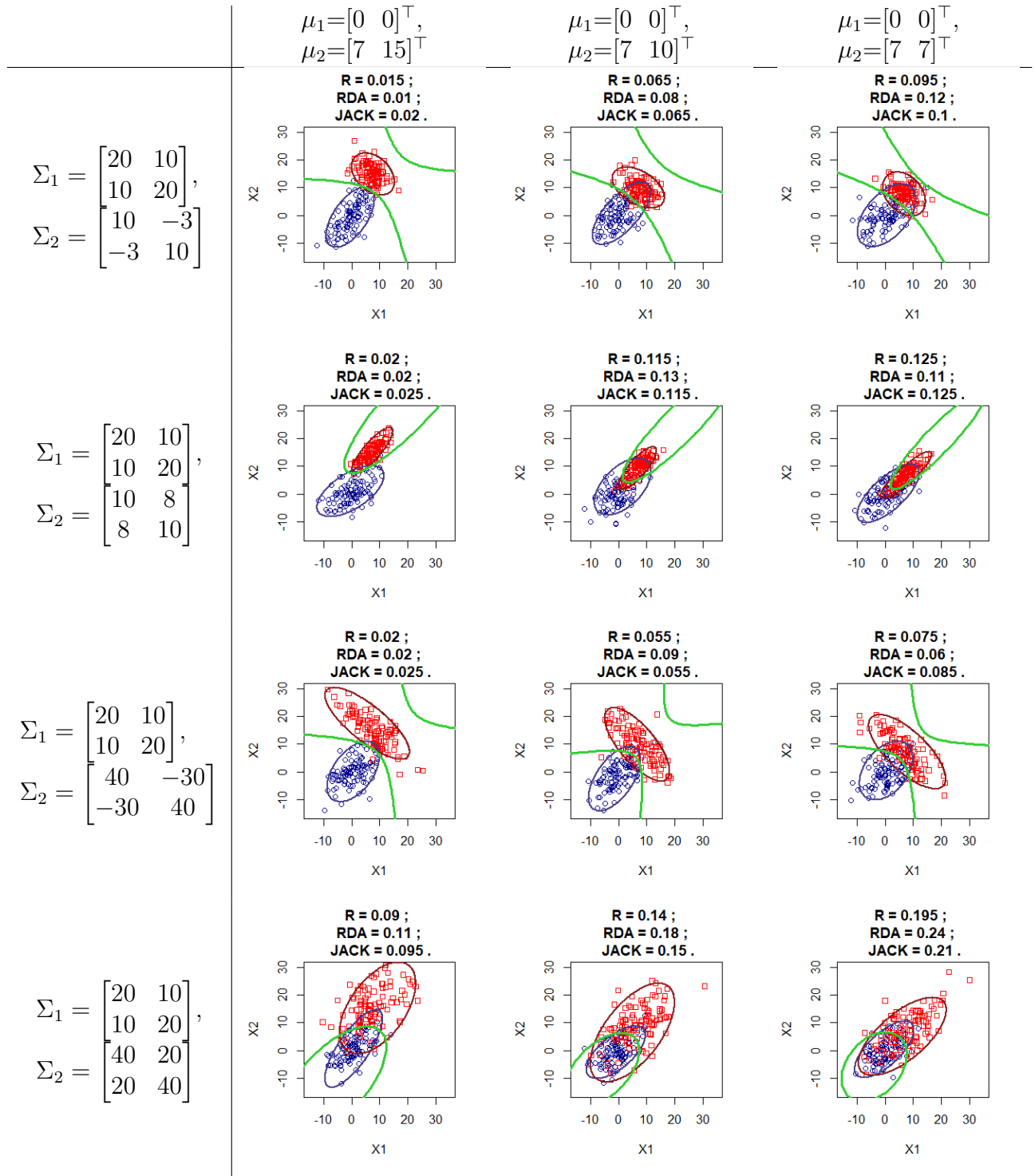
Tabela 14: Cenários possíveis no caso em que as matrizes de covariância são diferentes e há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 20 & 10 \\ 10 & 20 \end{bmatrix}$ .

Tabela 15: Cenários possíveis no caso em que as matrizes de covariância são diferentes e há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 30 & -20 \\ -20 & 30 \end{bmatrix}$ .

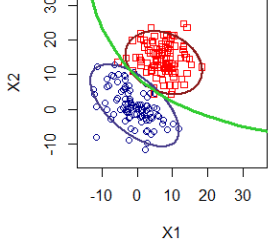
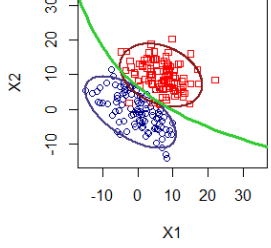
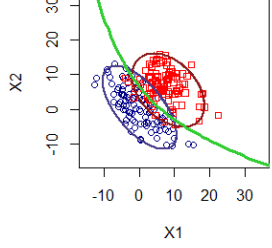
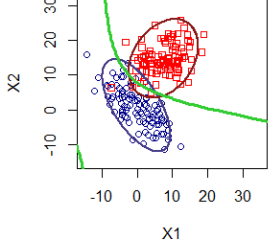
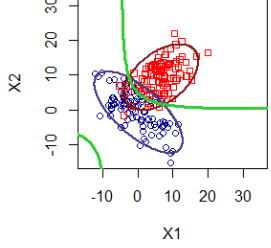
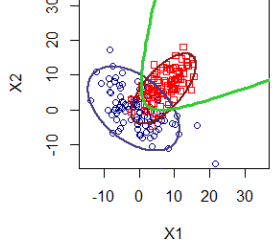
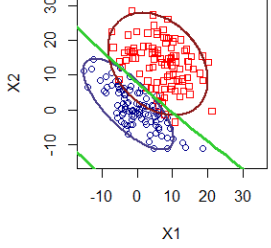
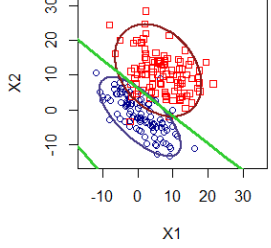
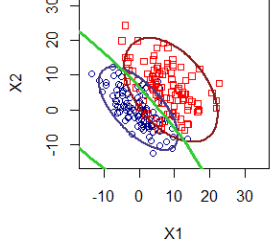
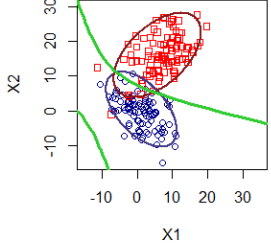
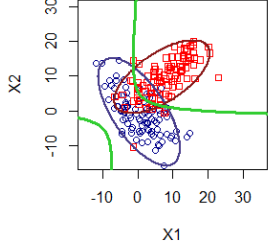
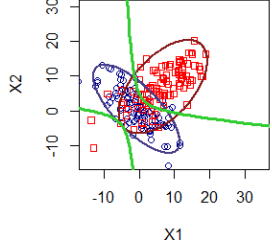
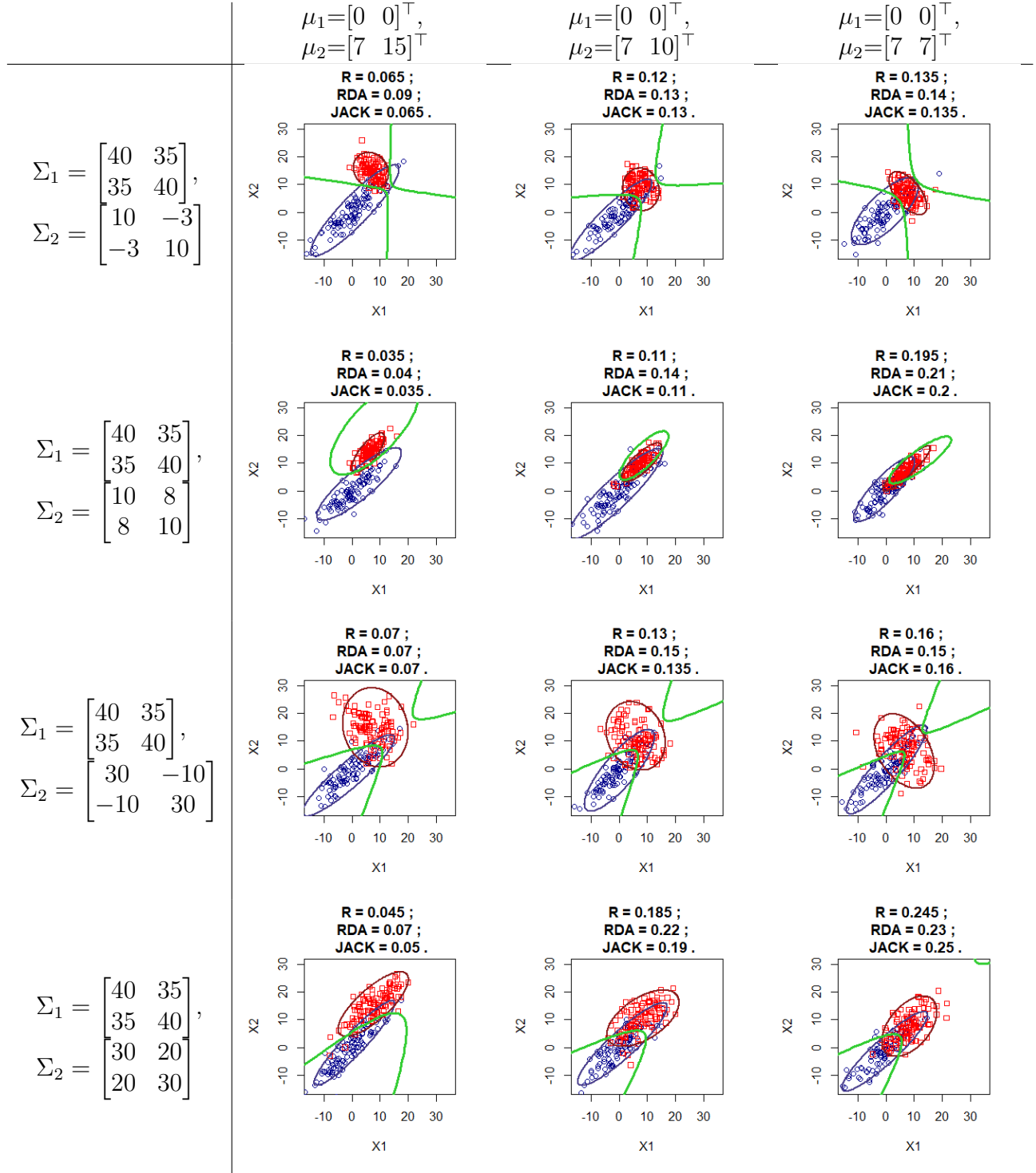
	$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ , $\mu_2 = \begin{bmatrix} 7 & 15 \end{bmatrix}^\top$	$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ , $\mu_2 = \begin{bmatrix} 7 & 10 \end{bmatrix}^\top$	$\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top$ , $\mu_2 = \begin{bmatrix} 7 & 7 \end{bmatrix}^\top$
$\Sigma_1 = \begin{bmatrix} 30 & -20 \\ -20 & 30 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 20 & -5 \\ -5 & 20 \end{bmatrix}$	<b>R = 0.02 ; RDA = 0.02 ; JACK = 0.02 .</b> 	<b>R = 0.03 ; RDA = 0.01 ; JACK = 0.03 .</b> 	<b>R = 0.08 ; RDA = 0.08 ; JACK = 0.085 .</b> 
$\Sigma_1 = \begin{bmatrix} 30 & -20 \\ -20 & 30 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 20 & 10 \\ 10 & 20 \end{bmatrix}$	<b>R = 0.03 ; RDA = 0.01 ; JACK = 0.03 .</b> 	<b>R = 0.085 ; RDA = 0.1 ; JACK = 0.085 .</b> 	<b>R = 0.105 ; RDA = 0.11 ; JACK = 0.11 .</b> 
$\Sigma_1 = \begin{bmatrix} 30 & -20 \\ -20 & 30 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 40 & -15 \\ -15 & 40 \end{bmatrix}$	<b>R = 0.03 ; RDA = 0.03 ; JACK = 0.035 .</b> 	<b>R = 0.05 ; RDA = 0.07 ; JACK = 0.05 .</b> 	<b>R = 0.095 ; RDA = 0.06 ; JACK = 0.095 .</b> 
$\Sigma_1 = \begin{bmatrix} 30 & -20 \\ -20 & 30 \end{bmatrix}$ , $\Sigma_2 = \begin{bmatrix} 40 & 25 \\ 25 & 40 \end{bmatrix}$	<b>R = 0.03 ; RDA = 0.05 ; JACK = 0.035 .</b> 	<b>R = 0.135 ; RDA = 0.12 ; JACK = 0.14 .</b> 	<b>R = 0.125 ; RDA = 0.15 ; JACK = 0.135 .</b> 

Tabela 16: Cenários possíveis no caso em que as matrizes de covariância são diferentes e há existência de correlação entre as variáveis, fixando  $\Sigma_1 = \begin{bmatrix} 40 & 35 \\ 35 & 40 \end{bmatrix}$ .

## 5 Resultados

Com o intuito de comparar a população de livros infantojuvenis atuais,  $\pi_1$ , e antigos,  $\pi_2$ , foram selecionados dois livros e para cada um deles foram calculadas a quantidade de palavras diferentes por lauda e o tamanho médio das frases por lauda, sendo considerada como uma lauda 1200 caracteres. Os livros utilizados são “Como treinar seu dragão”, um livro infantojuvenil lançado em 2010 escrito por Cressida Cowell e “A ilha do tesouro”, um livro infantojuvenil de 1883 escrito por Robert Louis Stevenson. Com o objetivo de testar a hipótese de que tal transformação tenha acontecido de forma gradual, foi selecionado o livro “O escaravelho do diabo”, escrito por Lúcia Machado de Almeida e publicado em 1956. É considerado um nível de significância 1% nos testes realizados.

### 5.1 Verificação de normalidade multivariada

Com a finalidade de aplicar o teste Box’s M para testar igualdade das matrizes de covariância, é necessário garantir que os dados possuam distribuição normal multivariada, já que este teste é sensível à suposição de normalidade multivariada. Apesar de não garantir, uma situação em que todas as variáveis exibem normalidade univariada ajuda a obter a normalidade multivariada; por isso, primeiro é testado se todas as variáveis possuem distribuição normal univariada, caso a normalidade univariada não ocorra em alguma variável, é aplicada uma transformação nesta variável com a finalidade de aproximá-la da normalidade e então testar-se a normalidade multivariada.

#### 5.1.1 Como treinar seu dragão

O primeiro livro, Como treinar seu dragão, possui 146 laudas, gerando, portanto, 146 valores de quantidade de palavras diferentes e 146 valores de tamanho médio de frases. Na Figura 5, é possível observar o histograma e o Q-Q plot da variável quantidade de palavras diferentes do livro “Como treinar seu dragão”.



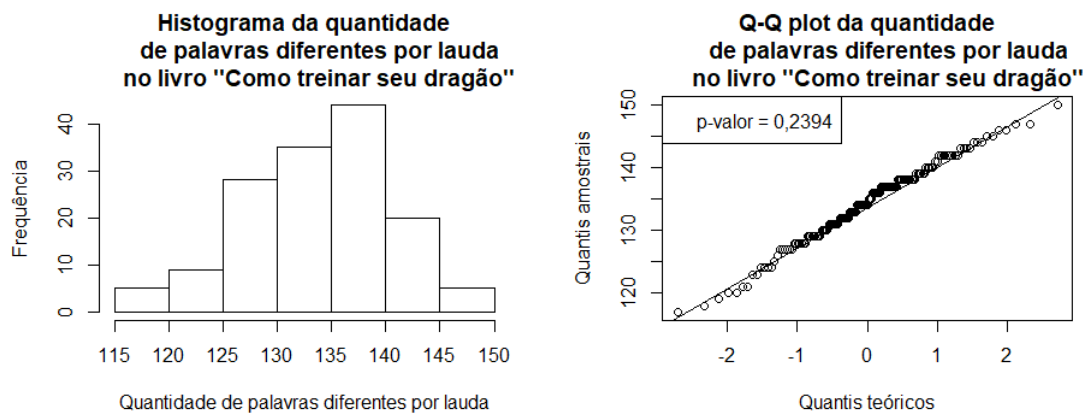


Figura 5: Histograma e Q-Q plot da quantidade de palavras diferentes por lauda do livro “Como treinar seu dragão”.

Através da análise do histograma e do Q-Q plot da quantidade de palavras diferentes por lauda, pode-se observar que os dados sugerem que esta variável segue uma distribuição normal univariada. Tal suposição é reafirmada através da realização do teste de normalidade univariada Shapiro-Wilk, que gera um p-valor igual a 0,239.

Neste livro, o tamanho médio de frases por lauda aparenta não seguir distribuição normal, como pode ser observado na Figura 6. Como o histograma apresenta uma cauda direita longa, as observações também são transformadas usando logaritmo para que a quantidade de valores grandes aumente, aumentando a simetria em relação à média. A Figura 7 ilustra o histograma dos dados transformados e sugere que a variável segue uma distribuição normal univariada, o que é confirmado com o teste Shapiro-Wilk, que gera um p-valor igual a 0,217. Dessa forma, nas análises são considerados os dados transformados ao invés dos dados originais.

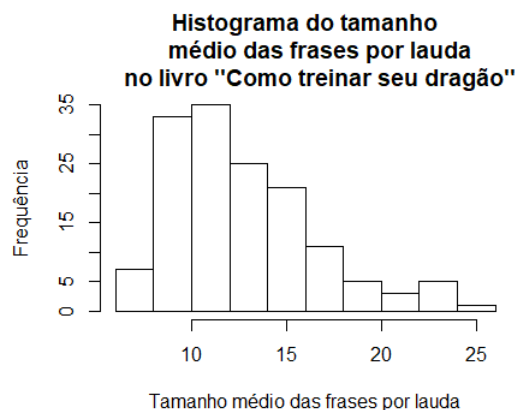


Figura 6: Histograma do tamanho médio de frases por lauda do livro “Como treinar seu dragão”.

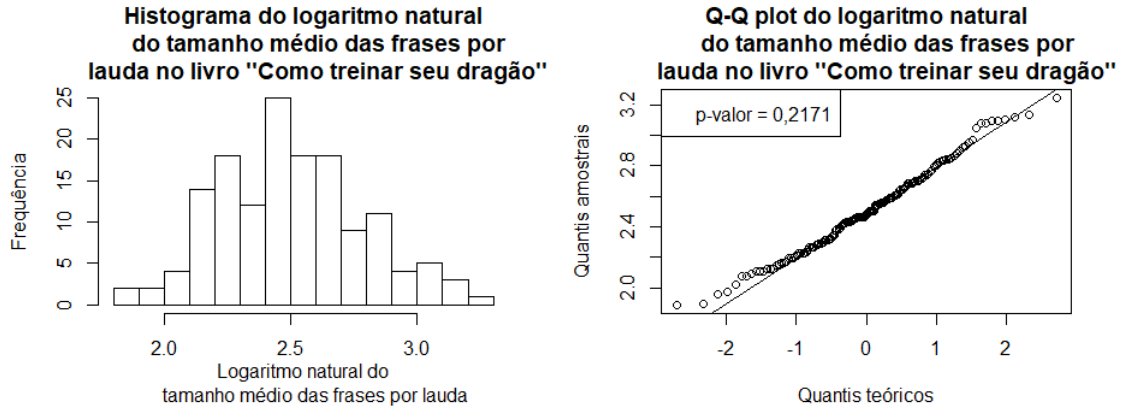


Figura 7: Histograma e Q-Q plot do logaritmo natural do tamanho médio das frases por lauda do livro “Como treinar seu dragão”.

As duas variáveis consideradas no livro “Como treinar seu dragão”, quantidade de palavras diferentes por lauda e logaritmo natural do tamanho médio das frases por lauda, serão denotadas por  $X_1$  e  $X_2$ , respectivamente, e formam uma variável aleatória multivariada denotada por  $\mathbf{X}^\top = [X_1, X_2]$ . Assim, a matriz de dados  $\mathbf{X}_1$  será da forma:

$$\mathbf{X}_{1(146 \times 2)} = \begin{bmatrix} \mathbf{x}_{11}^\top \\ \mathbf{x}_{12}^\top \\ \vdots \\ \mathbf{x}_{1146}^\top \end{bmatrix}$$

Observando a Figura 8, pode ser afirmado que o Q-Q plot sugere que a variável aleatória multivariada segue uma distribuição normal multivariada, tal hipótese é testada através do teste Shapiro-Wilk multivariado, que gera um p-valor igual a 0,0714. Logo, é razoável afirmar que a variável aleatória multivariada possui distribuição normal multivariada.

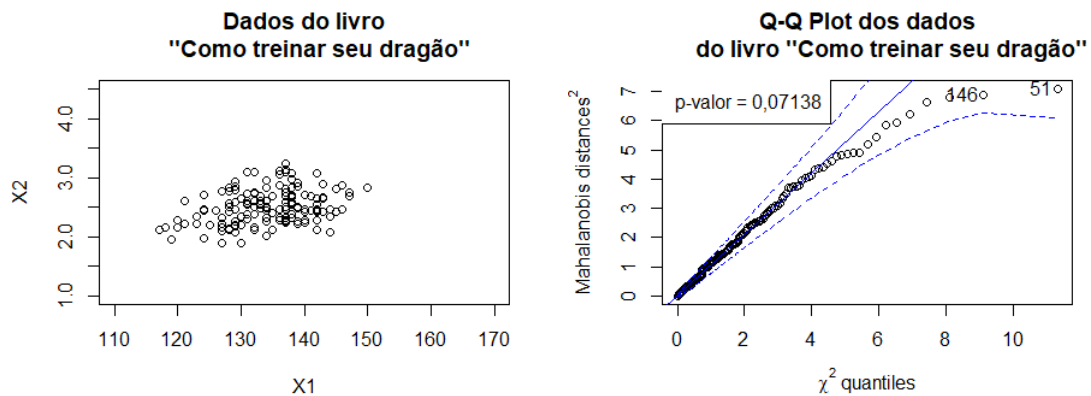


Figura 8: Gráfico de dispersão e Q-Q plot dos dados do livro “Como treinar seu dragão”.

### 5.1.2 A ilha do tesouro

Já o segundo livro, A ilha do tesouro, possui 294 laudas, gerando, portanto, 294 valores de quantidade de palavras diferentes e 294 valores de tamanho médio de frases. Na Figura 9, é possível observar o histograma e o Q-Q plot da variável quantidade de palavras diferentes do livro “A ilha do tesouro”.

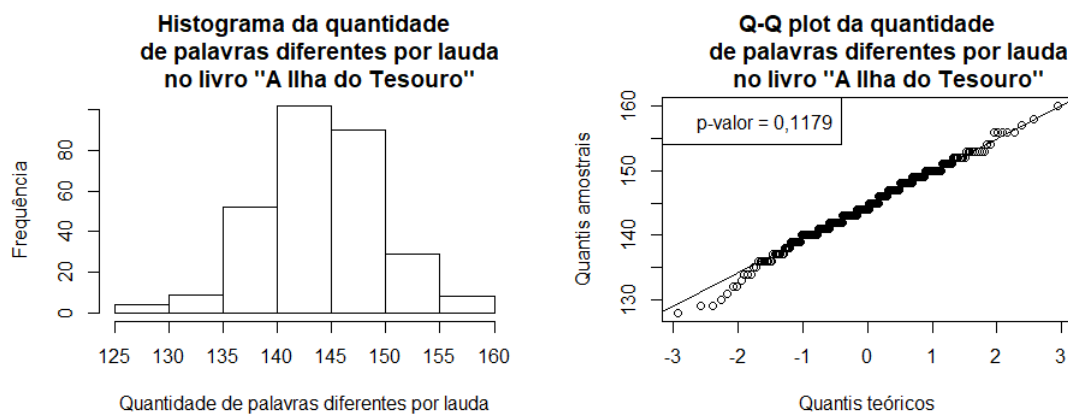


Figura 9: Histograma e Q-Q plot da quantidade de palavras diferentes por lauda do livro “A ilha do tesouro”.

Analisando o histograma e o Q-Q plot da quantidade de palavras diferentes por lauda, é possível observar que os dados sugerem que esta variável segue uma distribuição normal univariada. Tal suposição é reafirmada através da realização do teste de normalidade univariada Shapiro-Wilk, que gera um p-valor igual a 0,118.

O tamanho médio de frases por lauda aparenta não seguir distribuição normal, como pode ser observado na Figura 10. Como o histograma apresenta uma cauda direita longa,

as observações são transformadas usando logaritmo para que a quantidade de valores grandes aumente, aumentando a simetria em relação à média. A Figura 11 ilustra o histograma dos dados transformados e sugere que a variável segue uma distribuição normal univariada, o que é confirmado com o teste Shapiro-Wilk, que gera um p-valor igual a 0,099. Dessa forma, nos cálculos são considerados os dados transformados ao invés dos dados originais.

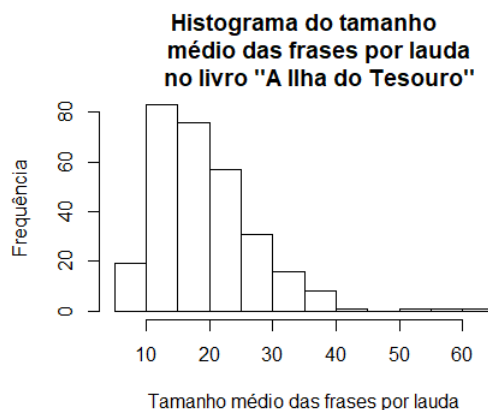


Figura 10: Histograma do tamanho médio de frases por lauda do livro “A ilha do tesouro”.

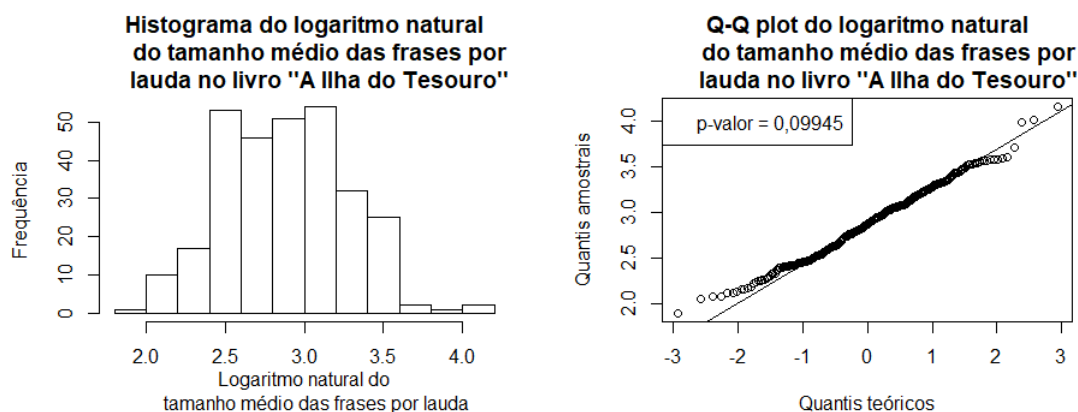


Figura 11: Histograma e Q-Q plot do logaritmo natural do tamanho médio das frases por lauda do livro “A ilha do tesouro”.

As variáveis consideradas no livro “A ilha do tesouro” são as mesmas que as consideradas no livro “Como treinar seu dragão” e serão denotadas da mesma forma. Assim, a matriz de dados  $\mathbf{X}_2$  será da forma:

$$\mathbf{X}_{2(294 \times 2)} = \begin{bmatrix} \mathbf{x}_{11}^\top \\ \mathbf{x}_{12}^\top \\ \vdots \\ \mathbf{x}_{1294}^\top \end{bmatrix}$$

Observando a Figura 12, pode ser afirmado que o Q-Q plot sugere que a variável aleatória multivariada segue uma distribuição normal multivariada, tal hipótese é testada através do teste Shapiro-Wilk multivariado, que gera um p-valor igual a 0,031. Logo, é razoável afirmar que a variável aleatória multivariada possui distribuição normal multivariada.

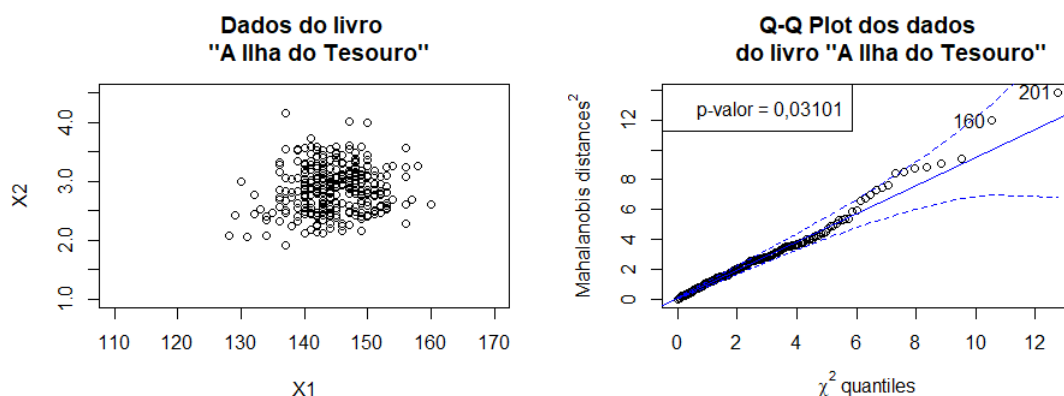


Figura 12: Gráfico de dispersão e Q-Q plot dos dados do livro “A ilha do tesouro”.

## 5.2 Teste de igualdade de matrizes de covariância

Posteriormente, deve-se verificar se as matrizes de covariâncias dos dados de cada livro são iguais ou diferentes. A partir da Figura 13 observa-se que os dados do primeiro gráfico se apresentam mais concentrados. As matrizes de covariância amostrais dos dados do primeiro e do segundo livro são, respectivamente,  $\mathbf{S}_1 = \begin{bmatrix} 43,786 & 0,621 \\ 0,621 & 0,081 \end{bmatrix}$  e  $\mathbf{S}_2 = \begin{bmatrix} 29,772 & 0,261 \\ 0,261 & 0,155 \end{bmatrix}$ . O teste de Box's M gerou um p-valor igual a  $4,51 \times 10^{-7}$ ; portanto, é sensato admitir que as matrizes de covariância são diferentes.

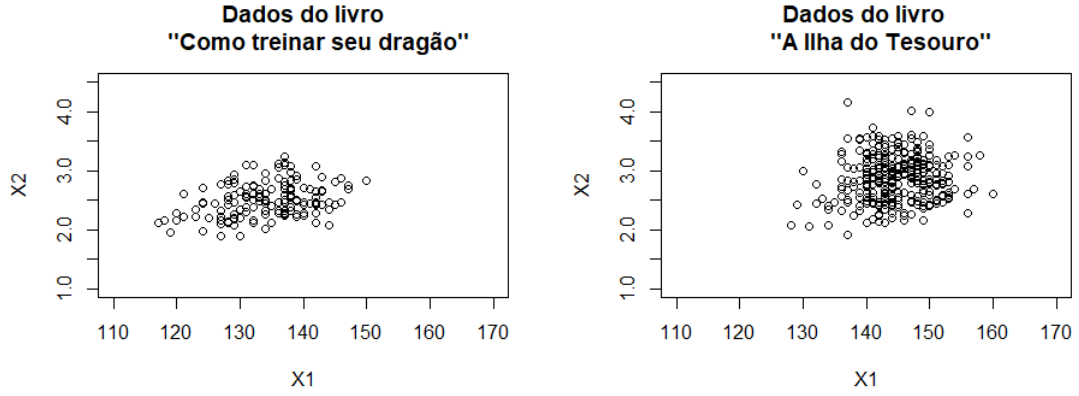


Figura 13: Gráfico de dispersão dos dados dos livros “Como treinar seu dragão” e “A ilha do tesouro”.

### 5.3 Regra de Classificação

Sejam  $f_1(\mathbf{x})$  e  $f_2(\mathbf{x})$  as funções densidade de probabilidade associadas à variável aleatória multivariada  $\mathbf{X}$  para os livros “Como treinar seu dragão” e “A ilha do tesouro”, respectivamente. Assim,  $f_1(\mathbf{x})$  e  $f_2(\mathbf{x})$  possuem distribuição normal multivariada com parâmetros  $\boldsymbol{\mu}_i$  e  $\boldsymbol{\Sigma}_i, i = 1, 2$ , sendo  $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ . Como os parâmetros são desconhecidos, para a obtenção de uma regra de classificação estimada, substitui-se os parâmetros  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1$  e  $\boldsymbol{\Sigma}_2$  pelas estimativas  $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_1$  e  $\mathbf{S}_2$ , respectivamente.

As estimativas encontradas são:

- $\bar{\mathbf{x}}_1^\top = \begin{bmatrix} 134,219 & 2,505 \end{bmatrix}, \bar{\mathbf{x}}_2^\top = \begin{bmatrix} 144,537 & 2,878 \end{bmatrix};$
- $\mathbf{S}_1 = \begin{bmatrix} 43,786 & 0,621 \\ 0,621 & 0,081 \end{bmatrix}, \mathbf{S}_2 = \begin{bmatrix} 29,772 & 0,261 \\ 0,261 & 0,155 \end{bmatrix}.$

Para encontrar a regra de classificação 3.12 para este exemplo, calcula-se:

- $\bar{\mathbf{x}}_1 = (\bar{\mathbf{x}}_1^\top)^\top = \begin{bmatrix} 134,219 \\ 2,505 \end{bmatrix}, \bar{\mathbf{x}}_2 = (\bar{\mathbf{x}}_2^\top)^\top = \begin{bmatrix} 144,537 \\ 2,878 \end{bmatrix};$
- $\mathbf{S}_1^{-1} = \begin{bmatrix} 0,026 & -0,195 \\ -0,195 & 13,778 \end{bmatrix}, \mathbf{S}_2^{-1} = \begin{bmatrix} 0,034 & -0,058 \\ -0,058 & 6,563 \end{bmatrix}.$

Assim, obtém-se

$$\begin{aligned}
\hat{\delta} &= \frac{1}{2} \ln \left( \frac{|\mathbf{S}_1|}{|\mathbf{S}_2|} \right) + \frac{1}{2} (\bar{\mathbf{x}}_1^\top \mathbf{S}_1^{-1} \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^\top \mathbf{S}_2^{-1} \bar{\mathbf{x}}_2) \\
&= \frac{1}{2} \ln \left( \frac{\begin{vmatrix} 43,786 & 0,621 \\ 0,621 & 0,081 \end{vmatrix}}{\begin{vmatrix} 29,772 & 0,261 \\ 0,261 & 0,155 \end{vmatrix}} \right) \\
&\quad + \frac{1}{2} \left( \begin{bmatrix} 134,219 & 2,505 \end{bmatrix} \begin{bmatrix} 0,026 & -0,195 \\ -0,195 & 13,778 \end{bmatrix} \begin{bmatrix} 134,219 \\ 2,505 \end{bmatrix} \right. \\
&\quad \left. - \begin{bmatrix} 144,537 & 2,878 \end{bmatrix} \begin{bmatrix} 0,034 & -0,058 \\ -0,058 & 6,563 \end{bmatrix} \begin{bmatrix} 144,537 \\ 2,878 \end{bmatrix} \right) \\
&= \frac{1}{2} \ln \left( \frac{3,178}{4,536} \right) + \frac{1}{2} (416,424 - 718,684) \\
&= -0,178 + -151,130 \\
&= -151,308
\end{aligned}$$

e, com isso,

$$\begin{aligned}
& -\frac{1}{2} \mathbf{x}^\top (\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1}) \mathbf{x} + (\bar{\mathbf{x}}_1^\top \mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2^\top \mathbf{S}_2^{-1}) \mathbf{x} - \hat{\delta} \\
&= -\frac{1}{2} \mathbf{x}^\top \left( \begin{bmatrix} 0,026 & -0,195 \\ -0,195 & 13,778 \end{bmatrix} - \begin{bmatrix} 0,034 & -0,058 \\ -0,058 & 6,563 \end{bmatrix} \right) \mathbf{x} \\
&\quad + \left( \begin{bmatrix} 134,219 & 2,505 \end{bmatrix} \begin{bmatrix} 0,026 & -0,195 \\ -0,195 & 13,778 \end{bmatrix} \right. \\
&\quad \left. - \begin{bmatrix} 144,537 & 2,878 \end{bmatrix} \begin{bmatrix} 0,034 & -0,058 \\ -0,058 & 6,563 \end{bmatrix} \right) \mathbf{x} + 151,308 \\
&= -\frac{1}{2} \mathbf{x}^\top \begin{bmatrix} -0,008 & -0,138 \\ -0,138 & 7,215 \end{bmatrix} \mathbf{x} + \begin{bmatrix} -1,814 & -2,257 \end{bmatrix} \mathbf{x} + 151,308
\end{aligned}$$

Portanto,  $\mathbf{x}$  deve ser alocado na população  $\pi_1$  se

$$\begin{aligned}
& -\frac{1}{2} \mathbf{x}^\top \begin{bmatrix} -0,008 & -0,138 \\ -0,138 & 7,215 \end{bmatrix} \mathbf{x} + \begin{bmatrix} -1,814 & -2,257 \end{bmatrix} \mathbf{x} + 151,308 \\
&\geq \ln \left[ \left( \frac{c(1|2)}{c(2|1)} \right) \left( \frac{p_2}{p_1} \right) \right]
\end{aligned}$$

e em  $\pi_2$  caso contrário.

Apesar do tamanho das amostras serem diferentes, não existem evidências para considerar que o tamanho da população de livros infantojuvenis de 1883 seja diferente do tamanho da população de livros infantojuvenis de 2010, sendo considerados, portanto, custos iguais e prioris iguais. Seja  $\mathbf{x}^\top = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$  uma observação a ser classificada como pertencente a  $\pi_1$  ou  $\pi_2$ , a regra de classificação pode ser representada pela curva correspondente à equação

$$\begin{aligned}
 & -\frac{1}{2}\mathbf{x}^\top \begin{bmatrix} -0,008 & -0,138 \\ -0,138 & 7,215 \end{bmatrix} \mathbf{x} + \begin{bmatrix} -1,814 & -2,257 \end{bmatrix} \mathbf{x} + 151,308 = 0 \\
 \Leftrightarrow & -\frac{1}{2} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} -0,008 & -0,138 \\ -0,138 & 7,215 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
 & + \begin{bmatrix} -1,814 & -2,257 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 151,308 = 0 \\
 \Leftrightarrow & -\frac{1}{2}(-0,008 x_1^2 - 0,275 x_1 x_2 + 7,215 x_2^2) \\
 & -1,814 x_1 - 2,257 x_2 + 151,308 = 0 \\
 \Leftrightarrow & 0,004 x_1^2 - 3,607 x_2^2 + 0,138 x_1 x_2 \\
 & -1,814 x_1 - 2,257 x_2 + 151,308 = 0
 \end{aligned}$$

A Figura 14 ilustra a hipérbole que representa a regra de classificação, observações pertencentes às regiões indicadas como  $R_1$  e  $R_2$  são classificadas como pertencentes à população de livros infantojuvenis antigos e atuais, respectivamente.

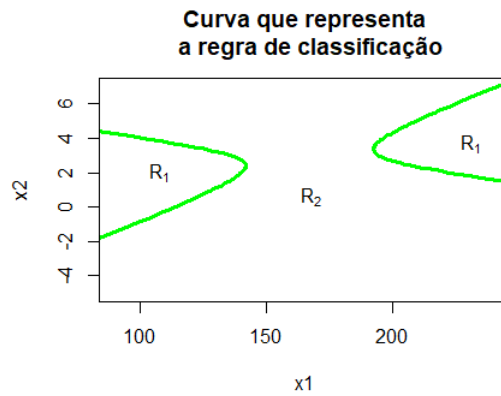


Figura 14: Curva que representa a regra de classificação



Considerando os dois livros selecionados, é possível observar na Figura 15 o gráfico de dispersão dos dados de cada livro, as elipses que cobrem 95% dos valores de cada amostra e a regra de classificação em verde.

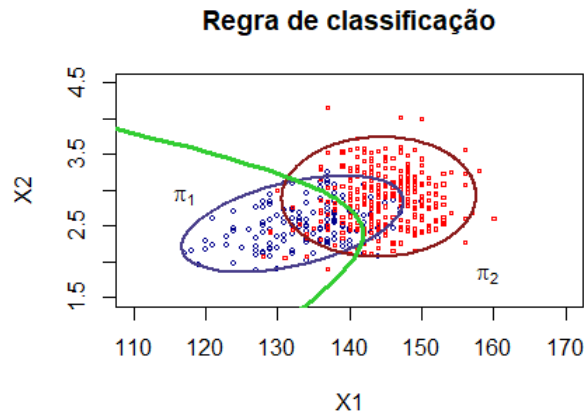


Figura 15: Regra de classificação

## 5.4 Avaliação da regra de classificação

Como as matrizes de covariância são diferentes, a regra de classificação é avaliada através dos estimadores de  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$ ,  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  e da PTCI calculados a partir dos métodos Ressubstituição (R), Ressubstituição com Divisão Amostral (RDA) e Pseudo-*jackknife*.

### 5.4.1 Ressubstituição (R)

A partir da regra de classificação estimada anteriormente, cada observação amostral é classificada em uma das duas populações e as classificações podem estar corretas ou incorretas. A Tabela 17 apresentada a seguir resume os resultados em uma tabela de contingência.

Tabela 17: Tabela de contingência			
População real	População classificada		Total
	$\pi_1$	$\pi_2$	
$\pi_1$	113	33	146
$\pi_2$	44	250	294
			440

O estimador da *PTCI*, denominado de *TEA*, é dado por:

$$TEA = \frac{33 + 44}{146 + 294} = \frac{77}{440} = 0,175.$$

As probabilidades  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  são estimadas por

$$\hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \frac{33}{146} = 0,226 \quad \text{e} \quad \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \frac{44}{194} = 0,227.$$

### 5.4.2 Ressubstituição com Divisão Amostral (RDA)

Dividindo a amostra em duas partes, a amostra de treinamento é formada pela metade dos dados de cada população e amostra de validação pela outra metade. A primeira é utilizada na estimação da função discriminante e especificação da regra estimada de classificação, enquanto na segunda as observações são classificadas de acordo com a regra estimada na amostra de treinamento e a partir do resultado final são estimadas as taxas de erro. A Figura 16 ilustra o gráfico de dispersão da amostra de treinamento e a regra de classificação gerada através desta.

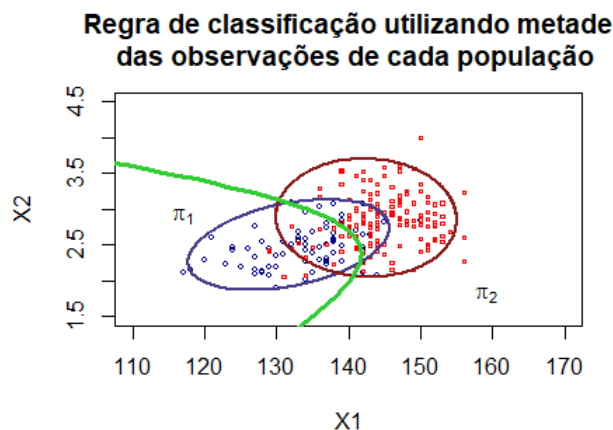


Figura 16: Regra de classificação utilizando metade dos dados de cada população

A partir desta regra de classificação, os dados da amostra de validação são classificados como pertencentes a  $\pi_1$  ou a  $\pi_2$  e as classificações podem estar corretas ou incorretas. A Tabela 18 resume os resultados em uma tabela de contingência.

O estimador da *PTCI*, denominado de *TEA*, é dado por:

$$TEA = \frac{17 + 20}{73 + 147} = \frac{37}{220} = 0,168.$$

Tabela 18: Tabela de contingência			
População real	População classificada		Total
	$\pi_1$	$\pi_2$	
$\pi_1$	56	17	73
$\pi_2$	20	127	147
			220

As probabilidades  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  são estimadas por

$$\hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) = \frac{17}{73} = 0,233 \quad \text{e} \quad \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) = \frac{20}{147} = 0,136.$$

### 5.4.3 Pseudo-*jackknife* (JACK)

Omitindo uma observação por vez das 440 observações originais na amostra de treinamento, as 439 observações remanescentes são utilizadas para estimar as regras de classificação. Cada observação omitida é classificada em uma das populações e é possível determinar se tal observação foi classificada correta ou incorretamente, já que sabe-se de qual população ela foi originalmente amostrada.

As probabilidades  $P(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1)$  e  $P(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2)$  são estimadas por

$$\begin{aligned} \hat{P}(\mathbf{X} \in R_2 | \mathbf{X} \in \pi_1) &= \frac{n_{12}^{(j)}}{n_1} = \frac{33}{146} = 0,226 \text{ e} \\ \hat{P}(\mathbf{X} \in R_1 | \mathbf{X} \in \pi_2) &= \frac{n_{21}^{(j)}}{n_2} = \frac{44}{294} = 0,150, \end{aligned}$$

em que o superescrito  $(j)$  denota o procedimento *Jackknife*.

A *TEA* estimada pelo método é

$$TEA^{(j)} = \frac{n_{12}^{(j)} + n_{21}^{(j)}}{n_1 + n_2} = \frac{33 + 44}{146 + 294} = 0,175$$

A Figura 17 ilustra o gráfico de dispersão dos dados de cada livro, as elipses que cobrem 95% dos valores de cada amostra, a regra de classificação em verde e as estimativas da PTCI calculadas por cada método acima do gráfico. Como as probabilidades de erros são pequenas, diz-se que as variáveis utilizadas discriminam bem as duas populações.

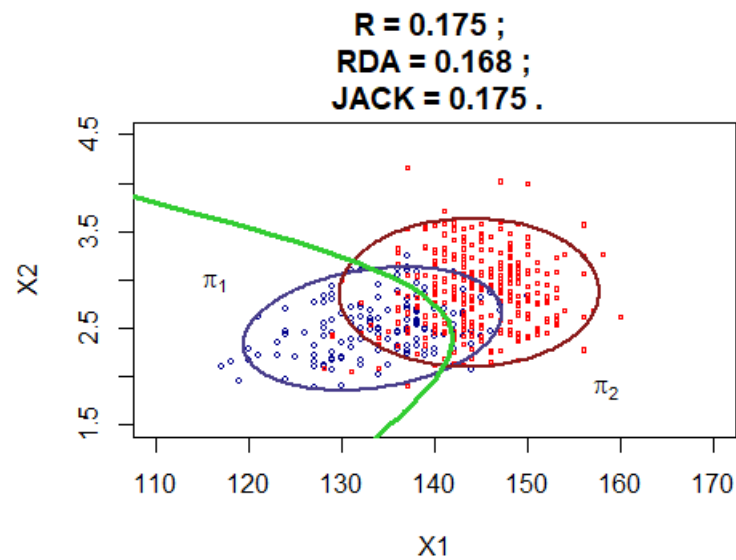


Figura 17: Regra de classificação

## 5.5 Classificação do livro “O escaravelho do diabo”

Utilizando a regra de classificação construída para classificar as observações de um terceiro livro, O escaravelho do diabo, escrito por Lúcia Machado de Almeida e publicado em 1956, obtém-se que 55,8% das observações são classificadas como pertencentes a  $\pi_1$  e 44,2% como pertencentes a  $\pi_2$ , seu gráfico de dispersão é observado em preto na Figura 18.

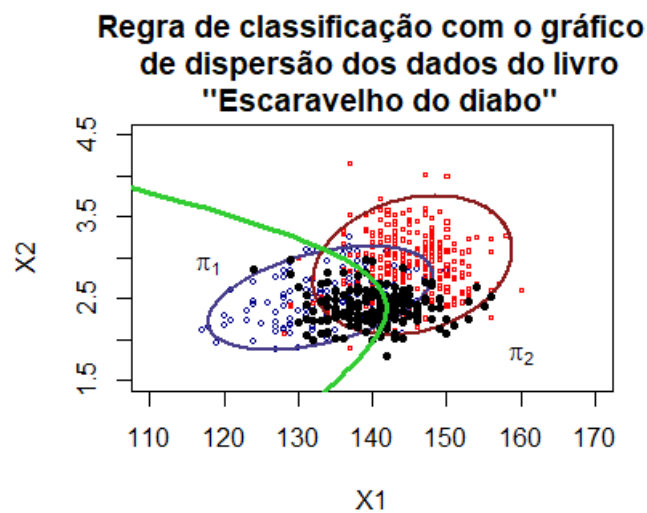


Figura 18: Regra de classificação

## 6 Considerações finais e trabalhos futuros

O presente trabalho teve como objetivo apresentar a análise discriminante, seu método de estimação e a avaliação da regra de classificação estimada. Além disso, os métodos de estimação e avaliação da análise discriminante foram implementadas no software R. Com o intuito de verificar se as estimativas calculadas são eficientes, foram realizadas simulações para o caso de duas variáveis e duas populações multivariadas; considerando que as matrizes de covariância são iguais e não há correlação entre as variáveis, simula-se também 12 cenários variando a distância entre as médias e as variâncias das variáveis consideradas; considerando o caso em que as matrizes de covariância são iguais e há correlação entre as variáveis, realizou-se 48 simulações variando a distância entre as médias e as variâncias das variáveis consideradas; considerando matrizes de covariância diferentes e não existência de correlação entre as variáveis, foram feitas 48 simulações variando a distância entre as médias e as matrizes de covariância de cada população; e considerando que as matrizes de covariância são diferentes e existe correlação entre as variáveis, foram realizadas 48 simulações variando a distância entre as médias e as matrizes de covariância de cada população. Para cada simulação, as regras de classificação foram estimadas e podem ser observadas nos respectivos gráficos e, para o caso em que as matrizes de covariância são iguais, os métodos de ressubstituição, ressubstituição por divisão amostral, *pseudo-jackknife*, probabilidade de classificação incorreta estimada e método dois de Lachenbruch e Mickey foram utilizados para estimar a probabilidade total de classificação incorreta (PTCI); já no caso em que as matrizes de covariância são diferentes a PTCI foi estimada apenas através dos métodos de ressubstituição, ressubstituição por divisão amostral, *pseudo-jackknife*. É possível observar que no geral as estimativas da PTCI são maiores nos casos em que as médias são próximas ou quando os dados apresentam alta variabilidade; entretanto, a posição das populações no gráfico de dispersão também influencia no quanto as populações se confundem.

A análise discriminante foi aplicada neste trabalho com o objetivo de criar uma regra de classificação que discrimine bem a população de livros do século XIX de livros atuais; para tal objetivo, foram selecionados os livros “Como treinar seu dragão”, representando uma amostra de livro infantojuvenil antigo, e “A ilha do tesouro”, representando uma amostra de livro infantojuvenil atual. Com o intuito de comparar a população de livros infantojuvenis atuais,  $\pi_1$ , e antigos,  $\pi_2$ , para cada um dos livros selecionados foram calculadas a quantidade de palavras diferentes por lauda e o tamanho médio das frases por lauda, sendo considerada como uma lauda 1200 caracteres.

Considera-se que os livros infantojuvenis do século XIX possuem características diferentes dos livros infantojuvenis atuais, já que a preocupação com a produção de livros infantojuvenis atualmente é maior. Na Figura 17, observa-se que, conforme esperado, o livro de 1883 apresentou em média uma quantidade maior de palavras diferentes por lauda e um tamanho médio de frases por lauda maior. A regra de classificação criada discriminar bem os dois livros sugere que ocorreu uma simplificação nos livros infantojuvenis do século XIX para os dias atuais.

O livro “O escaravelho do diabo” foi lançado depois do livro de 1883 e antes do de 2010, os valores médios da quantidade de palavra diferente por lauda e do tamanho médio de frases por lauda se apresentam distribuídos entre os valores dos outros dois livros, tal resultado sugere que a simplificação dos livros infantojuvenis ocorreu de forma gradual.

Em trabalhos futuros, seria recomendável que fossem utilizados livros diferentes, tanto no caso dos livros utilizados na análise discriminante quanto no caso do terceiro livro selecionado para confirmar a hipótese de que ocorreu uma simplificação gradual, com o intuito de confirmar as conclusões obtidas neste trabalho.

# Referências

- [1] ZILBERMAN, R. *A literatura infantil na escola*. [S.l.]: Global Editora e Distribuidora Ltda, 2015.
- [2] ASSMANN, S. O mundo da fantasia, a subjetividade e o atraso mental. 2012.
- [3] RECHOU, B.-A. R.; LÓPEZ, I. S.; RODRÍGUEZ, M. N. *A poesía infantil no século XXI (2000-2008)*. [S.l.]: Xerais de Galicia,S.A., 2009.
- [4] SCHEFFER, C. S. A literatura no contexto da educação infantil. 2010.
- [5] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <<https://www.R-project.org/>>.
- [6] MAROCO, J. *Análise estatística - Com utilização do SPSS*. 3. ed. [S.l.]: Sílabo, 2007.
- [7] MUYLDER, C. F. D. et al. Principais aplicações de análise discriminante na área de marketing: uma pesquisa bibliométrica. *Revista Gestão & Tecnologia*, v. 12, n. 2, p. 217–242, 2012.
- [8] CRASK, M. R.; JR, W. D. P. Validation of discriminant analysis in marketing research. *Journal of Marketing Research*, JSTOR, p. 60–68, 1977.
- [9] JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. 6. ed. [S.l.]: Pearson Prentice Hall, 2007.
- [10] JR WILLIAM C. BLACK, B. J. B. J. F. H.; ANDERSON, R. E. *Multivariate data analysis*. 7. ed. [S.l.]: Pearson, 2010.
- [11] WALD, A. On a statistical problem arising in the classification of an individual into one of two groups. v. 15, n. 2, p. 145–162, jun. 1944. ISSN 0003-4851. Disponível em: <<http://projecteuclid.org/euclid.aoms/1177731280>>.
- [12] LACHENBRUCH, P. A.; MICKEY, M. R. Estimation of error rates in discriminant analysis. *Technometrics*, Taylor & Francis, v. 10, n. 1, p. 1–11, 1968.