

UNIVERSIDADE FEDERAL DE VIÇOSA

WAGNER FARIA BARBOSA

ANÁLISE DE VARIÁVEIS CANÔNICAS NO SOFTWARE R

**VIÇOSA – MINAS GERAIS
2020**

WAGNER FARIA BARBOSA

ANÁLISE DE VARIÁVEIS CANÔNICAS NO SOFTWARE R

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

Orientador: José Ivo Ribeiro Júnior

VIÇOSA – MINAS GERAIS
2020

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade
Federal de Viçosa - Campus Viçosa**

T

B238a
2020
Barbosa, Wagner Faria, 1985-
Análise de variáveis canônicas no software R / Wagner
Faria Barbosa. – Viçosa, MG, 2020.
73 f. : il. (algumas color.) ; 29 cm.

Orientador: José Ivo Ribeiro Júnior.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 72-73.

1. Análise multivariada. 2. Delineamento experimental.
3. Análise de variância. 4. Autovalores. 5. Funções algébricas.
I. Universidade Federal de Viçosa. Departamento de Estatística.
Programa de Pós-Graduação em Estatística Aplicada e
Biometria. II. Título.

CDD 22. ed. 519.535

WAGNER FARIA BARBOSA

ANÁLISE DE VARIÁVEIS CANÔNICAS NO SOFTWARE R

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

APROVADA: 26 de outubro de 2020.

Assentimento:



Wagner Faria Barbosa
Autor



José Ivo Ribeiro Júnior
Orientador

*A Deus seja dada
toda honra e
glória para
sempre.*

AGRADECIMENTOS

Agradeço a Deus pela sua revelação especial, pela sua Palavra que é viva e transformadora, pelo seu amor, pelo dom gracioso da vida em Cristo Jesus, pela mensagem do Evangelho que chegou a minha alma, pelas suas misericórdias que se renovam sobre minha vida todos os dias, pelas portas abertas e portas fechadas, pelas oportunidades que incluem este mestrado, pelas dores e alegrias, pela minha família e amigos, pela consolação em tempos de angústia, pelas repreensões, pela minha saúde, pela força e pela fraqueza. Agradeço a Deus porque todas as coisas cooperam para o meu bem, para que eu conheça cada dia sua boa, perfeita e agradável vontade.

Agradeço aos meus pais, pelo amor, pelo carinho, pela dedicação e pelo incentivo aos meus estudos.

Agradeço aos meus novos amigos que fiz durante o curso e que partilharam comigo os desafios, mas também as alegrias da Pós-Graduação.

Agradeço a Universidade Federal de Viçosa pela oportunidade concedida neste curso.

Agradeço aos professores do curso de Pós-Graduação em Estatística Aplicada e Biometria pelo ensino e dedicação.

Agradeço ao secretário Júnior José Pires pela prontidão e zelo em auxiliar nas diversas atividades burocráticas do Programa de Pós-Graduação.

Agradeço aos membros da banca, Profa. Camila e Prof. Rodrigo, pelas correções e sugestões.

Agradeço ao meu orientador Prof. José Ivo, pela orientação, apoio, conselhos e confiança.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

BARBOSA, Wagner Faria, M.Sc., Universidade Federal de Viçosa, outubro de 2020. **Análise de variáveis canônicas no software R**. Orientador: José Ivo Ribeiro Júnior.

As variáveis canônicas (*VCs*) são combinações lineares de variáveis quantitativas utilizadas na discriminação de grupos. No contexto de experimentos executados com tratamentos repetidos sob diferentes delineamentos de casualização, a obtenção das *VCs* pressupõe a realização da MANOVA, uma vez que elas são obtidas pela maximização da relação entre a variação entre e dentro dos tratamentos utilizando, respectivamente, as matrizes de soma de quadrados e produtos do tratamento (**H**) e do resíduo (**E**). Funções discriminantes lineares de Fisher (*FDS*), por outro lado, são combinações lineares de variáveis cujo principal objetivo é a classificação de observações em populações distintamente estabelecidas e conhecidas *a priori*. Uma vez que existe uma equivalência matemática para obtenção das *FDS* e das *VCs*, a obtenção das *VCs* pode ser facilitada pela obtenção de *FDS*, que é um método mais comumente encontrado em softwares estatísticos. Neste sentido, o objetivo deste trabalho foi de analisar, por meio da simulação de dados, a equivalência entre as *FDS* e as *VCs*, a partir de experimentos instalados sob o DIC, DBC e DQL. A obtenção das *VCs* por meio das *FDS* foi realizada utilizando a função *lda* implementada no software R com o uso dos valores transformados das observações de cada variável-resposta pela subtração dos efeitos perturbadores (ou seja, de bloco para o DBC, ou de linha e coluna para o DQL) quando eles existiam no modelo estatístico. Além disso, duas padronizações foram utilizadas na obtenção das *VCs*. Os resultados proporcionados por ambos os métodos foram comparados por diferentes medidas que incluíram, autovetores, autovalores, importância relativa, além de correlações e da análise de variância e inspeção visual de gráficos bidimensionais, e comprovaram a similaridade das *VCs* com as *FDS*, uma vez que não houve prejuízo nem na interpretabilidade nem nas análises realizadas com ambas as funções lineares. Conclui-se, portanto, que as *VCs* podem ser eficientemente obtidas por meio das *FDS* com o uso da função *lda* e ambos os métodos são equivalentes.

Palavras-chave: Análise multivariada. Delineamentos experimentais. Análise discriminante.

ABSTRACT

BARBOSA, Wagner Faria, M.Sc., Universidade Federal de Viçosa, October, 2020. **Analysis of canonical variates in software R**. Adviser: José Ivo Ribeiro Júnior.

Canonical variables (*VCs*) are linear combinations of quantitative variables used to discriminate groups. In the context of experiments performed with repeated treatments under different randomization designs, obtaining the *VCs* presupposes the realization of MANOVA, since they are obtained by maximizing the relationship between the variance among and within the treatments using, respectively, the matrices of sum squares and sum products of treatments (*H*) and residues (*E*). Fisher's linear discriminating functions (*FDs*), on the other hand, are linear combinations of variables whose main objective is the classification of observations in distinctly established and *a priori* known populations. Since there is a mathematical equivalence for obtaining *FDs* and *VCs*, obtaining *VCs* can be facilitated by obtaining *FDs*, which is a method most found in statistical software. In this sense, the objective of this work was to analyze, through data simulation, the equivalence between *FDs* and *VCs*, based on experiments installed under DIC, DBC and DQL. The acquisition of *VCs* by means of *FDs* was performed using the *lda* function implemented in software R and using the transformed values of the observations of each response variable by subtracting the disturbing effects (that is, from block to DBC, or from row and column for DQL) when they existed in the statistical model. In addition, two normalizations were used for the eigenvectors. The results provided by both methods were compared by different measures that included, eigenvectors, eigenvalues, relative importance, in addition to correlations and analysis of variance and visual inspection of bi-dimensional graphs and proved the similarity of the *VCs* with the *FDs*. This is because there was no impairment in either interpretability or analysis performed with both linear functions. It is concluded, therefore, that *VCs* can be efficiently obtained by means of *FDs* using the *lda* function and both methods are similar.

Keywords: Multivariate analysis. Experimental designs. Discriminant analysis.

LISTA DE FIGURAS

Figura 5.1 - Média dos escores por tratamento em função da VC_1 e VC_2 pela padronização 1 ($g' Eg = 1$) (A) e 2 ($e' Se = 1$) (B) no DIC.....	54
Figura 5.2 - Média dos escores por tratamento em função da VC_1 e VC_2 pela padronização 1 ($g' Eg = 1$) (A) e 2 ($e' Se = 1$) (B) no DBC.....	58
Figura 5.3 - Média dos escores por tratamento em função da VC_1 e VC_2 pela padronização 1 ($g' Eg = 1$) (A) e 2 ($e' Se = 1$) (B) no DQL.....	63

LISTA DE TABELAS

Tabela 4.1 - Análise de variância multivariada sob o DIC	32
Tabela 4.2 - Variáveis canônicas padronizadas	34
Tabela 4.3 - Funções discriminantes lineares de Fisher padronizadas	37
Tabela 4.4 - Análise de variância multivariada sob o DBC.....	40
Tabela 4.5 - Análise de variância multivariada sob o DQL.....	47
Tabela 5.1 - Resultado parcial da análise de variância multivariada sob o DIC	52
Tabela 5.2 - Variáveis canônicas padronizadas obtidas posteriormente à MANOVA de um experimento instalado sob o DIC com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4	53
Tabela 5.3 - Correlações obtidas entre as duas primeiras variáveis canônicas (<i>VCs</i>) padronizadas e as variáveis-resposta originais Y_1 , Y_2 , Y_3 e Y_4 de um experimento instalado sob o DIC.....	53
Tabela 5.4 - Análise de variância (ANOVA) das variáveis canônicas padronizadas (<i>VCs</i>) obtidas posteriormente a MANOVA de um experimento instalado sob o DIC	53
Tabela 5.5 - Funções discriminantes (<i>FDS</i>) padronizadas obtidas de um experimento instalado sob o DIC.....	55
Tabela 5.6 - Resultado parcial da análise de variância multivariada (MANOVA) de um experimento hipoteticamente instalado sob o DBC com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4	56
Tabela 5.7 - Variáveis canônicas padronizadas obtidas posteriormente a MANOVA de um experimento instalado sob o DBC com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4	57
Tabela 5.8 - Correlação obtida entre as variáveis canônicas (<i>VCs</i>) padronizadas obtidas posteriormente a MANOVA e as variáveis-resposta de um experimento instalado sob o DBC	57
Tabela 5.9 - Análise de variância (ANOVA) das variáveis canônicas (<i>VCs</i>) padronizadas obtidas posteriormente a MANOVA de um experimento instalado sob o DBC	58
Tabela 5.10 - Funções discriminantes (<i>FDS</i>) padronizadas obtidas de um experimento instalado sob o DBC	59
Tabela 5.11 - Análise de variância (ANOVA) das funções discriminantes lineares (<i>FDS</i>) padronizadas obtidas com os escores dos dados originais de um experimento instalado sob o DBC.....	60

Tabela 5.12 - Resultado parcial da análise de variância multivariada (MANOVA) de um experimento hipoteticamente instalado sob o DQL com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4	61
Tabela 5.13 - Variáveis canônicas padronizadas (VCs) obtidas posteriormente a MANOVA de um experimento instalado sob o DQL com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4	61
Tabela 5.14 - Correlação obtida entre as variáveis canônicas (VCs) padronizadas obtidas posteriormente a MANOVA e as variáveis-resposta de um experimento instalado sob o DQL	62
Tabela 5.15 - Análise de variância (ANOVA) das variáveis canônicas (VCs) padronizadas obtidas posteriormente a MANOVA de um experimento instalado sob o DQL	62
Tabela 5.16 - Funções discriminantes (FDS) padronizadas obtidas de um experimento instalado sob o DQL	64
Tabela 5.17 - Análise de variância (ANOVA) das funções discriminantes lineares (FDS) padronizadas obtidas dos dados transformados de um experimento instalado sob o DQL	65

SUMÁRIO

1. INTRODUÇÃO GERAL	12
2. OBJETIVOS	13
2.1. Geral	13
2.2. Específicos	13
3. REVISÃO BIBLIOGRÁFICA	14
3.1. Variáveis Canônicas	14
3.2. Análise Discriminante	18
3.2.1. Duas Populações	18
3.2.2. Duas ou Mais de Duas Populações	22
3.3. Delineamentos Experimentais.....	26
3.4. Software Estatístico.....	27
4. MATERIAL E MÉTODOS.....	29
4.1. Simulação dos Dados	29
4.2. Delineamento Inteiramente Casualizado.....	31
4.2.1. Variáveis Canônicas	32
4.2.2. Funções Discriminantes	35
4.3. Delineamento em Blocos Casualizados	39
4.3.1. Variáveis Canônicas	40
4.3.2. Funções Discriminantes	42
4.4. Delineamento em Quadrado Latino	45
4.4.1. Variáveis Canônicas	46
4.4.2. Funções Discriminantes	48
4.5. Medidas de Avaliação	51
5. RESULTADOS	51
5.1. Simulação dos Dados	51
5.2. Delineamento Inteiramente Casualizado.....	52

5.2.1.	Variáveis Canônicas	52
5.2.2.	Funções Discriminantes	54
5.3.	Delineamento em Blocos Casualizados	56
5.3.1.	Variáveis Canônicas	56
5.3.2.	Funções Discriminantes	58
5.4.	Delineamento em Quadrado Latino	60
5.4.1.	Variáveis Canônicas	60
5.4.2.	Funções Discriminantes	63
6.	DISCUSSÃO	65
7.	CONCLUSÕES	71
8.	REFERÊNCIAS BIBLIOGRÁFICAS	72

1. INTRODUÇÃO GERAL

No contexto de experimentos com repetições realizados sob diferentes delineamentos experimentais, tais como: inteiramente casualizado (DIC), em blocos casualizados (DBC) e em quadrado latino (DQL), as variáveis canônicas (*VCs*) são funções latentes obtidas a partir da transformação linear de diferentes variáveis-resposta com o propósito de separação de tratamentos (RENCHEER, 2002). As *VCs* são obtidas através da maximização da relação das matrizes que compõe as somas de quadrados e de produtos das fontes de variação que representam, respectivamente, a variação entre e dentro dos tratamentos e são extraídas da decomposição da variância total através da análise de variância multivariada (MANOVA). Um dos principais benefícios desta técnica, é a redução da dimensionalidade dos experimentos multivariados para facilitar a comparação dos tratamentos com análises univariadas, além de possibilitar a inspeção visual dos resultados experimentais através de gráficos bi ou tridimensionais (RENCHEER, 2002).

Curiosamente, apesar de não explicitado, as *VCs* têm analogia com as funções discriminantes lineares de Fisher (*FDS*), que são funções matemáticas utilizadas para estabelecer regras de discriminação, a partir de um conjunto de informações ou de características (quantitativas), e classificar novos elementos amostrais (objetos) nas populações já existentes (McLACHLAN, 1992; FERREIRA, 2018). Basicamente, a análise discriminante é procedida de forma que, inicialmente, o perfil geral das diferentes populações pré-existentes seja identificado através de uma análise estatística do comportamento de *p*-variáveis aleatórias. Posteriormente, se houver algum novo objeto cuja origem é incerta, será possível compará-lo com o perfil de cada população e classificá-lo à população, cujo perfil seja o mais semelhante possível com aquele do novo objeto (MINGOTI, 2005).

O método para a análise discriminante proposto pelo estatístico e geneticista Ronald Fisher em 1936 é um dos mais proeminentes até os dias de hoje, embora algumas derivações dele já existam. As funções discriminantes estabelecidas por Fisher têm, como única pressuposição, que todas as populações sejam homocedásticas sem que haja, no entanto, a necessidade de que as variáveis sejam normalmente distribuídas (FISHER, 1936, 1938). Inicialmente, as *FDS* foram pensadas para apenas duas populações. Posteriormente, Fisher as estendeu para mais de duas. Resumidamente, o que Fisher fez foi criar uma transformação linear do vetor de observações multivariado (\mathbf{y}) em uma observação univariada ($fd = \mathbf{a}'\mathbf{y}$), de tal forma que o índice de separação entre as populações consideradas na análise fosse maximizado. Isso foi concretizado através da

maximização da razão das variâncias entre as populações e entre os objetos dentro de cada uma delas (FISHER, 1936, 1938).

Diante do exposto, é fácil inferir que, na ausência de outras fontes de variação experimentais além daquelas relativas às variações entre e dentro dos tratamentos, como é o caso de experimentos realizados sob o DIC, as *FDs* podem ser também obtidas para representarem as *VCs*, ou seja, as *FDs* tornam-se em si, as próprias *VCs*. Isso porque, a variação causada pelos tratamentos e a variação devida às causas aleatórias dentro deles em experimentos instalados sob o DIC funcionam, respectivamente, como as variações entre e dentro das populações para as *FDs* (MONTGOMERY, 2012). Conforme já atestado por Nørgaard et al. (2006) e Riba et al. (2020), a obtenção das *FDs* seguem uma dedução matemática semelhante à das *VCs*.

Apesar da semelhança entre os dois métodos multivariados, a literatura para as *FDs* é muito mais expressiva do que para as *VCs*, sendo também comum a incorporação das *FDs* em softwares estatísticos, o que não se observa, comumente, para as *VCs* (EVERITT, 2005).

2. OBJETIVOS

2.1. Geral

Estabelecer soluções para obtenção das *VCs* em experimentos instalados sob o DIC, DBC e DQL, por meio de procedimentos disponíveis no software R para as *FDs*.

2.2. Específicos

Estabelecer mudanças nos modelos estatísticos do DBC e do DQL para conterem apenas as fontes de variação relativas ao DIC (entre e dentro dos tratamentos), a fim de utilizá-los para obtenção das *VCs* por meio das *FDs*;

Demonstrar a obtenção das *VCs* por meio das *FDs* obtidas pela função *lda* do software R a partir dos valores originais do DIC e dos transformados do DBC e DQL;

Demonstrar a obtenção das *VCs* subsequente à MANOVA obtidas pelas funções *manova*, *solve* e *eigen* do software R; e

Comparar diferentes padronizações para obtenção dos autovetores utilizados como coeficientes de ponderação das variáveis originais nas *VCs*.

3. REVISÃO BIBLIOGRÁFICA

3.1. Variáveis Canônicas

Uma das formas de facilitar a compreensão dos resultados de experimentos com muitas variáveis é utilizar o método multivariado da análise de variáveis canônicas (*VCs*). Isso porque é possível reduzir a dimensionalidade dos dados pela obtenção de variáveis ou de funções latentes que possam representar grande parte da variabilidade contida nas diferentes variáveis. Além disso, as *VCs* podem ser utilizadas para representações gráficas e testes univariados que possam indicar os tratamentos mais adequados. Resumidamente, no contexto de experimentos com repetições dos tratamentos e executados sob diferentes delineamentos de casualização tais como o inteiramente casualizado (DIC), em blocos casualizados (DBC), em quadrado latino (DQL), as *VCs* nada mais são do que combinações lineares das p variáveis ($vc = \hat{\mathbf{a}}' \mathbf{y}$) (CRUZ, 2006; CRUZ et al., 2012; EVERITT, 2005; MARDIA et al., 1997).

Matematicamente, as *VCs* são obtidas pela maximização da razão da variação entre os tratamentos com a variação dentro deles (ou residual) com um número s de soluções possíveis. E isto, realizado *a posteriori* à análise de variância multivariada (MANOVA). As *VCs* são independentes entre si e conservam, em ordem decrescente de obtenção, a variabilidade da matriz composta pela razão entre as matrizes das somas de quadrados e de produtos entre (\mathbf{H}) e dentro (\mathbf{E}) dos tratamentos referentes às p variáveis-resposta Ys consideradas no experimento. Essa ordem de variância na obtenção das *VCs* é útil, pois permite substituir, de forma aproximada, a informação contida nas p variáveis-resposta Ys pela informação contida pelas s *VCs*.

O número s de *VCs* é um mínimo entre o número p de variáveis originais (variáveis-resposta Ys) e o número de graus de liberdade dos tratamentos da MANOVA ($t - 1$), ou seja, $s = \text{mín.}(p, t - 1)$. Assim, para $p \leq t - 1$, tem-se $s = p$. Por outro lado, quando o número de Ys for superior aos graus de liberdade dos tratamentos, o número s de *VCs* necessário para a análise é, no máximo, igual a $t - 1$.

A solução para obtenção dos escores das s *VCs* ($\widehat{VC}_1, \widehat{VC}_2, \dots, \widehat{VC}_s$) passa pela busca de um vetor estimado $\hat{\mathbf{a}}$ que maximize a relação da variância entre os t tratamentos com a variância comum dentro de todos eles (residual). A solução é um problema de maximização da razão de formas quadráticas, como segue:

$$\frac{\hat{\mathbf{a}}' \mathbf{H} \hat{\mathbf{a}}}{\hat{\mathbf{a}}' \mathbf{E} \hat{\mathbf{a}}}$$

A variação entre os tratamentos é expressa pela matriz amostral \mathbf{H} ($p \times p$) composta pelas somas de quadrados e de produtos dos tratamentos entre as p variáveis-resposta Y_s ; e a variação residual é expressa pela matriz amostral \mathbf{E} ($p \times p$) composta pelas somas de quadrados e de produtos do resíduo entre as p variáveis-resposta Y_s . Para o caso particular das VCs , os s autovetores e os s autovalores serão sempre estimados a partir de amostras experimentais e, portanto, são do tipo $\hat{\mathbf{a}}_m$ e $\hat{\lambda}_m$, respectivamente ($m = 1, 2, \dots, s$).

O autovalor de ordem m ($\hat{\lambda}_m$) é estimado por meio do determinante da expressão baseada na matriz $\mathbf{E}^{-1}\mathbf{H}$, tal que:

$$|\mathbf{E}^{-1}\mathbf{H} - \hat{\lambda}_m\mathbf{I}| = 0, \text{ em que:}$$

$\hat{\lambda}_m$ = estimativa do autovalor de ordem m ($m = 1, 2, \dots, s$) da matriz $\mathbf{E}^{-1}\mathbf{H}$, tal que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_s$; e

\mathbf{I} = matriz identidade $s \times s$.

E, de acordo com a estimativa do autovalor de ordem m ($\hat{\lambda}_m$) da matriz $\mathbf{E}^{-1}\mathbf{H}$, obtém-se a estimativa do autovetor não padronizado ($p \times 1$) de ordem m ($\hat{\mathbf{a}}_m$), de forma que o máximo da razão $\frac{\hat{\mathbf{a}}'\mathbf{H}\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\mathbf{E}\hat{\mathbf{a}}}$ seja obtido pela solução do sistema de equações homogêneas, como segue:

$$(\mathbf{E}^{-1}\mathbf{H} - \hat{\lambda}_m\mathbf{I})\hat{\mathbf{a}}_m = \mathbf{0}, \text{ em que:}$$

$\hat{\mathbf{a}}_m = [\hat{a}_{m1} \quad \hat{a}_{m2} \quad \dots \quad \hat{a}_{ms}]$ = estimativa do autovetor não padronizado de ordem m ($m = 1, 2, \dots, s$) associado à estimativa do autovalor de ordem m ($m = 1, 2, \dots, s$) da matriz $\mathbf{E}^{-1}\mathbf{H}$; e

$\mathbf{0}$ = vetor nulo $s \times 1$.

Desse modo, para p variáveis-resposta Y_s , têm-se as seguintes VCs :

$$\widehat{VC}_1 = \hat{a}_{11}Y_1 + \hat{a}_{12}Y_2 + \dots + \hat{a}_{1p}Y_p;$$

$$\widehat{VC}_2 = \hat{a}_{21}Y_1 + \hat{a}_{22}Y_2 + \dots + \hat{a}_{2p}Y_p;$$

...

$$\widehat{VC}_s = \hat{a}_{s1}Y_1 + \hat{a}_{s2}Y_2 + \dots + \hat{a}_{sp}Y_p.$$

Como $\frac{\hat{\mathbf{a}}'\mathbf{H}\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\mathbf{E}\hat{\mathbf{a}}}$ é uma razão, ela não varia para qualquer padronização de $\hat{\mathbf{a}}$, ou seja, $\frac{\hat{\mathbf{a}}'\mathbf{H}\hat{\mathbf{a}}}{\hat{\mathbf{a}}'\mathbf{E}\hat{\mathbf{a}}}$ continua constante para $\hat{\mathbf{a}}$ ou para $c\hat{\mathbf{a}}$, sendo c uma constante. Dessa forma, diferentes padronizações para $\hat{\mathbf{a}}$ podem ser utilizadas para obtenção das VCs .

Uma das padronizações é utilizar a matriz residual das somas de quadrados e de produtos \mathbf{E} para ponderar a estimativa de um autovetor $\hat{\mathbf{g}}_m$ ($p \times 1$) padronizado de ordem m , de tal modo que:

$$\hat{\mathbf{g}}_m' \mathbf{E} \hat{\mathbf{g}}_m = 1; \text{ e}$$

$$\hat{\mathbf{g}}_m' \mathbf{E} \hat{\mathbf{g}}_{m'} = 0, \text{ para } m \neq m' (m, m' = 1, 2, \dots, s).$$

Portanto, o vetor padronizado é:

$$\hat{\mathbf{g}}_m = [\hat{g}_{m1} \quad \hat{g}_{m2} \quad \dots \quad \hat{g}_{mp}].$$

Assim, as s VCS ($\widehat{VC}_1, \widehat{VC}_2, \dots, \widehat{VC}_s$) são obtidas por:

$$\widehat{VC}_1 = \hat{g}_{11}Y_1 + \hat{g}_{12}Y_2 + \dots + \hat{g}_{1p}Y_p;$$

$$\widehat{VC}_2 = \hat{g}_{21}Y_1 + \hat{g}_{22}Y_2 + \dots + \hat{g}_{2p}Y_p;$$

...

$$\widehat{VC}_s = \hat{g}_{s1}Y_1 + \hat{g}_{s2}Y_2 + \dots + \hat{g}_{sp}Y_p.$$

Mardia et al. (1997) recomendam utilizar a matriz de variâncias e covariâncias residuais \mathbf{S} para ponderar a estimativa de um autovetor $\hat{\mathbf{e}}_m$ ($p \times 1$) padronizado de ordem m , de tal modo que:

$$\hat{\mathbf{e}}_m' \mathbf{S} \hat{\mathbf{e}}_m = 1; \text{ e}$$

$$\hat{\mathbf{e}}_m' \mathbf{S} \hat{\mathbf{e}}_{m'} = 0, \text{ para } m \neq m' (m, m' = 1, 2, \dots, s), \text{ em que:}$$

$$\mathbf{S} = \frac{\mathbf{E}}{n_E}; \text{ e}$$

n_E = número de graus de liberdade do resíduo da MANOVA.

Portanto, o vetor padronizado é:

$$\hat{\mathbf{e}}_m = [\hat{e}_{m1} \quad \hat{e}_{m2} \quad \dots \quad \hat{e}_{mp}].$$

Assim, as s VCS ($\widehat{VC}_1, \widehat{VC}_2, \dots, \widehat{VC}_s$) são obtidas por:

$$\widehat{VC}_1 = \hat{e}_{11}Y_1 + \hat{e}_{12}Y_2 + \dots + \hat{e}_{1p}Y_p;$$

$$\widehat{VC}_2 = \hat{e}_{21}Y_1 + \hat{e}_{22}Y_2 + \dots + \hat{e}_{2p}Y_p;$$

...

$$\widehat{VC}_s = \hat{e}_{s1}Y_1 + \hat{e}_{s2}Y_2 + \dots + \hat{e}_{sp}Y_p.$$

Para obtenção dessa segunda estimativa do autovetor padronizado ($p \times 1$) de ordem m , pode-se utilizar a seguinte expressão:

$$\hat{\mathbf{e}}_m = \frac{1}{\sqrt{\hat{\mathbf{a}}_m' \mathbf{S} \hat{\mathbf{a}}_m}} \hat{\mathbf{a}}_m.$$

E para obtenção da primeira estimativa, tem-se:

$$\hat{\mathbf{g}}_m = \frac{1}{\sqrt{n}} \hat{\mathbf{e}}_m.$$

Seja qual for a padronização dos autovetores utilizados para obtenção dos escores das *VCs*, os testes e as interpretações serão equivalentes, visto que:

$$\frac{\hat{\mathbf{g}}'_m \mathbf{H} \hat{\mathbf{g}}_m}{\hat{\mathbf{g}}'_m \mathbf{E} \hat{\mathbf{g}}_m} = \frac{\hat{\mathbf{e}}'_m \mathbf{H} \hat{\mathbf{e}}_m}{\hat{\mathbf{e}}'_m \mathbf{E} \hat{\mathbf{e}}_m}.$$

Após obtenção dos escores das *VCs*, o número d de dimensões ($\leq s$) pode ser escolhido pela aplicação do teste de Bartlett, revisado por Mardia et al. (1997), que considera a dimensionalidade como sendo o número de autovalores da matriz $\mathbf{E}^{-1}\mathbf{H}$ significativamente diferentes de zero. Esse teste é realizado, sequencialmente, para $d = 0$, $d = 1$, e assim sucessivamente, até que um resultado não significativo apareça. Assim, se as hipóteses forem rejeitadas para $d = 0$ e $d = 1$, mas se não forem para $d = 2$, a dimensionalidade será igual a 2. Portanto, se até em $d - 1$ forem obtidos resultados significativos (rejeita-se H_0), mas em d não (não se rejeita H_0), infere-se que a dimensionalidade é igual a d . Se todas as hipóteses para $d = 0, 1, 2, \dots, s$ forem rejeitadas, a dimensionalidade d será igual a s . Consequentemente, a hipótese H_0 para cada valor d , separadamente, é dada por:

H_0 : a dimensionalidade do conjunto de dados é igual a d ($d = 0, 1, 2, \dots, s$).

A estatística do teste segue, de forma aproximada, a distribuição qui-quadrado (χ^2) e apresenta o seguinte valor calculado:

$$D_d^2 = \left(n_E - \frac{p-t+2}{2} \right) \sum_{m=d+1}^s \ln(1 + \hat{\lambda}_m), \text{ em que:}$$

$\hat{\lambda}_m$ = estimativa do autovalor de ordem m ($m = 1, 2, \dots, s$) da matriz $\mathbf{E}^{-1}\mathbf{H}$; e

n_E = número de graus de liberdade do resíduo da MANOVA.

Se $D_d^2 \geq \chi_{\text{tab}}^2$ ($p\text{-valor} \leq \alpha$), rejeita-se H_0 , sendo χ_{tab}^2 o valor obtido da distribuição χ^2 com $(p - d)(t - 1 - d)$ graus de liberdade que deixa uma probabilidade α na extremidade da cauda à direita.

Outra forma de escolher o valor d é fazê-lo em função das primeiras *VCs* que acumularem pelo menos 70% ou 80% da importância relativa (IR). A IR é a porcentagem da variação entre os t tratamentos em relação à variação residual, explicada pelas d ($d \leq s$) primeiras *VCs*, ou seja, é o resultado em percentual da divisão da soma dos d primeiros autovalores estimados pela soma de todos os k autovalores estimados, como segue:

$$IR(\%) = 100 \times \frac{\sum_{m=1}^d \hat{\lambda}_m}{\sum_{m=1}^s \hat{\lambda}_m}, \text{ em que:}$$

d = número de *VCs* analisadas; e

s = mín. ($p, t - 1$).

3.2. Análise Discriminante

É interessante constatar uma densa literatura científica que relata o uso da análise discriminante como método estatístico multivariado ao longo dos últimos 80 anos. Os campos aplicados são os mais diversos e incluem as ciências biológicas, ecologia, agrárias, sociais, engenharias, física e medicina (McLACHLAN, 1992; MINGOTI, 2005; FERREIRA, 2018).

A análise discriminante envolve basicamente a relação de um conjunto de variáveis quantitativas interrelacionadas para discriminação e/ou classificação de observações (elementos amostrais) em grupos (populações) distintos estabelecidos e conhecidos *a priori* (McLACHLAN, 1992; FERREIRA, 2018). É importante ressaltar que a análise discriminante se distingue da análise de agrupamentos, na qual o número de populações não é conhecido de antemão e o objetivo é agrupar as populações por suas similaridades (FERREIRA, 2018).

Assim, seja $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_p]$ um vetor contendo as medidas de p variáveis-resposta Y_s de um elemento amostral (objeto), o qual se deseja classificar em uma das t diferentes populações conhecidas *a priori*. Considerando um caso contínuo – embora o discreto possa ser construído de forma análoga – cada população x_i é caracterizada pela sua função densidade de probabilidade $f_i(\mathbf{y})$, sendo $i = 1, 2, \dots, t$. Seja também $f_i(\mathbf{y})$ diferente de $f_{i'}(\mathbf{y})$ apenas pelos seus parâmetros e não pela família de densidade ($\forall i \neq i'$). Isso significa que as formas funcionais de $f_i(\mathbf{y})$ e $f_{i'}(\mathbf{y})$ são idênticas. O desafio metodológico da análise discriminante é estipular uma regra de decisão para atribuição definitiva do objeto em uma das t populações com base nas suas medidas (McLACHLAN, 1992; FERREIRA, 2018). É importante ressaltar que qualquer regra que seja criada pode incorrer em classificações incorretas e, portanto, o objetivo é adequar uma regra que contemple o risco de uma decisão incorreta e o minimize (FERREIRA, 2018).

3.2.1. Duas Populações

Um dos métodos mais simples e objetivos criados para discriminar populações *a priori* e classificar objetos de origem incerta em uma delas foi proposto pelo matemático e geneticista Fisher. Inicialmente, o que Fisher fez foi criar uma função discriminante linear que fosse utilizada para discriminação e classificação considerando, inicialmente, a existência de apenas duas populações, sob a única pressuposição de que estas fossem homocedásticas (ou seja, $\Sigma_1 = \Sigma_2$) sem que necessariamente assumissem normalidade

multivariada (FISHER, 1936, 1938). Embora, normalidade multivariada seja requerida quando inferências estatísticas são realizadas (OWEN & CHMIELEWSKI, 1985).

Para recriar a FD para duas populações, considere um vetor aleatório \mathbf{y} p -dimensional $[y_1 \ y_2 \ \dots \ y_p]$ proveniente de uma entre duas populações quaisquer com funções densidades de probabilidades $f_1(\mathbf{y})$ e $f_2(\mathbf{y})$. Os vetores de médias populacionais são $\boldsymbol{\mu}_1 = [\mu_{11} \ \mu_{12} \ \dots \ \mu_{1p}]$ e $\boldsymbol{\mu}_2 = [\mu_{21} \ \mu_{22} \ \dots \ \mu_{2p}]$ para as populações x_1 e x_2 , respectivamente. As matrizes de variâncias e covariâncias são iguais em ambas as populações ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$), tal que:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}.$$

Assim, a combinação linear das variáveis $FD = \mathbf{a}'\mathbf{Y}$ estabelecida pelo vetor p -dimensional \mathbf{a} fornece as seguintes médias condicionais da variável FD :

$$\mu_{1FD} = E(FD|x_1) = E(\mathbf{a}'\mathbf{Y}|x_1) = \mathbf{a}'\boldsymbol{\mu}_1; \text{ e}$$

$$\mu_{2FD} = E(FD|x_2) = E(\mathbf{a}'\mathbf{Y}|x_2) = \mathbf{a}'\boldsymbol{\mu}_2.$$

Já a variância é dada, indistintamente, para as duas populações, por:

$$\sigma_{FD}^2 = Var(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'Var(\mathbf{Y})\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}.$$

O índice representando a distância quadrática Δ^2 é então definido para maximizar a magnitude da diferença entre as médias condicionais da variável FD combinada com sua variância comum assumida dentro das duas populações. Assim, tem-se:

$$\Delta^2 = \frac{(\mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\boldsymbol{\mu}_2)^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} = \frac{[\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}.$$

A situação de maior discriminação entre as duas populações irá corresponder àquela que maximizar a combinação linear Δ^2 . Para isso, deve-se determinar o vetor \mathbf{a} tal que:

$$\text{máx.}_a(\Delta^2) = \text{máx.}_a \frac{[\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}.$$

Essa condição é determinada através da primeira derivada da combinação linear Δ^2 em relação ao vetor \mathbf{a} , cujo resultado deve ser igualado a zero para obtenção da solução:

$$\begin{aligned} \frac{\partial \Delta^2}{\partial \mathbf{a}} &= \frac{2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{a}(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}) - 2[\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2\boldsymbol{\Sigma}\mathbf{a}}{(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})^2} = \\ &= \frac{2}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{a} - \Delta^2\boldsymbol{\Sigma}\mathbf{a}]. \end{aligned}$$

Igualando-se a $\mathbf{0}$, tem-se o seguinte sistema de equações homogêneas:

$$[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' - \Delta^2 \boldsymbol{\Sigma}] \mathbf{a} = \mathbf{0}.$$

Porém, ele terá solução diferente da solução trivial $\mathbf{0}$ se:

$$|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' - \Delta^2 \boldsymbol{\Sigma}| = 0.$$

A solução do sistema é obtida através da aplicação das propriedades de determinantes de matrizes particionadas, como segue:

$$|(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' - \Delta^2 \boldsymbol{\Sigma}| = 1 - \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\Delta^2} = 0.$$

Portanto, o máximo da distância quadrática entre as médias das populações univariadas resultantes da transformação linear estabelecida por \mathbf{a} que melhor discrimina as duas populações é dado por:

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

O vetor \mathbf{a} de combinações lineares que maximiza a distância quadrática é, então, obtido por:

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Tal resultado surge após a aplicação da restrição $\mathbf{a}' \boldsymbol{\Sigma}^{-1} \mathbf{a} = \Delta^2$, como segue:

$$\Delta^2 = \frac{[\mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\mathbf{a}' \boldsymbol{\Sigma}^{-1} \mathbf{a}};$$

$$[\mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2 = \Delta^2 \mathbf{a}' \boldsymbol{\Sigma}^{-1} \mathbf{a};$$

$$[\mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2 = (\Delta^2)^2;$$

$$\mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Delta^2;$$

$$\mathbf{a}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'; \text{ e}$$

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Assim, a combinação linear do vetor \mathbf{y} conhecida como *FD* é dada por:

$$fd = \mathbf{a}' \mathbf{y} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{y}.$$

Desse modo, para p variáveis-resposta Y_s , tem-se a seguinte *FD*:

$$FD_1 = a_{11} Y_1 + a_{12} Y_2 + \dots + a_{1p} Y_p.$$

Uma vez que $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ é o máximo da distância entre as médias populacionais das distribuições condicionais de *FD*, o ponto médio entre as médias univariadas dessa distribuição condicional servirá como um critério adequado para classificar uma observação \mathbf{y} em uma das duas populações. Logo, denomina-se m como esse ponto médio, tal que:

$$m = \frac{1}{2} (\mu_{1_{FD}} + \mu_{2_{FD}}) = \frac{1}{2} (\mathbf{a}' \boldsymbol{\mu}_1 + \mathbf{a}' \boldsymbol{\mu}_2) = \frac{1}{2} \mathbf{a}' (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) =$$

$$\frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Considerando a variável aleatória $FD = \mathbf{a}'\mathbf{Y}$, tem-se:

$$\begin{aligned} E(FD|x_1) - m &= E[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{Y} | x_1] - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \\ &= E[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1] - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \\ &= \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2} \Delta^2. \end{aligned}$$

A quantidade $\frac{1}{2} \Delta^2$ representa a metade da distância de Mahalanobis entre os centroides das duas populações e, portanto, é positiva. Dessa forma, uma observação multivariada \mathbf{y} será classificada em x_1 se o resultado de $fd - m \geq 0$.

Se a esperança condicional de FD for em relação a x_2 , tem-se:

$$\begin{aligned} E(FD|x_2) - m &= E[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{Y} | x_2] - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \\ &= E[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2] - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \\ &= -\frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2} \Delta^2. \end{aligned}$$

Uma vez que o valor esperado do desvio de FD em relação a m é negativo, uma observação multivariada \mathbf{y} será classificada em x_2 se o resultado de $fd - m < 0$.

Finalmente, a regra de classificação pode ser resumida em classificar \mathbf{y} em x_1 se $fd \geq m$, ou em x_2 , caso contrário, sendo:

$$fd = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{y} \text{ e } m = 0,5 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

A FD foi desenvolvida supondo o conhecimento dos parâmetros populacionais. Porém, não é comum ter a informação real sobre esses parâmetros e, portanto, é necessário obtê-los de amostras. Assim, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ podem ser substituídos pelas suas respectivas estimativas $\bar{\mathbf{y}}_1$, $\bar{\mathbf{y}}_2$ e \mathbf{S} , sendo n_1 e n_2 o número de observações p -variadas amostradas das populações x_1 e x_2 , respectivamente:

$$\begin{aligned} \bar{\mathbf{y}}_1 &= [\bar{y}_{11} \quad \bar{y}_{12} \quad \cdots \quad \bar{y}_{1p}] \text{ e } \bar{\mathbf{y}}_2 = [\bar{y}_{21} \quad \bar{y}_{22} \quad \cdots \quad \bar{y}_{2p}]; \\ \mathbf{S}_1 &= \begin{bmatrix} S_{111} & S_{112} & \cdots & S_{11p} \\ S_{121} & S_{122} & \cdots & S_{12p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1p1} & S_{1p2} & \cdots & S_{1pp} \end{bmatrix} \text{ e } \mathbf{S}_2 = \begin{bmatrix} S_{211} & S_{212} & \cdots & S_{21p} \\ S_{221} & S_{222} & \cdots & S_{22p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{2p1} & S_{2p2} & \cdots & S_{2pp} \end{bmatrix}; \text{ e} \\ \mathbf{S} &= \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}. \end{aligned}$$

Portanto, a FD estimada é:

$$\widehat{fd} = \widehat{\mathbf{a}}' \mathbf{y} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}^{-1} \mathbf{y}.$$

E o ponto médio entre as médias amostrais das observações univariadas \widehat{fd} é:

$$\widehat{m} = \frac{1}{2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2).$$

Desse modo, deve-se classificar y em x_1 se $\widehat{fd} - \widehat{m} \geq 0$; caso contrário, a observação multivariada deve ser classificada em x_2 .

3.2.2. Duas ou Mais de Duas Populações

A análise discriminante proposta por Fisher também foi generalizada, ou seja, recriada para quando fossem consideradas duas ou mais de duas populações e, novamente, o método pressupunha que as matrizes de variâncias e covariâncias para as t populações conhecidas fossem homogêneas ($\Sigma_1 = \Sigma_2 = \dots = \Sigma_t = \Sigma$), sem que houvesse a necessidade de normalidade p -variada (FISHER, 1936, 1938). Dessa forma, a matriz de variâncias e covariâncias considerada para o método é única, de tal forma que:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}.$$

Basicamente, o objetivo do método criado por Fisher foi conservar a maior parte da variabilidade original dos dados em apenas algumas poucas transformações (funções discriminantes) lineares das p variáveis-resposta Y s, que pudessem ser utilizadas para separação de um ou mais objetos em uma das t populações conhecidas (FERREIRA, 2018).

O caso geral para a análise discriminante linear de Fisher pode ser descrito de acordo com os seguintes passos:

Inicialmente, considere que cada população x_i possua um vetor de médias μ_i ($i = 1, 2, \dots, t$). Portanto, o vetor geral de médias das t populações é:

$$\mu = \frac{1}{t} \sum_{i=1}^t \mu_i.$$

A matriz das somas de quadrados e de produtos entre os t vetores de médias populacionais é dada pela matriz B_μ ($p \times p$):

$$B_\mu = \sum_{i=1}^t (\mu_i - \mu)(\mu_i - \mu)'$$

Considerando a combinação linear $fd = \mathbf{a}'\mathbf{y}$ do vetor aleatório p -dimensional \mathbf{y} , a esperança condicional da FD em cada população x_i ($i = 1, 2, \dots, t$) é dada por:

$$\mu_{i_{FD}} = E(FD|x_i) = E(\mathbf{a}'\mathbf{Y}|x_i) = \mathbf{a}'E(\mathbf{Y}|x_i) = \mathbf{a}'\mu_i.$$

Já a variância da FD , por:

$$\sigma_{i_{FD}}^2 = \sigma_{FD}^2 = Var(\mathbf{a}'\mathbf{Y}|x_i) = \mathbf{a}'Var(\mathbf{Y}|x_i)\mathbf{a} = \mathbf{a}'\Sigma\mathbf{a}.$$

E a média geral da FD é dada por:

$$\mu_{FD} = \frac{1}{t} \sum_{i=1}^t \mu_{iFD} = \frac{1}{t} \sum_{i=1}^t \mathbf{a}' \boldsymbol{\mu}_i = \mathbf{a}' \left(\frac{1}{t} \sum_{i=1}^t \boldsymbol{\mu}_i \right) = \mathbf{a}' \boldsymbol{\mu}.$$

Estabelecidos os parâmetros das médias das t populações, um vetor de combinações lineares \mathbf{a} é utilizado, a fim de maximizar a soma das distâncias generalizadas entre as médias de cada população x_i (μ_{iFD}) e a média geral das t populações (μ_{FD}) na escala univariada. Essa distância pode ser caracterizada como se segue:

$$\begin{aligned} \frac{\sum_{i=1}^t (\mu_{iFD} - \mu_{FD})^2}{\sigma_{FD}^2} &= \frac{\sum_{i=1}^t [\mathbf{a}'(\boldsymbol{\mu}_i - \boldsymbol{\mu})]^2}{\mathbf{a}' \boldsymbol{\Sigma} \mathbf{a}} = \frac{\sum_{i=1}^t [\mathbf{a}'(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' \mathbf{a}]}{\mathbf{a}' \boldsymbol{\Sigma} \mathbf{a}} = \\ &= \frac{\mathbf{a}' \sum_{i=1}^t [(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'] \mathbf{a}}{\mathbf{a}' \boldsymbol{\Sigma} \mathbf{a}} = \frac{\mathbf{a}' \mathbf{B}_\mu \mathbf{a}}{\mathbf{a}' \boldsymbol{\Sigma} \mathbf{a}}. \end{aligned}$$

A expressão $\frac{\mathbf{a}' \mathbf{B}_\mu \mathbf{a}}{\mathbf{a}' \boldsymbol{\Sigma} \mathbf{a}}$ representa a razão entre a variabilidade entre as t populações e a variabilidade comum dentro delas. A solução para o vetor \mathbf{a} que maximiza essa razão é obtida a partir da maximização da razão de formas quadráticas e possui um número s de soluções, tal que: $s = \text{mín.}(p, t - 1)$. Os s vetores latentes \mathbf{a}_m são obtidos do sistema de equações homogêneas:

$$(\mathbf{B}_\mu - \lambda_m \boldsymbol{\Sigma}) \mathbf{a}_m = \mathbf{0} \quad (m = 1, 2, \dots, s).$$

Consequentemente, são geradas s FDs :

$$fd_m = \mathbf{a}'_m \mathbf{y} \quad (m = 1, 2, \dots, s).$$

Desse modo, as s FDs possíveis de serem obtidas para uso na discriminação e/ou classificação, são:

$$FD_1 = a_{11}Y_1 + a_{12}Y_2 + \dots + a_{1p}Y_p;$$

$$FD_2 = a_{21}Y_1 + a_{22}Y_2 + \dots + a_{2p}Y_p;$$

...

$$FD_s = a_{s1}Y_1 + a_{s2}Y_2 + \dots + a_{sp}Y_p.$$

Vale ressaltar aqui que os passos anteriores para obtenção das FDs se assemelham com aqueles utilizados para obtenção das VCs . Essa equivalência surge porque ambos os procedimentos são realizados de forma a encontrar um vetor \mathbf{a} que maximize uma razão de variâncias entre e dentro dos tratamentos ou das populações. Dessa forma, as VCs podem ser representadas por FDs obtidas de experimentos multivariados instalados sob o DIC, onde os t tratamentos funcionam como as t populações para as FDs e onde apenas as variações entre e dentro dos tratamentos estão presentes.

Todavia, para o caso da análise discriminante, uma regra de classificação ainda deve ser gerada para atender o objetivo da análise, que consiste em classificar um novo objeto \mathbf{y} em uma das t populações. Essa regra de classificação pode ser estabelecida a partir da definição dos vetores \mathbf{fd} e $\boldsymbol{\mu}_{i_{FD}}$ ($s \times 1$), os quais são gerados com os valores das s FDs , respectivamente, a partir do vetor de observações \mathbf{y} e do vetor de médias $\boldsymbol{\mu}_i$ de cada população x_i ($i = 1, 2, \dots, t$). Assim, se a matriz \mathbf{Q} ($p \times s$) for definida como uma matriz cujas colunas são formadas pelos s vetores latentes \mathbf{a}_m ($m = 1, 2, \dots, s$), tem-se:

$$\mathbf{Q} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_s].$$

Portanto, \mathbf{fd} e $\boldsymbol{\mu}_{i_{FD}}$ são formados, respectivamente, por $\mathbf{Q}'\mathbf{y}$ e $\mathbf{Q}'\boldsymbol{\mu}_i$, tal que:

$$\mathbf{fd} = \mathbf{Q}'\mathbf{y} = \begin{bmatrix} fd_1 \\ fd_2 \\ \vdots \\ fd_s \end{bmatrix} \text{ e } \boldsymbol{\mu}_{i_{FD}} = \mathbf{Q}'\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i_{FD1}} \\ \mu_{i_{FD2}} \\ \vdots \\ \mu_{i_{FDs}} \end{bmatrix}.$$

Segundo Fisher (1938), o objeto correspondente ao vetor \mathbf{y} será alocado na população x_i se a distância entre o vetor \mathbf{fd} e o vetor $\boldsymbol{\mu}_{i_{FD}}$ for menor que a distância entre o mesmo vetor \mathbf{fd} e $\boldsymbol{\mu}_{i'_{FD}}$, para $i \neq i'$ ($i, i' = 1, 2, \dots, t$).

Antes, sejam estabelecidas a esperança matemática e a variância condicionais em cada população x_i obtidas a partir do vetor aleatório \mathbf{FD} ($s \times 1$), tal que:

$$\mathbf{FD} = \mathbf{Q}'\mathbf{Y} = \begin{bmatrix} \mathbf{a}'_1\mathbf{Y} \\ \mathbf{a}'_2\mathbf{Y} \\ \vdots \\ \mathbf{a}'_s\mathbf{Y} \end{bmatrix} = \begin{bmatrix} FD_1 \\ FD_2 \\ \vdots \\ FD_s \end{bmatrix}.$$

Desse modo, verifica-se que a média da FD é dada por:

$$E(\mathbf{FD}|x_i) = E(\mathbf{Q}'\mathbf{Y}|x_i) = \begin{bmatrix} \mathbf{a}'_1\boldsymbol{\mu}_i \\ \mathbf{a}'_2\boldsymbol{\mu}_i \\ \vdots \\ \mathbf{a}'_s\boldsymbol{\mu}_i \end{bmatrix} = \mathbf{Q}'\boldsymbol{\mu}_i = \boldsymbol{\mu}_{i_{FD}}.$$

E a covariância entre duas FDs , por:

$$Cov(\mathbf{FD}|x_i) = Cov(\mathbf{Q}'\mathbf{Y}|x_i) = \begin{bmatrix} \mathbf{a}'_1\boldsymbol{\Sigma}\mathbf{a}_1 & \mathbf{a}'_1\boldsymbol{\Sigma}\mathbf{a}_2 & \cdots & \mathbf{a}'_1\boldsymbol{\Sigma}\mathbf{a}_s \\ \mathbf{a}'_2\boldsymbol{\Sigma}\mathbf{a}_1 & \mathbf{a}'_2\boldsymbol{\Sigma}\mathbf{a}_2 & \cdots & \mathbf{a}'_2\boldsymbol{\Sigma}\mathbf{a}_s \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}'_s\boldsymbol{\Sigma}\mathbf{a}_1 & \mathbf{a}'_s\boldsymbol{\Sigma}\mathbf{a}_2 & \cdots & \mathbf{a}'_s\boldsymbol{\Sigma}\mathbf{a}_s \end{bmatrix} = \mathbf{I}_s.$$

Uma vez que os componentes da FD possuem variâncias unitárias e não são correlacionados, uma medida apropriada de distância quadrática entre o vetor de observações \mathbf{fd} e o vetor $\boldsymbol{\mu}_{i_{FD}}$ será suficiente para estabelecer a regra de decisão:

$$(\mathbf{fd} - \boldsymbol{\mu}_{i_{FD}})'(\mathbf{fd} - \boldsymbol{\mu}_{i_{FD}}) = \sum_{m=1}^s [\mathbf{a}'_m(\mathbf{y} - \boldsymbol{\mu}_i)]^2 = \sum_{m=1}^s (fd_m - \mu_{i_{FDm}})^2.$$

Portanto, pode-se alocar o vetor \mathbf{y} na população x_i se:

$\sum_{m=1}^s [\mathbf{a}'_m (\mathbf{y} - \boldsymbol{\mu}_i)]^2 = \sum_{m=1}^s (fd_m - \mu_{iFD_m})^2 = \min_l \sum_{m=1}^s [\mathbf{a}'_m (\mathbf{y} - \boldsymbol{\mu}_l)]^2$, para $l = 1, 2, \dots, t$.

A *FD* pode ainda ser reformulada da seguinte maneira:

$$\begin{aligned} (\mathbf{fd} - \boldsymbol{\mu}_{iFD})' (\mathbf{fd} - \boldsymbol{\mu}_{iFD}) &= \sum_{m=1}^s [\mathbf{a}'_m (\mathbf{y} - \boldsymbol{\mu}_i)]^2 = \sum_{m=1}^s (fd_m - \mu_{iFD_m})^2 = \\ &= (\mathbf{Q}' \mathbf{y} - \mathbf{Q}' \boldsymbol{\mu}_i)' (\mathbf{Q}' \mathbf{y} - \mathbf{Q}' \boldsymbol{\mu}_i) = \\ &= [\mathbf{Q}' (\mathbf{y} - \boldsymbol{\mu}_i)]' \mathbf{Q}' (\mathbf{y} - \boldsymbol{\mu}_i) = \\ &= (\mathbf{y} - \boldsymbol{\mu}_i)' \mathbf{Q} \mathbf{Q}' (\mathbf{y} - \boldsymbol{\mu}_i). \end{aligned}$$

Contudo, como a matriz $\mathbf{Q} = \sum^{-\frac{1}{2}} \mathbf{P}$, sendo \mathbf{P} uma matriz ortogonal cujas colunas são formadas pelos autovetores de $\sum^{-\frac{1}{2}} \mathbf{B}_\mu \sum^{-\frac{1}{2}}$, tem-se, então:

$$\begin{aligned} (\mathbf{fd} - \boldsymbol{\mu}_{iFD})' (\mathbf{fd} - \boldsymbol{\mu}_{iFD}) &= (\mathbf{y} - \boldsymbol{\mu}_i)' \sum^{-\frac{1}{2}} \mathbf{P} \left(\sum^{-\frac{1}{2}} \mathbf{P} \right)' (\mathbf{y} - \boldsymbol{\mu}_i) = \\ &= (\mathbf{y} - \boldsymbol{\mu}_i)' \sum^{-\frac{1}{2}} \mathbf{P} \mathbf{P}' \sum^{-\frac{1}{2}} (\mathbf{y} - \boldsymbol{\mu}_i) = \\ &= (\mathbf{y} - \boldsymbol{\mu}_i)' \sum^{-\frac{1}{2}} \sum^{-\frac{1}{2}} (\mathbf{y} - \boldsymbol{\mu}_i) = \\ &= (\mathbf{y} - \boldsymbol{\mu}_i)' \sum^{-1} (\mathbf{y} - \boldsymbol{\mu}_i). \end{aligned}$$

Portanto, pelo critério de Fisher (1938), minimizar a distância quadrática $\sum_{m=1}^s (fd_m - \mu_{iFD_m})^2$ equivale a maximizar a distância de Mahalanobis entre a observação multivariada \mathbf{y} e o vetor de médias da i -ésima população:

$$(\mathbf{y} - \boldsymbol{\mu}_i)' \sum^{-1} (\mathbf{y} - \boldsymbol{\mu}_i).$$

Essa distância pode ser expandida de tal forma que:

$$\begin{aligned} (\mathbf{y} - \boldsymbol{\mu}_i)' \sum^{-1} (\mathbf{y} - \boldsymbol{\mu}_i) &= (\mathbf{y} - \boldsymbol{\mu}_i)' \sum^{-1} (\mathbf{y} - \boldsymbol{\mu}_i); \\ \mathbf{y}' \sum^{-1} \mathbf{y} - 2 \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} + \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i &= \mathbf{y}' \sum^{-1} \mathbf{y} - 2 \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} + \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i; \text{ e} \\ -2 \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} + \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i &= -2 \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} + \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i. \end{aligned}$$

Multiplicando-se a última equação anterior por (-1) e dividindo-a por (2) , tem-se:

$$\begin{aligned} \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} - \frac{1}{2} \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i &= \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} - \frac{1}{2} \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i; \text{ e} \\ \boldsymbol{\mu}'_i \sum^{-1} \left(\mathbf{y} - \frac{1}{2} \boldsymbol{\mu}_i \right) &= \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} - \frac{1}{2} \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i. \end{aligned}$$

Se \mathbf{a}_i for determinado por $\sum^{-1} \boldsymbol{\mu}_i$, então:

$$\begin{aligned} \mathbf{a}'_i \left(\mathbf{y} - \frac{1}{2} \boldsymbol{\mu}_i \right) &= \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} - \frac{1}{2} \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i; \text{ e} \\ \mathbf{a}'_i \mathbf{y} - \frac{1}{2} \mathbf{a}'_i \boldsymbol{\mu}_i &= \boldsymbol{\mu}'_i \sum^{-1} \mathbf{y} - \frac{1}{2} \boldsymbol{\mu}'_i \sum^{-1} \boldsymbol{\mu}_i. \end{aligned}$$

A observação multivariada \mathbf{y} será classificada em x_i , para o qual o valor de:

$$\mathbf{a}'_i \mathbf{y} - \frac{1}{2} \mathbf{a}'_i \boldsymbol{\mu}_i$$

seja o máximo em relação a todos os possíveis valores de $i = 1, 2, \dots, t$.

As *FDS* foram desenvolvidas supondo o conhecimento dos parâmetros populacionais. Porém, não é comum ter a informação real sobre esses parâmetros e, frequentemente, é necessário obtê-los de amostras. Assim, cada média populacional μ_i pode ser substituída pela sua respectiva estimativa \bar{y}_i .

$$\bar{\mathbf{y}}_i = [\bar{y}_{i1} \quad \bar{y}_{i2} \quad \cdots \quad \bar{y}_{ip}].$$

E a estimativa da variância global por:

$$\mathbf{S} = \frac{\sum_{i=1}^t (n_i - 1) \mathbf{S}_i}{\sum_{i=1}^t (n_i - 1)}, \text{ em que:}$$

$$\mathbf{S}_i = \begin{bmatrix} S_{i11} & S_{i12} & \cdots & S_{i1p} \\ S_{i21} & S_{i22} & \cdots & S_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{ip1} & S_{ip2} & \cdots & S_{ipp} \end{bmatrix}.$$

Se \mathbf{a}_i for determinado por $\mathbf{S}^{-1} \bar{\mathbf{y}}_i$, então, a observação multivariada \mathbf{y} será classificada em x_i , para o qual o valor de $\hat{\mathbf{a}}_i' \mathbf{y} - \frac{1}{2} \hat{\mathbf{a}}_i' \bar{\mathbf{y}}_i$ seja o máximo em relação a todos os possíveis valores de $i = 1, 2, \dots, t$.

3.3. Delineamentos Experimentais

Os delineamentos experimentais são as formas em que os tratamentos são dispostos às unidades experimentais ou parcelas. Cada delineamento experimental possui um modelo estatístico específico que o descreve e que é baseado nos princípios básicos da repetição, da casualização e, mas não necessariamente, do controle local (MONTGOMERY, 2012).

A diferença básica entre os modelos será dada pelas possíveis fontes de variações sistemáticas (perturbadoras) ou não intencionais, mas de natureza conhecida e que podem ser controladas pelo pesquisador como, por exemplo, a heterogeneidade de uma área em um experimento agrônômico. Nesses casos, os modelos estatísticos são arquitetados com a finalidade de remover, do erro experimental, as variações sistemáticas e, portanto, considerando-as como fontes de variações adicionais.

O modelo estatístico mais simples que pode ser escolhido é o do DIC, que considera apenas a repetição e a casualização em sua formulação, de modo que:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \text{ em que:}$$

y_{ij} = valor observado da variável-resposta Y no tratamento x_i ($i = 1, 2, \dots, t$) e na repetição j ($j = 1, 2, \dots, r$);

μ = média geral populacional de Y ;

τ_i = efeito do tratamento x_i ; e

ε_{ij} = erro experimental associado a y_{ij} que segue $N(0, \sigma^2)$.

Quando apenas uma fonte de perturbação está presente, o modelo estatístico que descreve essa condição é o do DBC, como segue:

$$y_{ij} = \mu + \tau_i + \omega_j + \varepsilon_{ij}, \text{ em que:}$$

y_{ij} = valor observado da variável-resposta Y no tratamento x_i ($i = 1, 2, \dots, t$) e no bloco b_j ($j = 1, 2, \dots, r$);

μ = média geral populacional da Y ;

τ_i = efeito do tratamento x_i ;

ω_j = efeito do bloco b_j ; e

ε_{ij} = erro experimental associado a y_{ij} que segue $N(0, \sigma^2)$.

Quando os tratamentos são distribuídos às repetições de duas formas distintas, de modo que se possam eliminar os efeitos de duas fontes de variação – denominadas linhas e colunas – do erro experimental, tem-se o DQL. Assim, na ocorrência de t tratamentos, um quadrado com t linhas e t colunas será formado, contendo t^2 parcelas, cujo modelo estatístico é, tal que:

$$y_{(i)jk} = \mu + \tau_i + \omega_j + \gamma_k + \varepsilon_{(i)jk}, \text{ em que:}$$

$y_{(i)jk}$ = valor observado da variável-resposta Y no tratamento x_i ($i = 1, 2, \dots, t$), na linha b_j ($j = 1, 2, \dots, t$) e na coluna c_k ($k = 1, 2, \dots, t$);

μ = média geral populacional de Y ;

τ_i = efeito do tratamento x_i ;

ω_j = efeito da linha b_j ;

γ_k = efeito da coluna c_k ; e

$\varepsilon_{(i)jk}$ = erro experimental associado a $y_{(i)jk}$ que segue $N(0, \sigma^2)$.

3.4. Software Estatístico

Os procedimentos para obtenção das VCs são incomuns nos softwares estatísticos. O que se consegue, muitas vezes, é obter as matrizes \mathbf{H} e \mathbf{E} de acordo com a MANOVA. No entanto, eles são frequentes para as FDS , incluindo aqueles que as utilizam como métodos de classificação. Na atualidade, o software R é um dos mais proeminentes ambientes computacionais e, também, uma linguagem de programação utilizada na manipulação, análise e visualização gráfica de dados (R CORE TEAM, 2020). O software R tem progressivamente ganhado expressividade no mundo, especialmente, na academia

para o uso em análises de dados de pesquisas científicas (TIPPMANN, 2015; MUENCHEN, 2019). A instalação básica do R disponibiliza uma biblioteca com diferentes pacotes, os quais contém diversas funções com seus respectivos argumentos. No R, um recurso de ajuda (*help*) é disponibilizado para cada função e contém normalmente a definição das funções, seus argumentos e alguns exemplos de aplicação.

A função *manova*, por exemplo, utilizada para análise de variância multivariada, faz parte do pacote *stats* (R CORE TEAM, 2020) o qual é carregado automaticamente na inicialização do R. A função *manova* possui os seguintes principais argumentos:

- *formula*: representa o modelo estatístico multivariado que é inserido na função. Nesse caso, as variáveis dependentes entram como uma matriz à esquerda do sinal “~” e as variáveis independentes entram à direita deste mesmo sinal na forma de uma soma. Portanto, é possível definir na função *manova* diferentes modelos estatísticos tais como DIC, DBC e DQL; e,
- *data*: matriz que fornece os dados.

Já a função *eigen* computa os autovalores e autovetores de matrizes numéricas ou complexas e faz parte do pacote *base* do R (R CORE TEAM, 2020). Os argumentos da função *eigen* são:

- *x*: matriz quadrada numérica ou complexa cuja decomposição espectral deve ser calculada (as matrizes lógicas são convertidas em numéricas);
- *symmetric*: se verdadeiro (*TRUE*), define a matriz que está sendo considerada como simétrica (ou Hermitiana se complexa) e, então, apenas seu triângulo inferior com a diagonal incluída da matriz será utilizado. Nesse caso, se este argumento não for definido, a função *isSymmetric* será utilizada em *x*; e,
- *only.values*: se verdadeiro (*TRUE*), apenas os autovalores são calculados e retornados, caso contrário, tanto os autovalores quanto os autovetores são retornados.

A função *lda* realiza a análise discriminante e faz parte do pacote *MASS* (VENABLES & RIPLEY, 2002) do R, que não é carregado automaticamente na inicialização do R e, portanto, requer abertura para uso de suas funções. Os argumentos da função *lda* são:

- *formula*: a fórmula dos grupos na forma $\sim x_1 + x_2 + \dots$. Portanto, o fator de agrupamento fica do lado esquerdo do sinal “~” e do lado direito são especificadas as variáveis discriminadoras. Neste caso, não é possível inserir na função fatores que representem variações além daquela que representa o agrupamento de

tratamentos; no caso de experimentos com repetições, a função *lda* não permite o uso de delineamentos com fatores perturbadores tais como DBC e DQL; e,

- *data*: define o banco de dados de onde as variáveis especificadas na fórmula são preferencialmente retiradas.

4. MATERIAL E MÉTODOS

4.1. Simulação dos Dados

A simulação dos dados foi realizada de modo que as análises propostas nesse trabalho pudessem ser realizadas e os métodos comparáveis. Isso foi feito considerando a simulação dos erros experimentais $\varepsilon_1, \varepsilon_2, \varepsilon_3$ e ε_4 das variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , respectivamente, avaliadas em nove tratamentos (x_1, x_2, \dots, x_9) com nove repetições cada.

Para que as pressuposições, tanto da regra de decisão da análise discriminante linear como da MANOVA fossem atendidas, os erros experimentais $\varepsilon_1, \varepsilon_2, \varepsilon_3$ e ε_4 foram simulados sob observância da multinormalidade, da independência e da homogeneidade das matrizes populacionais de variâncias e covariâncias residuais que ocorrem dentro de cada um dos nove tratamentos considerados ($\Sigma_1 = \Sigma_2 = \dots = \Sigma_9 = \Sigma$). Assim sendo, tem-se que:

$$\varepsilon_1, \varepsilon_2, \varepsilon_3 \text{ e } \varepsilon_4 \sim N_4(\mathbf{0}, \mathbf{M}).$$

Nesse caso, $\mathbf{0}$ e Σ são, respectivamente, o vetor populacional das médias e a matriz populacional comum das variâncias e covariâncias dos erros experimentais $\varepsilon_1, \varepsilon_2, \varepsilon_3$ e ε_4 , de acordo com os nove tratamentos com nove repetições cada, tal que:

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ e } \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}.$$

Para a simulação dos erros experimentais $\varepsilon_1, \varepsilon_2, \varepsilon_3$ e ε_4 , foram consideradas as seguintes correlações populacionais: $\rho_{12} = \rho_{34} = 0,9$ e $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = 0$. Desse modo, a covariância populacional entre os erros experimentais ε_w e $\varepsilon_{w'}$ ($\sigma_{ww'}$), para $w \neq w'$ ($w, w' = 1, 2, 3$ e 4), foi obtida por:

$$\sigma_{ww'} = \rho_{ww'} \sqrt{\sigma_{ww} \times \sigma_{w'w'}}, \text{ em que:}$$

$\rho_{ww'}$ = correlação populacional entre os erros experimentais ε_w e $\varepsilon_{w'}$, para $w \neq w'$ ($w, w' = 1, 2, 3$ e 4);

σ_{ww} = variância populacional do erro experimental ε_w ; e

$\sigma_{w'w'}$ = variância populacional do erro experimental $\varepsilon_{w'}$;

Por outro lado, a variância do erro experimental ε_w (σ_{ww}), para $w = 1, 2, 3$ e 4 , foi definida a fim de conferir um coeficiente de variação (CV_w) igual a 5%, como segue:

$$\sigma_{ww} = \left(\frac{5\mu_w}{100}\right)^2, \text{ em que:}$$

μ_w = média geral populacional da variável-resposta Y_w .

No presente trabalho, o vetor de médias gerais populacionais das variáveis-resposta Y_1, Y_2, Y_3 e Y_4 foi definido por:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 = 60 \\ \mu_2 = 60 \\ \mu_3 = 180 \\ \mu_4 = 180 \end{bmatrix}.$$

Portanto, tem-se:

$$\sigma_{11} = \sigma_{22} = \left(\frac{5 \times 60}{100}\right)^2 = 9; \text{ e}$$

$$\sigma_{33} = \sigma_{44} = \left(\frac{5 \times 180}{100}\right)^2 = 81.$$

Desse modo, as covariâncias populacionais entre os erros experimentais $\varepsilon_1, \varepsilon_2, \varepsilon_3$ e ε_4 , foram as seguintes:

$$\sigma_{12} = \sigma_{21} = 0,9 \times \sqrt{9 \times 9} = 8,1;$$

$$\sigma_{13} = \sigma_{14} = \sigma_{23} = \sigma_{24} = \sigma_{31} = \sigma_{41} = \sigma_{32} = \sigma_{42} = 0; \text{ e}$$

$$\sigma_{34} = \sigma_{43} = 0,9 \times \sqrt{81 \times 81} = 72,9.$$

Assim sendo, tem-se:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 9 & 8,1 & 0 & 0 \\ 8,1 & 9 & 0 & 0 \\ 0 & 0 & 81 & 72,9 \\ 0 & 0 & 72,9 & 81 \end{bmatrix}.$$

Com as definições da matriz $\boldsymbol{\Sigma}$ e do vetor de médias $\mathbf{0}$, uma única matriz (81×4) contendo os resíduos experimentais inerentes aos valores observados nos delineamentos DIC, DBC e DQL, foi obtida utilizando a função *mvnorm* do pacote *MASS* (VENABLES & RIPLEY, 2002) como apresentado no *script* a seguir:

```
library(MASS)

# Definindo a matriz sigma
sigma <- matrix(c(9, 8.1, 0, 0,
                 8.1, 9, 0, 0,
                 0, 0, 81, 72.9,
                 0, 0, 72.9, 81), nrow = 4, ncol = 4)

# Definindo o vetor nulo de médias
m0 <- c(0, 0, 0, 0)
# Simulando os resíduos experimentais
erros <- mvnorm(n = 81, m0, sigma)
```

4.2. Delineamento Inteiramente Casualizado

Os valores observados de cada uma das variáveis-resposta Y_1, Y_2, Y_3 e Y_4 de um experimento simuladamente realizado sob o DIC foram obtidos de acordo com o seguinte modelo:

$$y_{ijw} = \mu_w + \tau_{i_w} + e_{ijw}, \text{ em que:}$$

y_{ijw} = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$) e na repetição j ($j = 1, 2, \dots, 9$);

μ_w = média geral de Y_w ;

τ_{i_w} = efeito do tratamento x_i para Y_w ; e

e_{ijw} = resíduo associado a y_{ijw} para Y_w .

Assim, o vetor de efeitos $\tau_i = [\tau_{i_1} \ \tau_{i_2} \ \tau_{i_3} \ \tau_{i_4}]$ do tratamento x_i ($i = 1, 2, \dots, 9$) para o conjunto das quatro variáveis-resposta Y_1, Y_2, Y_3 e Y_4 foi definido como:

$$\tau_1 = \tau_2 = [-15 \ -15 \ -45 \ -45];$$

$$\tau_3 = \tau_4 = [-15 \ -15 \ 45 \ 45];$$

$$\tau_5 = [0 \ 0 \ 0 \ 0];$$

$$\tau_6 = \tau_7 = [15 \ 15 \ -45 \ -45]; \text{ e}$$

$$\tau_8 = \tau_9 = [15 \ 15 \ 45 \ 45].$$

E o vetor de médias geral como:

$$\mu = \begin{bmatrix} \mu_{w=1} = 60 \\ \mu_{w=2} = 60 \\ \mu_{w=3} = 180 \\ \mu_{w=4} = 180 \end{bmatrix}$$

O procedimento para a construção da tabela de dados com os valores observados das variáveis sob o DIC foi realizado utilizando funções do pacote *tidyverse* (WICKHAM et al., 2019) de acordo com o *script* a seguir:

```
library(tidyverse)

# Adicionando o vetor de médias aos erros
erros_mu <- erros + matrix(c(60, 60, 180, 180), nrow = 81, ncol = 4, byrow = TRUE)

# Definindo os efeitos dos tratamentos para cada uma das quatro variáveis-resposta
t1 <- t2 <- c(-15, -15, -45, -45)
t3 <- t4 <- c(-15, -15, 45, 45)
t5 <- c( 0, 0, 0, 0)
t6 <- t7 <- c( 15, 15, -45, -45)
t8 <- t9 <- c( 15, 15, 45, 45)

# Efeitos dos tratamentos
ef_trats <- rbind(t1, t2, t3, t4, t5, t6, t7, t8, t9) %>%
  as_tibble() %>%
  add_column(trat = as.factor(seq(1:9)), .before = TRUE)
```

```
# A função %>% canaliza um objeto para uma função ou expressão. A função rbind
toma uma sequência de vetores e os combina pelas linhas. A função as_tibble
transforma o banco de dados em um tipo chamado tibble. A função add_column
adiciona colunas em um banco de dados.
```

```
# Adicionando a identificação dos tratamentos
erros_mu.id <- erros_mu %>%
  as_tibble() %>%
  add_column(
    trat = as.factor(rep(1:9, each = 9)),
    .before = TRUE)

# Somando os efeitos dos tratamentos ao quadro erros_mu
dados.dic <- erros_mu.id %>%
  gather(vars, valor_erro_mu, -c(trat)) %>%
  left_join(ef_trats %>%
    gather(vars, valor_trat, -trat)) %>%
  mutate(soma = valor_erro_mu + valor_trat) %>%
  select(-c(valor_erro_mu, valor_trat)) %>%
  pivot_wider(names_from = vars, values_from = soma) %>%
  unchop(everything()) %>%
  rename(y1 = V1, y2 = V2, y3 = V3, y4 = V4)
```

função gather reúne colunas em pares de valores-chave. A função left_join(x, y) adiciona colunas de y em x (incluindo todas as linhas em x), coincidindo as linhas com base nas chaves. A função mutate adiciona novas variáveis. # A função select seleciona variáveis em um banco de dados. A função pivot_wider "amplia" os dados, aumentando o número de colunas e diminuindo o número de linhas. # a função unchop torna o banco de dados mais longo ao expandir as colunas da lista de forma que cada elemento da coluna da lista tenha sua própria linha na saída. # função everything seleciona todas as variáveis. # a função rename renomeia individualmente as variáveis em um banco de dados.

4.2.1. Variáveis Canônicas

Para obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ VCs associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , os seguintes procedimentos foram realizados. Primeiro, as matrizes H e E foram obtidas da MANOVA sob o DIC com o auxílio da função *manova* do software estatístico R (Tabela 4.1) de acordo com o *script* a seguir:

```
# MANOVA dos dados provenientes do DIC
mod <- manova(cbind(y1, y2, y3, y4) ~ trat, data = dados.dic)

# Obtendo as matrizes H (trat), E (erro), e os graus de liberdade do residuo
(nE) e encontrando S
H <- summary(mod)$SS$trat
E <- summary(mod)$SS$Residuals
nE <- mod$df.residual
S <- E/nE
```

Tabela 4.1 - Análise de variância multivariada sob o DIC

FV	GL	SQ e SP	QM e PM
Trat	8	$H = \begin{bmatrix} SQTrat_1 & SPTrat_{12} & SPTrat_{13} & SPTrat_{14} \\ SPTrat_{12} & SQTrat_2 & SPTrat_{23} & SPTrat_{24} \\ SPTrat_{13} & SPTrat_{23} & SQTrat_3 & SPTrat_{34} \\ SPTrat_{14} & SPTrat_{24} & SPTrat_{34} & SQTrat_4 \end{bmatrix}$	-
Res	72	$E = \begin{bmatrix} SQRes_1 & SPRes_{12} & SPRes_{13} & SPRes_{14} \\ SPRes_{12} & SQRes_2 & SPRes_{23} & SPRes_{24} \\ SPRes_{13} & SPRes_{23} & SQRes_3 & SPRes_{34} \\ SPRes_{14} & SPRes_{24} & SPRes_{34} & SQRes_4 \end{bmatrix}$	$S = \begin{bmatrix} QMRes_1 & PMRes_{12} & PMRes_{13} & PMRes_{14} \\ PMRes_{12} & QMRes_2 & PMRes_{23} & PMRes_{24} \\ PMRes_{13} & PMRes_{23} & QMRes_3 & PMRes_{34} \\ PMRes_{14} & PMRes_{24} & PMRes_{34} & QMRes_4 \end{bmatrix}$

Depois, a matriz $E^{-1}H$ foi gerada para obtenção dos quatro autovalores e dos quatro respectivos autovetores. No mesmo software, a função *eigen* ($E^{-1}H$) retorna a seguinte solução do sistema de equações homogêneas:

$$(E^{-1}H - \hat{\lambda}_m I)\hat{v}_m = \mathbf{0}, \text{ para } m = 1, 2, 3 \text{ e } 4, \text{ em que:}$$

H = matriz das somas de quadrados e de produtos dos tratamentos (4×4);

E = matriz das somas de quadrados e de produtos do resíduo (4×4);

$\hat{\lambda}_m$ = estimativa do autovalor de ordem m ($m = 1, 2, 3$ e 4) da matriz $E^{-1}H$;

$\hat{v}_m = [\hat{v}_{m1} \quad \hat{v}_{m2} \quad \hat{v}_{m3} \quad \hat{v}_{m4}]$ = estimativa do autovetor normalizado de ordem m ($m = 1, 2, 3$ e 4) associado à estimativa do autovalor $\hat{\lambda}_m$ da matriz $E^{-1}H$; e

$\mathbf{0}$ = vetor nulo (4×1).

O autovetor normalizado e estimado \hat{v}_m (4×1) de ordem m é de tal forma que:

$$\hat{v}_m' \hat{v}_m = 1 \text{ e } \hat{v}_m' \hat{v}_{m'} = 0 \text{ para } m \neq m' \text{ (} m, m' = 1, 2, 3 \text{ e } 4 \text{), e,}$$

pode ser obtido de acordo com o *script* a seguir:

```
# Invertendo a matriz E e multiplicando pela H
E_inv_H <- solve (E) %*% H

# Obtendo os autovalores e autovetores de E_inv_H
auto_val_vet <- eigen(E_inv_H)

# Conferindo a padronização v'v
auto_vet_v <- auto_val_vet$vector
round(t(auto_vet_v)%*%auto_vet_v,2)
```

A normalização $\hat{v}_m' \hat{v}_m = 1$ não corrige apropriadamente as diferenças de escala das variáveis-resposta Ys para obtenção das VCs e, portanto, ela deve ser reescalada. Dessa forma, a obtenção das quatro VCs foi realizada adotando-se outros s tipos de estimativas dos autovetores padronizados como recomendado por Mardia et al. (1997), como seguem:

$$\hat{g}_m = [\hat{g}_{m1} \quad \hat{g}_{m2} \quad \hat{g}_{m3} \quad \hat{g}_{m4}]; \text{ e}$$

$$\hat{e}_m = [\hat{e}_{m1} \quad \hat{e}_{m2} \quad \hat{e}_{m3} \quad \hat{e}_{m4}], \text{ em que:}$$

$$\hat{g}_m' E \hat{g}_m = 1 \text{ e } \hat{g}_m' E \hat{g}_{m'} = 0, \text{ para } m \neq m' \text{ (} m, m' = 1, 2, 3 \text{ e } 4 \text{);}$$

$$\hat{e}_m' S \hat{e}_m = 1 \text{ e } \hat{e}_m' S \hat{e}_{m'} = 0, \text{ para } m \neq m' \text{ (} m, m' = 1, 2, 3 \text{ e } 4 \text{);}$$

$$S = \frac{E}{n_E} = \text{matriz de variâncias e covariâncias residuais; e}$$

n_E = número de graus de liberdade do resíduo da MANOVA.

Ambas as padronizações podem ser obtidas, respectivamente, com as seguintes relações algébricas: $\hat{g}_m = \frac{1}{\sqrt{\hat{v}_m' E \hat{v}_m}} \hat{v}_m$ e $\hat{e}_m = \hat{g}_m \times \sqrt{n_E}$ para $m = 1, 2, 3$ e 4 , e conforme o *script* a seguir:

```

# Corrigindo a padronização dos autovetores obtidos pela função eigen para g'Eg
escala_g <- diag(1/sqrt((t(auto_vet_v) %*% E %*% auto_vet_v)))
auto_vet_g <- auto_vet_v %*% diag(escala_g)

# Conferindo a padronização g'Eg
round(t(auto_vet_g) %*% E %*% auto_vet_g, 2)

# Para utilizar a padronização e'Se = 1, basta fazer:
auto_vet_e <- auto_vet_g * sqrt(nE)

# Conferindo a padronização e'Se
round(t(auto_vet_e) %*% S %*% auto_vet_e, 2)

```

Dessa forma, as quatro *VCs* padronizadas foram obtidas conforme a Tabela 4.2:

Tabela 4.2 - Variáveis canônicas padronizadas

VCs	Padronização 1	Padronização 2
\widehat{VC}_1	$\hat{g}_{11}Y_1 + \hat{g}_{12}Y_2 + \hat{g}_{13}Y_3 + \hat{g}_{14}Y_4$	$\hat{e}_{11}Y_1 + \hat{e}_{12}Y_2 + \hat{e}_{13}Y_3 + \hat{e}_{14}Y_4$
\widehat{VC}_2	$\hat{g}_{21}Y_1 + \hat{g}_{22}Y_2 + \hat{g}_{23}Y_3 + \hat{g}_{24}Y_4$	$\hat{e}_{21}Y_1 + \hat{e}_{22}Y_2 + \hat{e}_{23}Y_3 + \hat{e}_{24}Y_4$
\widehat{VC}_3	$\hat{g}_{31}Y_1 + \hat{g}_{32}Y_2 + \hat{g}_{33}Y_3 + \hat{g}_{34}Y_4$	$\hat{e}_{31}Y_1 + \hat{e}_{32}Y_2 + \hat{e}_{33}Y_3 + \hat{e}_{34}Y_4$
\widehat{VC}_4	$\hat{g}_{41}Y_1 + \hat{g}_{42}Y_2 + \hat{g}_{43}Y_3 + \hat{g}_{44}Y_4$	$\hat{e}_{41}Y_1 + \hat{e}_{42}Y_2 + \hat{e}_{43}Y_3 + \hat{e}_{44}Y_4$

Os escores das *VCs* padronizadas foram encontrados com base nas seguintes fórmulas:

$$z_{ijw_m} = y_{ijw} \times \hat{g}_{m_w}; \text{ ou,}$$

$$f_{ijw_m} = y_{ijw} \times \hat{e}_{m_w}; \text{ em que:}$$

z_{ijw_m} = escore do tratamento x_i ($i = 1, 2, \dots, 9$) e na repetição j ($j = 1, 2, \dots, 9$) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da variável canônica \widehat{VC}_m ($m = 1, 2, 3$ e 4) pela padronização 1;

f_{ijw_m} = escore 2 do tratamento x_i ($i = 1, 2, \dots, 9$) e na repetição j ($j = 1, 2, \dots, 9$) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da variável canônica \widehat{VC}_m ($m = 1, 2, 3$ e 4) pela padronização 2;

y_{ijw} = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$) e na repetição j ($j = 1, 2, \dots, 9$);

\hat{g}_{m_w} = coeficiente da \widehat{VC}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 1.

\hat{e}_{m_w} = coeficiente da \widehat{VC}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 2.

E, conforme o *script* a seguir:

```
# Obtendo os escores das VCs utilizando os vetores "e" (adaptar para os
vetores "g")
escores_vc_e <- as.matrix(dados.dic[,c(2:5)])*%auto_vet_e

# Identificação dos escores com os tratamentos
escores_vc_e.id <- escores_vc_e %>%
  as_tibble() %>%
  rename(vc = 1:4) %>%
  add_column(dados.dic[,1], .before = TRUE)
```

A função *eigen* fornece os autovalores $\hat{\lambda}_m$ correspondentes a cada VC quando os autovetores \hat{v}_m ($m = 1, 2, 3$ e 4) são estimados, porém, eles também podem ser encontrados pela relação entre a soma de quadrados de tratamentos ($SQ_{Trat.}$) e soma de quadrados dos resíduos ($SQ_{Res.}$) obtidas da ANOVA de cada variável canônica (MANLY & ALBERTO, 2016) de acordo com o seguinte:

$$\hat{\lambda}_m = \frac{SQ_{Trat.m}}{SQ_{Res.m}}; \text{ em que:}$$

$\hat{\lambda}_m$ = autovalor associado a \widehat{VC}_m ($m = 1, 2, 3$ e 4);

$SQ_{Trat.m}$ = soma de quadrados de tratamento associada a \widehat{VC}_m ($m = 1, 2, 3$ e 4);

$SQ_{Res.m}$ = soma de quadrados do resíduo associada a \widehat{VC}_m ($m = 1, 2, 3$ e 4).

O *script* a seguir demonstra a obtenção dos autovalores a partir dos escores de cada variável canônica \widehat{VC}_m ($m = 1, 2, 3$ e 4) no DIC:

```
# ANOVA dos escores obtidos pela padronização e' Se (adaptar para g'Eg)
vars <- names(escores_vc_e.id[2:5])
fit_anova_e <- map(paste(vars, '~', 'trat'), ~ anova(lm(., escores_vc_e.id)))

# Encontrando os autovalores pela divisão da SQTrat pela SQRes (adaptar para
g'Eg)
names(fit_anova_e) <- vars
auto_val_e <- map_df(fit_anova_e, broom::tidy, .id = 'vars') %>%
  select(vars, term, sumsq) %>%
  pivot_wider(names_from = term, values_from = c(sumsq)) %>%
  mutate(av = trat/Residuals) %>%
  select(-c(trat, Residuals))
auto_val_e
```

4.2.2. Funções Discriminantes

Nessa análise executada por meio da função *lda* do software estatístico R, foram obtidas $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ FDS associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , por meio da solução do seguinte sistema de equações homogêneas:

$$(\widehat{B}_\mu - \hat{\lambda}_m \mathbf{S}) \hat{a}_m = \mathbf{0}, \text{ para } m = 1, 2, 3 \text{ e } 4, \text{ em que:}$$

$$\widehat{B}_\mu = \sum_{i=1}^9 (\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})';$$

$\bar{y}_i = [\bar{y}_{1i} \ \bar{y}_{2i} \ \bar{y}_{3i} \ \bar{y}_{4i}]$ = estimativa do vetor de médias das variáveis-resposta Y_1, Y_2, Y_3 e Y_4 do tratamento x_i ($i = 1, 2, \dots, 9$);

$\bar{y} = [\bar{y}_1 \ \bar{y}_2 \ \bar{y}_3 \ \bar{y}_4] =$ estimativa do vetor de médias gerais das variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4 ;

$\hat{\lambda}_m =$ estimativa do autovalor de ordem m ($m = 1, 2, 3$ e 4);

$\hat{\mathbf{a}}_m =$ estimativa do autovetor não normalizado de ordem m ($m = 1, 2, 3$ e 4);

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix};$$

$s_{ww} =$ estimativa da variância do erro experimental ε_w ;

$s_{ww'} =$ estimativa da covariância entre os erros experimentais ε_w e $\varepsilon_{w'}$, para $w \neq w'$ ($w, w' = 1, 2, 3$ e 4); e

$\mathbf{0} =$ vetor nulo (4×1).

Desse modo, as quatro *FDS* foram obtidas por:

$$\widehat{FD}_1 = \hat{a}_{11}Y_1 + \hat{a}_{12}Y_2 + \hat{a}_{13}Y_3 + \hat{a}_{14}Y_4;$$

$$\widehat{FD}_2 = \hat{a}_{21}Y_1 + \hat{a}_{22}Y_2 + \hat{a}_{23}Y_3 + \hat{a}_{24}Y_4;$$

$$\widehat{FD}_3 = \hat{a}_{31}Y_1 + \hat{a}_{32}Y_2 + \hat{a}_{33}Y_3 + \hat{a}_{34}Y_4; \text{ e}$$

$$\widehat{FD}_4 = \hat{a}_{41}Y_1 + \hat{a}_{42}Y_2 + \hat{a}_{43}Y_3 + \hat{a}_{44}Y_4.$$

No entanto, as quatro *FDS* podem não ser informativas, caso as variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4 não sejam proporcionais, isto é, não apresentem as mesmas escalas e as mesmas variações. Por isso, os coeficientes do vetor $\hat{\mathbf{a}}_m$ foram padronizados conforme o seguinte:

$$\hat{\mathbf{e}}_m = [\hat{e}_{m1} \ \hat{e}_{m2} \ \hat{e}_{m3} \ \hat{e}_{m4}], \text{ em que:}$$

$$\hat{\mathbf{e}}'_m \mathbf{S} \hat{\mathbf{e}}_m = 1 \text{ e } \hat{\mathbf{e}}'_m \mathbf{S} \hat{\mathbf{e}}_{m'} = 0, \text{ para } m \neq m' \text{ (} m, m' = 1, 2, 3 \text{ e } 4 \text{);}$$

$$\mathbf{S} = \mathbf{E}/(n - t); \text{ e}$$

$n =$ número de observações; e,

$t =$ número de tratamentos.

E, podem ser obtidos pelo *script* a seguir:

```
library(MASS)
fdlf <- lda(trat ~ y1 + y2 + y3 + y4, data = dados.dic)

# Encontrando os autovetores
auto_vet_e <- fdlf$scaling
```

Esta padronização do tipo 2 é a que a função *lda* do software estatístico R produz, portanto, as quatro *FDS* associadas às variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4 , apresentaram os

seus coeficientes conforme mencionados pelo segundo tipo de padronização ($\hat{\mathbf{e}}_m$) como mostra a Tabela 4.3.

Tabela 4.3 - Funções discriminantes lineares de Fisher padronizadas

\widehat{FD}_m	Padronização 2
\widehat{FD}_1	$\hat{e}_{11}Y_1 + \hat{e}_{12}Y_2 + \hat{e}_{13}Y_3 + \hat{e}_{14}Y_4$
\widehat{FD}_2	$\hat{e}_{21}Y_1 + \hat{e}_{22}Y_2 + \hat{e}_{23}Y_3 + \hat{e}_{24}Y_4$
\widehat{FD}_3	$\hat{e}_{31}Y_1 + \hat{e}_{32}Y_2 + \hat{e}_{33}Y_3 + \hat{e}_{34}Y_4$
\widehat{FD}_4	$\hat{e}_{41}Y_1 + \hat{e}_{42}Y_2 + \hat{e}_{43}Y_3 + \hat{e}_{44}Y_4$

A função *lda* não fornece as matrizes de variâncias e covariâncias que representam a variação entre ($\widehat{\mathbf{B}}_\mu$) e dentro (\mathbf{S}) dos tratamentos; a última é necessária para verificar qual padronização é utilizada pela função *lda*. Ambas as matrizes podem ser encontradas pelos seguintes passos:

```
# Encontrando a matriz S
# Passo 1: encontrar os resíduos da variação dentro dos tratamentos
desvios.dentro <-
  dados.dic %>%
  group_by(trat) %>%
  mutate_all(funs(. - mean(.))) %>%
  ungroup %>%
  select(-trat) %>%
  as.matrix

# A função group_by agrupa um banco de dados pelos valores das variáveis
# selecionadas para que operações possam ser realizadas por grupo. A função
# mutate_all aplica uma mesma operação em todas as variáveis. A função ungroup
# desagrupa o banco de dados anteriormente agrupado.

# Passo 2: matriz de soma de quadrados da variação dentro dos tratamentos
E <- t(desvios.dentro) %*% desvios.dentro

# Passo 3: matriz de variâncias e covariâncias da variação dentro dos tratamentos
# (S)
n_obs <- nrow(dados.dic)
n_trat <- nlevels(dados.dic$trat)
S <- E / (n_obs - n_trat)
# Encontrando a matriz B
# Passo 4: desvios entre tratamentos
desvios.entre <- dados.dic %>%
  group_by(trat) %>%
  summarize_all(mean) %>%
  select(-trat) %>%
  mutate_all(funs(. - mean(.))) %>%
  as.matrix

#A função summarize_all() aplica uma mesma função em todas as variáveis.

# Passo 5: Soma de quadrados de tratamentos
H <- n_trat * t(desvios.entre) %*% desvios.entre
B <- H / (n_trat - 1)
```

```
# Verificando a padronização dos autovetores
round(t(auto_vet_e) %*% S %*% auto_vet_e, 2) # padronização e' Se = 1
```

Os escores das *FDs* padronizadas foram encontrados com base na seguinte fórmula:

$$f_{ijw_m} = y_{ijw} \times \hat{e}_{m_w}; \text{ em que:}$$

f_{ijw_m} = escore do tratamento x_i ($i = 1, 2, \dots, 9$), na repetição j ($j = 1, 2, \dots, 9$), da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da função discriminante \widehat{FD}_m ($m = 1, 2, 3$ e 4) pela padronização 2;

y_{ijw} = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$) e na repetição j ($j = 1, 2, \dots, 9$);

\hat{e}_{m_w} = coeficiente da \widehat{FD}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 2.

```
# Obtendo os escores das FDs utilizando os vetores "e"
escores_fd_e <- as.matrix(dados.dic[,c(2:5)]) %*% auto_vet_e
```

```
# Identificação dos escores com os tratamentos
escores_fd_e.id <- escores_fd_e %>%
  as_tibble() %>%
  rename(vc = 1:4) %>%
  add_column(dados.dic[,1], .before = TRUE)
```

A função *lda* não produz os autovalores ($\hat{\lambda}_m$) correspondentes a cada função discriminante \widehat{FD}_m ($m = 1, 2, 3$ e 4), mas eles podem ser obtidos de acordo com o seguinte (MANLY & ALBERTO, 2016):

$$\hat{\lambda}_m = \frac{SQ_{Trat.m}}{SQ_{Res.m}}; \text{ em que:}$$

$\hat{\lambda}_m$ = autovalor associado a \widehat{FD}_m ($m = 1, 2, 3$ e 4);

$SQ_{Trat.m}$ = soma de quadrado de tratamento associada a \widehat{FD}_m ($m = 1, 2, 3$ e 4);

$SQ_{Res.m}$ = soma de quadrado do resíduo associada a \widehat{FD}_m ($m = 1, 2, 3$ e 4).

O *script* a seguir demonstra a obtenção dos autovalores:

```
# ANOVA dos escores obtidos pela padronização e' Se
vars <- names(escores_fd_e.id[2:5])
fit_anova_e <- map(paste(vars, '~', 'trat'), ~ anova(lm(., escores_fd_e.id)))

# Encontrando os autovalores pela divisão da SQTrat pela SQRes
names(fit_anova_e) <- vars
auto_val_e <- map_df(fit_anova_e, broom::tidy, .id = 'vars') %>%
  select(vars, term, sumsq) %>%
  pivot_wider(names_from = term, values_from = c(sumsq)) %>%
  mutate(av = trat/Residuals) %>%
  select(-c(trat, Residuals))
auto_val_e
```

4.3. Delineamento em Blocos Casualizados

Os valores observados de cada uma das variáveis-resposta Y_1, Y_2, Y_3 e Y_4 de um experimento simuladamente realizado sob o DBC foram obtidos de acordo com o seguinte modelo:

$$y_{ijw} = \mu_w + \tau_{i_w} + \omega_{j_w} + e_{ijw}, \text{ em que:}$$

y_{ijw} = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$) e no bloco b_j ($j = 1, 2, \dots, 9$);

μ_w = média geral de Y_w ;

τ_{i_w} = efeito do tratamento x_i para Y_w ;

ω_{j_w} = efeito do bloco b_j para Y_w ; e

e_{ijw} = resíduo associado a y_{ijw} para Y_w .

Adotando o mesmo vetor de efeitos de tratamentos e o mesmo vetor de média geral utilizado para o experimento realizado sob o DIC, e definindo o vetor de efeitos dos blocos $\omega_j = [\omega_{j_1} \ \omega_{j_2} \ \omega_{j_3} \ \omega_{j_4}]$ ($j = 1, 2, \dots, 9$) para o conjunto das quatro variáveis-resposta Y_1, Y_2, Y_3 e Y_4 como:

$$\omega_1 = \omega_2 = [-10 \ -10 \ -30 \ -30];$$

$$\omega_3 = \omega_4 = [-5 \ -5 \ -15 \ -15];$$

$$\omega_5 = [0 \ 0 \ 0 \ 0];$$

$$\omega_6 = \omega_7 = [5 \ 5 \ 15 \ 15]; \text{ e}$$

$$\omega_8 = \omega_9 = [10 \ 10 \ 30 \ 30].$$

Os valores observados das variáveis do experimento realizado sob o DBC foram obtidos utilizando funções do pacote *tidyverse* (WICKHAM et al., 2019) de acordo com o *script* a seguir:

```
library(tidyverse)

# Adicionando o vetor de médias aos erros
erros_mu <- erros + matrix(c(60, 60, 180, 180), nrow = 81, ncol = 4, byrow =
TRUE)

# Os efeitos de tratamentos para cada uma das quatro variáveis-resposta já foram
definidos no item 4.2.

# Efeitos dos blocos para cada uma das quatro variáveis-resposta
w1 <- w2 <- c(-10, -10, -30, -30)
w3 <- w4 <- c(-5, -5, -15, -15)
w5 <- c(0, 0, 0, 0)
w6 <- w7 <- c(5, 5, 15, 15)
w8 <- w9 <- c(10, 10, 30, 30)

# Efeitos dos tratamentos já especificado no item 4.2.
```

```

# Quadro com os efeitos dos blocos
ef_blocs <- rbind(w1, w2, w3, w4, w5, w6, w7, w8, w9) %>%
  as_tibble() %>%
  add_column(bloc = as.factor(seq(1:9)),
             .before = TRUE)

# Criando o quadro de dados para o DBC com os valores das variáveis-resposta

# Adicionando a identificação dos tratamentos e blocos
erros_mu.id <- erros_mu %>%
  as_tibble() %>%
  add_column(trat = as.factor(rep(1:9, each = 9)),
             .before = TRUE) %>%
  add_column(bloc = as.factor(rep(1:9, 9)),
             .after = 'trat')

# Somando os efeitos dos tratamentos e dos blocos ao quadro erros_mu
dados.dbc <- erros_mu.id %>%
  gather(vars, valor_erro_mu, -c(bloc, trat)) %>%
  left_join(ef_trats %>%
            gather(vars, valor_trat, -trat)) %>%
  left_join(ef_blocs %>%
            gather(vars, valor_bloc, -bloc)) %>%
  mutate(soma = valor_erro_mu + valor_trat + valor_bloc) %>%
  select(-c(valor_erro_mu, valor_trat, valor_bloc)) %>%
  pivot_wider(names_from = vars, values_from = soma) %>%
  unchop(everything()) %>%
  rename(y1 = V1, y2 = V2, y3 = V3, y4 = V4)

```

4.3.1. Variáveis Canônicas

Para obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ VCs associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , foram realizados, do mesmo modo que no DIC, os seguintes procedimentos. Primeiro, as matrizes \mathbf{H} e \mathbf{E} foram obtidas da MANOVA sob o DBC utilizando a função *manova* do software estatístico R (Tabela 4.4) de acordo com o *script* a seguir:

```

# Realizando a MANOVA de dados provenientes do DBC
mod <- manova(cbind(y1, y2, y3, y4) ~ bloc + trat, data = dados.dbc)

# Os demais procedimentos seguem os passos do script como apresentado no item
4.2.1.

```

Tabela 4.4 - Análise de variância multivariada sob o DBC

FV	GL	SQ e SP	QM e PM
Bloco	8	—	—
Tratamento	8	H	—
Resíduo	64	E	S

Depois, a matriz $\mathbf{E}^{-1}\mathbf{H}$ foi gerada para obtenção dos quatro autovalores e dos quatro respectivos autovetores por meio da função *eigen* do software estatístico R. Os autovetores foram, então, reescalados para a padronização 1 ($\hat{\mathbf{g}}_m' \mathbf{E} \hat{\mathbf{g}}_m = 1$) e 2

($\hat{e}'_m S \hat{e}_m = 1$) utilizando, respectivamente, as seguintes relações algébricas: $\hat{g}_m = \frac{1}{\sqrt{\hat{v}'_m E \hat{v}_m}} \hat{v}_m$ e $\hat{e}_m = \hat{g}_m \times \sqrt{n_E}$ ($m = 1, 2, 3$ e 4) para a obtenção das VCs conforme a Tabela 4.2 e como descrito no item 4.2.1.

Os escores das VCs padronizadas foram obtidos com base na seguinte fórmula:

$$z_{ijw_m} = y_{ijw} \times \hat{g}_{m_w}; \text{ ou,}$$

$$f_{ijw_m} = y_{ijw} \times \hat{e}_{m_w}; \text{ em que:}$$

z_{ijw_m} = escore do tratamento x_i ($i = 1, 2, \dots, 9$), no bloco b_j ($j = 1, 2, \dots, 9$) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da variável canônica \widehat{VC}_m ($m = 1, 2, 3$ e 4) pela padronização 1;

f_{ijw_m} = escore 2 do tratamento x_i ($i = 1, 2, \dots, 9$), no bloco b_j ($j = 1, 2, \dots, 9$) da variável-resposta Y_w ($w = 1, 2, 3$ e 4), da variável canônica \widehat{VC}_m ($m = 1, 2, 3$ e 4) pela padronização 2;

y_{ijw} = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$) e no bloco b_j ($j = 1, 2, \dots, 9$);

\hat{g}_{m_w} = coeficiente da \widehat{VC}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 1.

\hat{e}_{m_w} = coeficiente da \widehat{VC}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 2.

E, conforme o *script* a seguir:

```
# Obtendo os escores das VCs utilizando os vetores "e" (adaptar para os vetores
#g")
escores_vc_e <- as.matrix(dados.dbc[,c(3:6)])%*%auto_vet_e

# Identificação dos escores com os tratamentos
escores_vc_e.id <- escores_vc_e %>%
  as_tibble() %>%
  rename(vc = 1:4) %>%
  add_column(dados.dbc[,1:2], .before = TRUE)
```

A função *eigen* fornece os autovalores ($\hat{\lambda}_m$) correspondentes a cada variável canônica (\widehat{VC}_m) quando os autovetores \hat{v}_m ($m = 1, 2, 3$ e 4) são estimados, porém, eles também podem ser encontrados de acordo com o seguinte (MANLY & ALBERTO, 2016):

$$\hat{\lambda}_m = \frac{SQ_{Trat.m}}{SQ_{Res.m}}; \text{ em que:}$$

$\hat{\lambda}_m$ = autovalor associado a \widehat{VC}_m ($m = 1, 2, 3$ e 4);

$SQ_{Trat.m}$ = soma de quadrados de tratamento associada a \widehat{VC}_m ($m = 1, 2, 3$ e 4);

$SQ_{Res.m}$ = soma de quadrados do resíduo associada a \widehat{VC}_m ($m = 1, 2, 3$ e 4).

O *script* a seguir demonstra a obtenção dos autovalores a partir dos escores da VC_m ($m = 1, 2, 3$ e 4) no DBC:

```
# ANOVA dos escores obtidos pela padronização e' Se (adaptar para g'Eg)
vars <- names(escores_vc_e.id[3:6])
fit_anova_e <- map(paste(vars, '~', 'bloc', '+', 'trat')), ~ anova(lm(.,
escores_vc_e.id))

# Encontrando os autovalores pela divisão da SQTrat pela SQRes (adaptar para
g'Eg)
names(fit_anova_e) <- vars
auto_val_e <- map_df(fit_anova_e, broom::tidy, .id = 'vars') %>%
  select(vars, term, sumsq) %>%
  pivot_wider(names_from = term, values_from = c(sumsq)) %>%
  mutate(av = trat/Residuals) %>%
  select(-c(trat, bloc, Residuals))
auto_val_e
```

4.3.2. Funções Discriminantes

Antes da obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ *FDs* associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , os seus valores observados foram transformados, a fim de apresentarem apenas duas fontes de variação (entre e dentro dos tratamentos). Dessa forma, para a variável-resposta Y_w ($w = 1, 2, 3$ e 4), tem-se:

$$y_{ij_w}^* = y_{ij_w} - \omega_{j_w} = \mu_w + \tau_{i_w} + e_{ij_w}.$$

Isso equivale em transformar o modelo estatístico do DBC no do DIC e, conseqüentemente, torná-lo adequado à utilização da função *lda* do software estatístico R. Portanto, os valores transformados das variáveis-resposta foram obtidos de acordo com o *script* a seguir:

```
# Encontrando os efeitos dos blocos
efeito_bloc <- dados.dbc %>%
  group_by(bloc) %>%
  summarise_if(
    .predicate = ~ is.numeric(),
    .funs = mean) %>%
  mutate_at(
    .vars = vars(matches('^y([1-4])$')),
    .funs = function(x) x - mean(x))

# Removendo os efeitos dos blocos e obtendo os valores transformados para o
DBC
dados.dbc_t <- dados.dbc %>%
  gather(vars, valor_trat, -c(bloc, trat)) %>%
  left_join(
    efeito_bloc %>%
      gather(vars, efeito_b, -bloc)) %>%
  mutate(
    dif = valor_trat - efeito_b) %>%
  select(-c(bloc, valor_trat, efeito_b)) %>%
  pivot_wider(names_from = vars, values_from = dif) %>%
  unchop(everything())
```

E, a obtenção das *FDs* foi conforme a seguinte solução do sistema de equações homogêneas:

$$(\widehat{\mathbf{B}}_{\mu} - \hat{\lambda}_m \mathbf{S}) \widehat{\mathbf{a}}_m = \mathbf{0}, \text{ para } m = 1, 2, 3 \text{ e } 4, \text{ em que:}$$

$$\widehat{\mathbf{B}}_{\mu} = \sum_{i=1}^9 (\bar{\mathbf{y}}_i^* - \bar{\mathbf{y}}^*) (\bar{\mathbf{y}}_i^* - \bar{\mathbf{y}}^*)';$$

$\bar{\mathbf{y}}_i^* = [\bar{y}_{1i}^* \ \bar{y}_{2i}^* \ \bar{y}_{3i}^* \ \bar{y}_{4i}^*]$ = estimativa do vetor de médias das variáveis-resposta transformadas Y_1^* , Y_2^* , Y_3^* e Y_4^* do tratamento x_i ($i = 1, 2, \dots, 9$);

$\bar{\mathbf{y}}^* = [\bar{y}_1^* \ \bar{y}_2^* \ \bar{y}_3^* \ \bar{y}_4^*]$ = estimativa do vetor de médias gerais das variáveis-resposta transformadas Y_1^* , Y_2^* , Y_3^* e Y_4^* ;

$\hat{\lambda}_m$ = estimativa do autovalor de ordem m ($m = 1, 2, 3 \text{ e } 4$);

$\widehat{\mathbf{a}}_m$ = estimativa do autovetor não normalizado de ordem m ($m = 1, 2, 3 \text{ e } 4$);

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix};$$

s_{ww} = estimativa da variância do erro experimental ε_w ;

$s_{ww'}$ = estimativa da covariância entre os erros experimentais ε_w e $\varepsilon_{w'}$, para $w \neq w'$ ($w, w' = 1, 2, 3 \text{ e } 4$); e

$\mathbf{0}$ = vetor nulo (4×1).

Desse modo, as quatro *FDs* foram obtidas por:

$$\widehat{FD}_1 = \hat{a}_{11}Y_1^* + \hat{a}_{12}Y_2^* + \hat{a}_{13}Y_3^* + \hat{a}_{14}Y_4^*;$$

$$\widehat{FD}_2 = \hat{a}_{21}Y_1^* + \hat{a}_{22}Y_2^* + \hat{a}_{23}Y_3^* + \hat{a}_{24}Y_4^*;$$

$$\widehat{FD}_3 = \hat{a}_{31}Y_1^* + \hat{a}_{32}Y_2^* + \hat{a}_{33}Y_3^* + \hat{a}_{34}Y_4^*;$$
 e

$$\widehat{FD}_4 = \hat{a}_{41}Y_1^* + \hat{a}_{42}Y_2^* + \hat{a}_{43}Y_3^* + \hat{a}_{44}Y_4^*.$$

Do mesmo modo como apresentando no item 4.2.2, as quatro *FDs* associadas às variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4 , apresentaram os seus coeficientes conforme a padronização 2 como mostra a Tabela 4.3, sendo obtidos de acordo com o *script* a seguir:

```
library(MASS)
fdlf <- lda(trat ~ y1 + y2 + y3 + y4, data = dados.dbc_t)

# Encontrando os autovetores
auto_vet_e <- fdlf$scaling
```

As matrizes de variâncias e covariâncias que representam a variação entre ($\widehat{\mathbf{B}}_{\mu}$) e dentro (\mathbf{S}) dos tratamentos foram obtidas conforme o *script* apresentado no item 4.2.2, com a ressalva de sua adaptação pela utilização dos valores observados transformados das variáveis do experimento no DBC, ou seja, trocando dados.dic por dados.dbc_t.

Para o DBC, os escores das *FDs* padronizadas foram obtidos utilizando os valores observados originais das variáveis-resposta de modo que o delineamento fosse preservado para as análises posteriores. Dessa forma, os escores foram obtidos com base na seguinte fórmula:

$$f_{ijw_m} = y_{ijw} \times \hat{e}_{m_w}; \text{ em que:}$$

f_{ijw_m} = escore do tratamento x_i ($i = 1, 2, \dots, 9$), no bloco b_j ($j = 1, 2, \dots, 9$), da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da variável canônica \widehat{FD}_m ($m = 1, 2, 3$ e 4) pela padronização 2;

y_{ijw} = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$) e no bloco b_j ($j = 1, 2, \dots, 9$);

\hat{e}_{m_w} = coeficiente da \widehat{FD}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 2.

```
# Obtendo os escores das FDs utilizando os vetores "e"
escores_fd_e <- as.matrix(dados.dbc[,c(3:6)])%%auto_vet_e

# Identificação dos escores com os tratamentos
escores_fd_e.id <- escores_fd_e %>%
  as_tibble() %>%
  rename(vc = 1:4) %>%
  add_column(dados.dbc[,1:2], .before = TRUE)
```

Embora os escores das *FDs* tenham sido obtidos com os valores observados das variáveis originais, eles também podem ser utilizados para obtenção dos autovalores ($\hat{\lambda}_m$) correspondentes a cada função discriminante \widehat{FD}_m a partir da relação entre a $SQ_{Trat.m}$ e $SQ_{Res.m}$ ($m = 1, 2, 3$ e 4) (MANLY & ALBERTO, 2016) como descrito no item 4.4.2 e de acordo com o *script* a seguir:

```
# ANOVA dos escores obtidos pela padronização e' Se (adaptar para g'Eg)
vars <- names(escores_fd_e.id[3:6])
fit_anova_e <- map(paste(vars, '~', 'bloc', '+', 'trat')),
  ~ anova(lm(., escores_fd_e.id))
# Encontrando os autovalores pela divisão SQTrat/SQRes (adaptar para g'Eg)
names(fit_anova_e) <- vars
auto_val_e <- map_df(fit_anova_e, broom::tidy, .id = 'vars') %>%
  select(vars, term, sumsq) %>%
  pivot_wider(names_from = term, values_from = c(sumsq)) %>%
  mutate(av = trat/Residuals) %>%
  select(-c(trat, bloc, Residuals))
auto_val_e
```

4.4. Delineamento em Quadrado Latino

Os valores observados de cada uma das variáveis-resposta Y_1, Y_2, Y_3 e Y_4 de um experimento simuladamente realizado sob o DQL foram obtidos de acordo com o seguinte modelo:

$$y_{(i)jk_w} = \mu_w + \tau_{i_w} + \omega_{j_w} + \gamma_{k_w} + e_{(i)jk_w}, \text{ em que:}$$

$y_{(i)jk_w}$ = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$), na linha b_j ($j = 1, 2, \dots, 9$) e na coluna c_k ($k = 1, 2, \dots, 9$);

μ_w = média geral de Y_w ;

τ_{i_w} = efeito do tratamento x_i para Y_w ;

ω_{j_w} = efeito da linha b_j para Y_w ;

γ_{k_w} = efeito da coluna c_k para Y_w ; e

$e_{(i)jk_w}$ = resíduo associado a $y_{(i)jk_w}$ para Y_w .

Tomando o mesmo vetor de efeitos de tratamentos e de média geral utilizado para o experimento realizado sob o DIC ou DBC, definiu-se também para o caso do experimento realizado sob o DQL, o vetor de efeitos das linhas $\omega_j = [\omega_{j_1} \ \omega_{j_2} \ \omega_{j_3} \ \omega_{j_4}]$ e das colunas $\gamma_k = [\gamma_{k_1} \ \gamma_{k_2} \ \gamma_{k_3} \ \gamma_{k_4}]$ ($j = k = 1, 2, \dots, 9$) para o conjunto das quatro variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , cujos valores foram os seguintes:

$$\omega_1 = \omega_2 = \gamma_1 = \gamma_2 = [-10 \ -10 \ -30 \ -30];$$

$$\omega_3 = \omega_4 = \gamma_3 = \gamma_4 = [-5 \ -5 \ -15 \ -15];$$

$$\omega_5 = \gamma_5 = [0 \ 0 \ 0 \ 0];$$

$$\omega_6 = \omega_7 = \gamma_6 = \gamma_7 = [5 \ 5 \ 15 \ 15]; \text{ e}$$

$$\omega_8 = \omega_9 = \gamma_8 = \gamma_9 = [10 \ 10 \ 30 \ 30].$$

Com isso, o procedimento para a construção do banco de dados com os valores das variáveis observadas do experimento realizado sob o DQL foi feito utilizando funções do pacote *tidyverse* (WICKHAM et al., 2019) conforme o seguinte *script*:

```
library(tidyverse)

# Os efeitos de tratamentos foram definidos no item 4.2

# Os efeitos das linhas e colunas no DQL são os mesmos dos blocos no DBC.

# Efeitos das linhas
ef_lin <- ef_blocs %>%
  rename(linha = bloc)

# Efeitos das colunas
ef_col <- ef_lin %>%
  rename(coluna = linha)
```

```

# Criando o quadro de dados para o DQL com os valores das quatro variáveis-
resposta
trat.lin.col <- agricolae::design.lsd(1:9)$book[,-1] %>%
  select("1:9", row, col) %>%
  rename(
    trat = "1:9",
    linha = row,
    coluna = col)

# A função design.lsd do pacote agricolae estabelece a aleatorização dos
tratamentos nas linhas e colunas de um experimento realizado sob o DQL

trat.lin.col.id <- arrange(trat.lin.col, trat) %>%
  as_tibble()

# A função arrange ordena as linhas de um banco de dados pelos valores da coluna
selecionada.

# Adicionando a identificação dos tratamentos, linhas e colunas ao quadro
erros_mu
erros_mu.id <- erros_mu %>%
  as_tibble() %>%
  add_column(trat.lin.col.id,
             .before = TRUE)

# Somando os efeitos dos tratamentos, das linhas e das colunas ao quadro
erros_mu.id
dados.dql <- erros_mu.id %>%
  gather(vars, valor_erro_mu, -c(coluna, linha, trat)) %>%
  left_join(ef_trats %>%
            gather(vars, valor_trat, -trat)) %>%
  left_join(ef_lin %>%
            gather(vars, valor_lin, -linha)) %>%
  left_join(ef_col %>%
            gather(vars, valor_col, -coluna)) %>%
  mutate(soma = valor_erro_mu + valor_trat + valor_lin + valor_col) %>%
  select(-c(valor_erro_mu, valor_trat, valor_lin, valor_col)) %>%
  pivot_wider(names_from = vars, values_from = soma) %>%
  unchop(everything()) %>%
  rename(y1 = V1, y2 = V2, y3 = V3, y4 = V4)

```

4.4.1. Variáveis Canônicas

Para obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ VCs associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , foram realizados, do mesmo modo que no DIC e no DBC, os seguintes procedimentos. Primeiro, as matrizes H e E foram obtidas da MANOVA sob o DQL com o auxílio da função *manova* do software estatístico R (Tabela 4.5) conforme o *script* a seguir:

```

# Realizando a MANOVA de dados provenientes do DQL
mod <- manova(cbind(y1, y2, y3, y4) ~ linha + coluna + trat, data = dados.dql)

# Os demais procedimentos seguem os passos do script como apresentado no item
4.2.1.

```

Tabela 4.5 - Análise de variância multivariada sob o DQL

FV	GL	SQ e SP	QM e PM
Linha	8	–	–
Coluna	8	–	–
Tratamento	8	H	–
Resíduo	56	E	S

Depois, a matriz $E^{-1}H$ foi gerada para obtenção dos quatro autovalores e dos quatro respectivos autovetores por meio da função *eigen* do software estatístico R. Os autovetores foram, então, reescalados para a padronização 1 ($\hat{g}'_m E \hat{g}_m = 1$) e 2 ($\hat{e}'_m S \hat{e}_m = 1$) utilizando, respectivamente, as seguintes relações algébricas: $\hat{g}_m = \frac{1}{\sqrt{\hat{v}'_m E \hat{v}_m}} \hat{v}_m$ e $\hat{e}_m = \hat{g}_m \times \sqrt{n_E}$ ($m = 1, 2, 3$ e 4) para a obtenção das VCs conforme a Tabela 4.2 e como descrito no item 4.2.1.

Os escores das VCs padronizadas foram obtidos com base na seguinte fórmula:

$$z_{(i)jk_{wm}} = y_{(i)jk_w} \times \hat{g}_{m_w}; \text{ ou,}$$

$$f_{(i)jk_{wm}} = y_{(i)jk_w} \times \hat{e}_{m_w}; \text{ em que:}$$

$z_{(i)jk_{wm}}$ = escore do tratamento x_i ($i = 1, 2, \dots, 9$), na linha b_j ($j = 1, 2, \dots, 9$), na coluna c_k ($k = 1, 2, \dots, 9$), da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da variável canônica \widehat{VC}_m ($m = 1, 2, 3$ e 4) pela padronização 1;

$f_{(i)jk_{wm}}$ = escore do tratamento x_i ($i = 1, 2, \dots, 9$), na linha b_j ($j = 1, 2, \dots, 9$), na coluna c_k ($k = 1, 2, \dots, 9$), da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da variável canônica VC_m ($m = 1, 2, 3$ e 4) pela padronização 2;

$y_{(i)jk_w}$ = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$), na linha b_j ($j = 1, 2, \dots, 9$) e na coluna c_k ($k = 1, 2, \dots, 9$);

\hat{g}_{m_w} = coeficiente da \widehat{VC}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 1.

\hat{e}_{m_w} = coeficiente da \widehat{VC}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 2.

E, conforme o *script* a seguir:

```
# Obtendo os escores das VCs utilizando os vetores "e" (adaptar para os vetores "g")
escores_vc_e <- as.matrix(dados.dql[,c(4:7)])%*%auto_vet_e
```

```
# Identificação dos escores com os tratamentos
escores_vc_e.id <- escores_vc_e %>%
  as_tibble() %>%
  rename(vc = 1:4) %>%
  add_column(dados.dql[,1:3], .before = TRUE)
```

A função *eigen* fornece os autovalores ($\hat{\lambda}_m$) correspondentes a cada variável canônica (\widehat{VC}_m) quando os autovetores \hat{v}_m ($m = 1, 2, 3$ e 4) são estimados, porém, eles também podem ser encontrados de acordo com o seguinte:

$$\hat{\lambda}_m = \frac{SQ_{Trat.m}}{SQ_{Res.m}}; \text{ em que:}$$

$\hat{\lambda}_m$ = autovalor associado \widehat{VC}_m ($m = 1, 2, 3$ e 4);

$SQ_{Trat.m}$ = soma de quadrados de tratamento associada a \widehat{VC}_m ($m = 1, 2, 3$ e 4);

$SQ_{Res.m}$ = soma de quadrados do resíduo associada a \widehat{VC}_m ($m = 1, 2, 3$ e 4).

O *script* a seguir demonstra a obtenção dos autovalores a partir dos escores da \widehat{VC}_m ($m = 1, 2, 3$ e 4) no DQL:

```
# ANOVA dos escores obtidos pela padronização e' Se (adaptar para g' Eg)
vars <- names(escores_vc_e.id[4:7])
fit_anova_e <- map(paste(vars, '~', 'linha', '+', 'coluna', '+', 'trat')), ~
anova(lm(., escores_vc_e.id))

# Encontrando os autovalores pela divisão da SQTrat pela SQRes (adaptar para
g' Eg)
names(fit_anova_e) <- vars
auto_val_e <- map_df(fit_anova_e, broom::tidy, .id = 'vars') %>%
  select(vars, term, sumsq) %>%
  pivot_wider(names_from = term, values_from = c(sumsq)) %>%
  mutate(av = trat/Residuals) %>%
  select(-c(trat, linha, coluna, Residuals))
auto_val_e
```

4.4.2. Funções Discriminantes

Do mesmo modo que no DBC, antes da obtenção das $s = \text{mín. } (p = 4, t - 1 = 8) = 4$ *FDs* associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 , os seus valores observados foram transformados, a fim de apresentarem apenas duas fontes de variação (entre e dentro dos tratamentos). Dessa forma, para a variável-resposta Y_w ($w = 1, 2, 3$ e 4), tem-se:

$$y_{(i)jk_w}^* = y_{(i)jk_w} - \omega_{j_w} - \gamma_{k_w} = \mu_w + \tau_{i_w} + e_{(i)jk_w}.$$

Isso equivale em transformar o modelo estatístico do DQL no do DIC e, conseqüentemente, torná-lo adequado à utilização da função *lda* do software estatístico R. Portanto, os valores transformados das variáveis-resposta foram obtidos de acordo com o *script* a seguir:

```

# Encontrando os efeitos das linhas
efeito_linha <- dados.dql %>%
  group_by(linha) %>%
  summarise_if(
    .predicate = ~ is.numeric(),
    .funs = mean) %>%
  mutate_at(
    .vars = vars(matches('^y([1-4])$')),
    .funs = function(x) x - mean(x))

# Encontrando os efeitos das colunas
efeito_coluna <- dados.dql %>%
  group_by(coluna) %>%
  summarise_if(
    .predicate = ~ is.numeric(),
    .funs = mean) %>%
  mutate_at(
    .vars = vars(matches('^y([1-4])$')),
    .funs = function(x) x - mean(x))

# Removendo o efeito das linhas e colunas e obtendo os novos valores para o DQL
dados.dql_t <- dados.dql %>%
  gather(vars, valor_trat, -c(linha, coluna, trat)) %>%
  left_join(
    efeito_linha %>%
      gather(vars, efeito_l, -linha)) %>%
  left_join(
    efeito_coluna %>%
      gather(vars, efeito_c, -coluna)) %>%
  mutate(
    dif = valor_trat - (efeito_l + efeito_c)) %>%
  select(-c(linha, coluna, valor_trat, efeito_l, efeito_c)) %>%
  pivot_wider(names_from = vars, values_from = dif) %>%
  unchop(everything())

```

E, as *FDs* foram obtidas pela solução do sistema de equações homogêneas da mesma forma que foram obtidas no DIC e no DBC. Além disso, do mesmo modo como apresentando no item 4.2.2, as quatro *FDs* associadas às variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4 apresentaram os seus coeficientes conforme a padronização 2 como mostra a Tabela 4.3, sendo obtidos de acordo com o *script* a seguir:

```

library(MASS)
fdlf <- lda(trat ~ y1 + y2 + y3 + y4, data = dados.dql_t)

# Encontrando os autovetores
auto_vet_e <- fdlf$scaling

```

As matrizes de variâncias e covariâncias que representam a variação entre ($\hat{\mathbf{B}}_{\mu}$) e dentro (\mathbf{S}) dos tratamentos foram obtidas conforme o *script* apresentado no item 4.2.2, com a ressalva de sua adaptação pela utilização dos valores observados transformados das variáveis do experimento no DQL, ou seja, trocando dados.dic por dados.dql_t.

```

library(MASS)

# Aplicando a função lda()
fdlf <- lda(trat ~ y1 + y2 + y3 + y4, data = dados.dql_t)

```

```
# Encontrando os autovetores
auto_vet_e <- fdlf$scaling
```

Para o DQL, os escores das *FDs* padronizadas foram obtidos utilizando os valores observados originais das variáveis-resposta de modo que o delineamento fosse preservado para as análises posteriores, assim, como foi feito no DBC. Dessa forma, os escores foram obtidos com base na seguinte fórmula:

$$f_{(i)jk_w m} = y_{(i)jk_w} \times \hat{e}_{m_w}; \text{ em que:}$$

$f_{(i)jk_w m}$ = escore do tratamento x_i ($i = 1, 2, \dots, 9$), na linha b_j ($j = 1, 2, \dots, 9$), na coluna c_k ($k = 1, 2, \dots, 9$), da variável-resposta Y_w ($w = 1, 2, 3$ e 4) e da variável canônica \widehat{FD}_m ($m = 1, 2, 3$ e 4) pela padronização 2;

$y_{(i)jk_w}$ = valor observado da variável-resposta Y_w ($w = 1, 2, 3$ e 4) no tratamento x_i ($i = 1, 2, \dots, 9$), na linha b_j ($j = 1, 2, \dots, 9$) e na coluna c_k ($k = 1, 2, \dots, 9$);

\hat{e}_{m_w} = coeficiente da \widehat{FD}_m ($m = 1, 2, 3$ e 4) da variável-resposta Y_w ($w = 1, 2, 3$ e 4) pela padronização 2.

```
# Obtendo os escores das FDs utilizando os vetores "e"
escores_fd_e <- as.matrix(dados.dql[,c(4:7)])*%auto_vet_e
```

```
# Identificação dos escores com os tratamentos
escores_fd_e.id <- escores_fd_e %>%
  as_tibble() %>%
  rename(vc = 1:4) %>%
  add_column(dados.dql[,1:3], .before = TRUE)
```

Embora os escores das *FDs* tenham sido obtidos com os valores observados originais das variáveis-resposta, eles também podem ser utilizados para obtenção dos autovalores ($\hat{\lambda}_m$) correspondentes a cada função discriminante (\widehat{FD}_m) a partir da relação entre a $SQ_{Trat.m}$ e a $SQ_{Res.m}$ ($m = 1, 2, 3$ e 4) (MANLY & ALBERTO, 2016) como descrito no item 4.4.2 e de acordo com o *script* a seguir:

```
# ANOVA dos escores obtidos pela padronização e' Se (adaptar para g'Eg)
vars <- names(escores_fd_e.id[4:7])
fit_anova_e <- map(paste(vars, '~', 'linha', '+', 'coluna', '+', 'trat')), ~
anova(lm(., escores_fd_e.id))

# Encontrando os autovalores pela divisão da SQTrat pela SQRes (adaptar para g'Eg)
names(fit_anova_e) <- vars
auto_val_e <- map_df(fit_anova_e, broom::tidy, .id = 'vars') %>%
  select(vars, term, sumsq) %>%
  pivot_wider(names_from = term, values_from = c(sumsq)) %>%
  mutate(av = trat/Residuals) %>%
  select(-c(trat, bloc, Residuals))
auto_val_e
```

4.5. Medidas de Avaliação

Inicialmente, a multinormalidade dos resíduos e igualdade das matrizes de variâncias e covariâncias residuais foram testadas, respectivamente, pelo teste de Mardia (MARDIA, 1970) e o teste M-Box (BOX, 1949). O teste de Mardia foi executado pela função *mvn* do pacote *mvn* e combina os resultados dos testes de assimetria e curtose multivariados ao mesmo tempo para inferir sobre a normalidade multivariada; se ambos os testes indicarem normalidade multivariada, ou seja, p -valor $> 0,05$, então, os dados seguem uma distribuição de normalidade multivariada ao nível de significância de 0,05 (KORKMAZ et al., 2014; MARDIA, 1970). Já o teste M-Box foi executado pela função *box_m* do pacote *rstatix*; se p -valor $> 0,05$, então, as matrizes de variâncias e covariâncias residuais são iguais entre os $t = 9$ tratamentos ao nível de significância de 0,05 (BOX, 1949; KASSAMBARA, 2020). Dadas satisfeitas as pressuposições, as quatro VCs e as quatro FDs foram obtidas e comparadas por diferentes indicadores como medidas de desempenho de ambas as combinações lineares, tais como: autovetores, autovalores, importância relativa (IR), correlações entre as funções lineares e as variáveis originais, análises de variância (ANOVA) e gráficos bidimensionais. Cada medida de avaliação ou análise foi realizada de acordo com cada delineamento experimental e padronização utilizada, no entanto, isso foi realizado na maioria das vezes apenas para as duas primeiras funções lineares de acordo com o teste de dimensionalidade de Bartlett (BARTLETT, 1941) e da representatividade pelo acúmulo de IR maior que 70%.

5. RESULTADOS

5.1. Simulação dos Dados

A simulação dos erros experimentais ($\varepsilon_1, \varepsilon_2, \varepsilon_3$ e ε_4) associados às variáveis respostas Y_1, Y_2, Y_3 e Y_4 , foi procedida de acordo com o item 4.1. e resultou em uma matriz de erros (81×4) com vetor de médias e matriz de variâncias e covariâncias, como se segue:

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} -0,20 \\ -0,28 \\ -0,42 \\ -0,23 \end{bmatrix} \quad \hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 8,9 & 8,1 & -0,7 & -0,6 \\ 8,1 & 9,2 & -1,6 & -1,1 \\ -0,7 & -1,6 & 80,4 & 72,9 \\ -0,6 & -1,1 & 72,9 & 81,9 \end{bmatrix}$$

E, a partir dos resíduos e_1, e_2, e_3 e e_4 associados às variáveis respostas Y_1, Y_2, Y_3 e Y_4 , os valores observados, considerando o modelo estatístico de cada delineamento

experimental, foram obtidos pela adição da média geral, dos efeitos de tratamentos e dos efeitos perturbadores quando estes últimos estavam presentes.

5.2. Delineamento Inteiramente Casualizado

5.2.1. Variáveis Canônicas

Os valores observados considerando um experimento multivariado simulado realizado sob o DIC foram obtidos e os resíduos da MANOVA apresentaram multinormalidade (*Assimetria*: $p > 0,05$; *Curtose*: $p > 0,05$) e igualdade das matrizes de covariâncias ($p > 0,05$), respectivamente, de acordo com o teste de Mardia e o teste M-Box. As matrizes **H** e **E** (Tabela 5.1) foram obtidas da análise de variância multivariada para posterior obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ VCs associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 .

Tabela 5.1 - Resultado parcial da análise de variância multivariada sob o DIC

FV	GL	SQ e SP				QM e PM
		H				
Trat	8	$= \begin{bmatrix} 16860,32 & 16851,96 & -2036,69 & -843,22 \\ 16851,96 & 16852,74 & -2192,76 & -996,36 \\ -2036,69 & -2192,76 & 141742,45 & 139671,19 \\ -843,22 & -996,36 & 139671,19 & 137750,94 \end{bmatrix}$				-
Res	72	$E = \begin{bmatrix} 675,69 & 616,05 & -102,65 & -121,31 \\ 616,05 & 697,39 & -164,10 & -156,84 \\ -102,65 & -164,10 & 6045,65 & 5578,45 \\ -121,31 & -156,84 & 5578,45 & 6279,07 \end{bmatrix}$				$S = \begin{bmatrix} 9,38 & 8,56 & -1,43 & -1,68 \\ 8,56 & 9,69 & -2,28 & -2,18 \\ -1,43 & -2,28 & 83,97 & 77,48 \\ -1,68 & -2,18 & 77,48 & 87,21 \end{bmatrix}$

Os autovetores (**v**) obtidos a partir da aplicação da função *eigen* do produto das matrizes $E^{-1}H$ foram reescalados e as VCs foram obtidas de acordo com as padronizações 1 ($g'Eg = 1$) e 2 ($e'Se = 1$). Os escores das VCs padronizadas foram utilizados para encontrar os autovalores e, conseqüentemente, a importância relativa de cada variável canônica foi também encontrada. A Tabela 5.2 apresenta os coeficientes para cada padronização bem como os autovalores e a importância relativa (IR) associada a cada variável canônica (\widehat{VC}_m , para $m = 1, 2, 3$ e 4). O teste de dimensionalidade de Bartlett indicou a \widehat{VC}_1 e \widehat{VC}_2 como suficientes para explicar a maior parte da variação total (primeiro valor da estatística do teste não significativo: $D_2^2 = 7,04 < \chi_{[(4-2)(9-1-2)]}^2 = 21,03$); estas duas variáveis explicam juntas quase 100% da variação total (Tabela 5.2).

Tabela 5.2 - Variáveis canônicas padronizadas obtidas posteriormente à MANOVA de um experimento instalado sob o DIC com quatro variáveis-resposta Y_1, Y_2, Y_3 e Y_4

\widehat{VC}_m	Padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)				Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)				$\hat{\lambda}_m$	IR (%)
	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4		
\widehat{VC}_1	0,020	0,017	0,001	0,003	0,173	0,145	0,013	0,028	26,351	52,47
\widehat{VC}_2	0,011	0,001	-0,009	-0,003	0,093	0,006	-0,075	-0,028	23,775	47,34
\widehat{VC}_3	0,084	-0,084	-0,001	-0,001	0,716	-0,717	-0,010	0,010	0,064	0,13
\widehat{VC}_4	0,001	-0,003	-0,029	0,029	0,006	-0,024	-0,246	0,249	0,035	0,07

As correlações das duas primeiras VCs com as variáveis-resposta originais Y_1, Y_2, Y_3 e Y_4 do experimento instalado sob o DIC também foram calculadas e são demonstradas na Tabela 5.3. Para ambas as padronizações, as correlações foram as mesmas.

Tabela 5.3 - Correlações obtidas entre as duas primeiras variáveis canônicas (VCs) padronizadas e as variáveis-resposta originais Y_1, Y_2, Y_3 e Y_4 de um experimento instalado sob o DIC

\widehat{VC}_m	Padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)				Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)			
	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
\widehat{VC}_1	0,94	0,94	0,30	0,33	0,94	0,94	0,30	0,33
\widehat{VC}_2	0,34	0,35	-0,95	-0,94	0,34	0,35	-0,95	-0,94

A análise de variância (ANOVA) das VCs obtidas do experimento instalado sob o DIC foi também procedida e a Tabela 5.4 demonstra os resultados para as estimativas das \widehat{VC}_1 e \widehat{VC}_2 . É importante destacar que as VCs na padronização 1 apresentaram $SQ_{Res.} = 1$ e o $QM_{Res.} = 0,01$, porém, na padronização 2 apresentaram a $SQ_{Res.} = 72$ e o $QM_{Res.} = 1$. Apesar disso, o valor da estatística F é o mesmo entre as duas padronizações para mesma VC .

Tabela 5.4 - Análise de variância (ANOVA) das variáveis canônicas padronizadas (VCs) obtidas posteriormente a MANOVA de um experimento instalado sob o DIC

Padronização	Fonte de variação	GL	\widehat{VC}_1			\widehat{VC}_2		
			SQ	QM	F	SQ	QM	F
1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)	Tratamento	8	26,3	3,29	237,2	23,78	2,97	214,0
	Resíduo	72	1,00	0,01		1,00	0,01	
2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)	Tratamento	8	1897,3	237,16	237,2	1711,8	213,97	214,0
	Resíduo	72	72,0	1,00		72,0	1,00	

Resultados na tabela estão arredondados.

A Figura 5.1 apresenta a média dos escores por tratamento em função da \widehat{VC}_1 e \widehat{VC}_2 pela padronização 1 (Figura 5.1 A) e 2 (Figura 5.1 B).

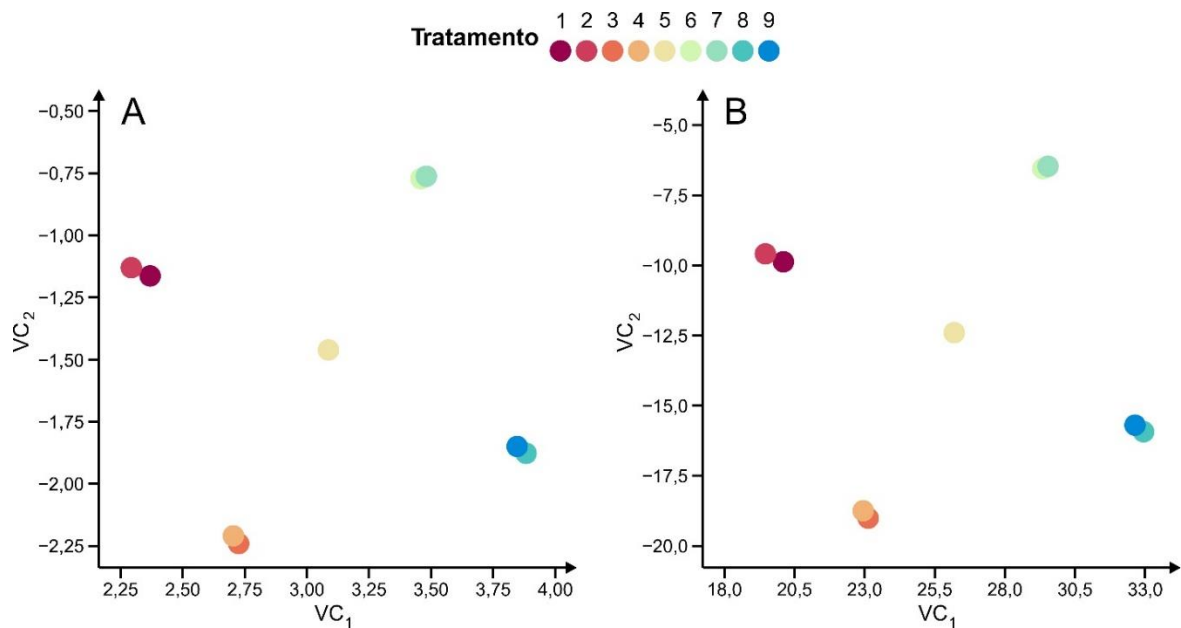


Figura 5.1 - Média dos escores por tratamento em função da \widehat{VC}_1 e \widehat{VC}_2 pela padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$) (A) e 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$) (B) no DIC.

5.2.2. Funções Discriminantes

Os valores observados do experimento multivariado hipotético sob o DIC utilizados para obtenção das $s = \min.(p = 4, t - 1 = 8) =$ funções discriminantes (FDs) foram os mesmos utilizados para obtenção das VCs posteriormente à MANOVA já que, neste caso, não é necessário a transformação dos dados. Sendo assim, como descrito anteriormente, os resíduos apresentaram multinormalidade e igualdade de matriz de covariâncias entre os tratamentos. A função *lda* do pacote *MASS* não informa as matrizes que representam a variação entre e dentro dos tratamentos, contudo, elas puderam ser obtidas de acordo com o apresentado no item 4.2.2. Dessa forma, tem-se a matriz $\widehat{\mathbf{B}}_{\mu}$ que representa a variação entre os tratamentos:

$$\widehat{\mathbf{B}}_{\mu} = \begin{bmatrix} 2107,5 & 2106,5 & -254,6 & -105,4 \\ 2106,5 & 2106,6 & -274,1 & -124,5 \\ -254,6 & -274,1 & 17717,8 & 17458,9 \\ -105,4 & -124,6 & 17458,9 & 17218,9 \end{bmatrix}$$

E, a matriz \mathbf{S} , que representa a variação dentro dos tratamentos:

$$\mathbf{S} = \begin{bmatrix} 9,38 & 8,56 & -1,43 & -1,68 \\ 8,56 & 9,69 & -2,28 & -2,18 \\ -1,43 & -2,28 & 83,97 & 77,48 \\ -1,68 & -2,18 & 77,48 & 87,21 \end{bmatrix}$$

Para o DIC, ambas as matrizes $\widehat{\mathbf{B}}_{\mu}$ e \mathbf{S} são idênticas àquelas que compõem as matrizes de quadrados médios e produtos médios, respectivamente, do tratamento e do resíduo da MANOVA.

Os escores das *FDs* padronizadas foram utilizados para encontrar os autovalores e, conseqüentemente, a importância relativa de cada função discriminante também foi encontrada. A Tabela 5.5 apresenta os coeficientes para cada padronização 2 bem como os autovalores e a importância relativa (IR) (que é, em termos absolutos, idêntica ao valor da proporção do traço que é gerada pela função *lda*) associada a cada função discriminante (\widehat{FD}_m , para $m = 1, 2, 3$ e 4).

O teste de dimensionalidade de Bartlett indicou a \widehat{FD}_1 e \widehat{FD}_2 como suficientes para explicar a maior parte da variação total (primeiro valor da estatística do teste não significativo: $D_2^2 = 7,04 < \chi_{[(4-2)(9-1-2)]}^2 = 21,03$); estas duas variáveis explicam juntas quase 100% da variação total (Tabela 5.5).

Tabela 5.5 - Funções discriminantes (*FDs*) padronizadas obtidas de um experimento instalado sob o DIC

\widehat{FD}_m	Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)				$\hat{\lambda}_m$	IR (%)	Proporção do traço
	Y_1	Y_2	Y_3	Y_4			
\widehat{FD}_1	-0,173	-0,145	-0,013	-0,028	26,351	52,47	0,5247
\widehat{FD}_2	-0,093	-0,006	0,075	0,028	23,775	47,34	0,4734
\widehat{FD}_3	0,716	-0,717	-0,010	0,010	0,064	0,13	0,0013
\widehat{FD}_4	0,006	-0,024	-0,246	0,249	0,035	0,07	0,0007

As correlações das duas primeiras *FDs* com as variáveis-resposta originais Y_1 , Y_2 , Y_3 e Y_4 apresentaram, em valor absoluto, os mesmos valores que as correlações obtidas com as duas primeiras *VCs* (Tabela 5.3). Além disso, as análises de variância das *FDs* obtidas do experimento instalado sob o DIC apresentaram o mesmo resultado que aquelas procedidas para *VCs* no mesmo delineamento. Os mesmos valores para as somas de quadrados, quadrados médios e do F calculado foram obtidos como mostra a Tabela 5.4. A disposição gráfica das médias dos escores por tratamento também foi a mesma como apresentado na Figura 5.1 considerando os mesmos sinais dos coeficientes para *FDs*.

5.3. Delineamento em Blocos Casualizados

5.3.1. Variáveis Canônicas

Os valores observados considerando um experimento multivariado simulado realizado sob o DBC foram obtidos e os resíduos apresentaram multinormalidade (*Assimetria*: $p > 0,05$; *Curtose*: $p > 0,05$) e igualdade das matrizes de covariâncias ($P > 0,05$), respectivamente, de acordo com o teste de Mardia e o teste M-Box. As matrizes \mathbf{H} e \mathbf{E} (Tabela 5.6) foram obtidas da análise de variância multivariada para posterior obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ VCs associadas às variáveis-resposta Y_1, Y_2, Y_3 e Y_4 .

Tabela 5.6 - Resultado parcial da análise de variância multivariada (MANOVA) de um experimento hipoteticamente instalado sob o DBC com quatro variáveis-resposta Y_1, Y_2, Y_3 e Y_4

FV	GL	SQ e SP	QM e PM
Bloco	8	–	–
Trat	8	$\mathbf{H} = \begin{bmatrix} 16860,32 & 16851,96 & -2036,69 & -843,22 \\ 16851,96 & 16852,74 & -2192,76 & -996,36 \\ -2036,69 & -2192,76 & 141742,45 & 139671,19 \\ -843,22 & -996,36 & 139671,19 & 137750,94 \end{bmatrix}$	–
Res	64	$\mathbf{E} = \begin{bmatrix} 562,83 & 553,04 & -72,88 & -43,95 \\ 553,04 & 650,67 & -143,22 & -79,25 \\ -72,88 & -143,22 & 5562,76 & 5223,29 \\ -43,95 & -79,25 & 5223,29 & 5858,91 \end{bmatrix}$	$\mathbf{s} = \begin{bmatrix} 8,79 & 8,64 & -1,14 & -0,69 \\ 8,64 & 10,17 & -2,24 & -1,24 \\ -1,14 & -2,24 & 86,92 & 81,61 \\ -0,69 & -1,24 & 81,61 & 91,55 \end{bmatrix}$

Os autovetores (\mathbf{v}) obtidos a partir da aplicação da função *eigen* do produto das matrizes $\mathbf{E}^{-1}\mathbf{H}$ foram reescalados e as VCs foram obtidas de acordo com as padronizações 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$) e 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$). Os escores das VCs padronizadas foram utilizados para encontrar os autovalores e, conseqüentemente, a importância relativa de cada VC também foi encontrada. A Tabela 5.7 apresenta os coeficientes para cada padronização bem como os autovalores e a importância relativa (IR) associada a cada variável canônica (\widehat{VC}_m , para $m = 1, 2, 3$ e 4).

O teste de dimensionalidade de Bartlett indicou a \widehat{VC}_1 e \widehat{VC}_2 como suficientes para explicar a maior parte da variação total (primeiro valor da estatística do teste não significativo: $D_2^2 = 7,97 < \chi_{[(4-2)(9-1-2)]}^2 = 21,03$); estas duas variáveis explicam juntas quase 100% da variação total (Tabela 5.7).

Tabela 5.7 - Variáveis canônicas padronizadas obtidas posteriormente a MANOVA de um experimento instalado sob o DBC com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4

\widehat{VC}_m	Padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)				Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)				$\hat{\lambda}_m$	IR (%)
	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4		
\widehat{VC}_1	0,0382	0,0040	0,0002	0,0005	0,305	0,032	0,002	0,004	30,017	53,53
\widehat{VC}_2	0,0076	-0,0070	-0,0102	-0,0034	0,060	-0,056	-0,082	-0,027	25,937	46,25
\widehat{VC}_3	0,0968	-0,0971	-0,0037	0,0036	0,775	-0,777	-0,029	0,029	0,084	0,15
\widehat{VC}_4	-0,0026	-0,0004	-0,0317	0,0322	-0,021	-0,003	-0,244	0,257	0,042	0,07

As correlações das duas primeiras VCs com as variáveis-resposta originais Y_1 , Y_2 , Y_3 e Y_4 do experimento instalado sob o DBC também foram calculadas e são demonstradas na Tabela 5.8. Para ambas as padronizações, as correlações foram as mesmas.

Tabela 5.8 - Correlação obtida entre as variáveis canônicas (VCs) padronizadas obtidas posteriormente a MANOVA e as variáveis-resposta de um experimento instalado sob o DBC

\widehat{VC}_m	Padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)				Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)			
	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
\widehat{VC}_1	1,00	1,00	0,23	0,25	1,00	1,00	0,23	0,25
\widehat{VC}_2	-0,17	0,17	-1,00	-1,00	-0,17	0,17	-1,00	-1,00

A análise de variância (ANOVA) das VCs obtidas do experimento instalado sob o DBC foi também procedida e a Tabela 5.9 demonstra os resultados para as estimativas das \widehat{VC}_1 e \widehat{VC}_2 . É importante destacar que as VCs na padronização 1 apresentaram $SQ_{Res.} = 1$ e o $QM_{Res.} = 0,016$, porém, na padronização 2, elas apresentaram a $SQ_{Res.} = 64$ e o $QM_{Res.} = 1$. Apesar disso, o valor da estatística F é o mesmo entre as duas padronizações para mesma VC .

Tabela 5.9 - Análise de variância (ANOVA) das variáveis canônicas (VCs) padronizadas obtidas posteriormente a MANOVA de um experimento instalado sob o DBC

Padronização	Fonte de variação	GL	\widehat{VC}_1			\widehat{VC}_2		
			SQ	QM	F	SQ	QM	F
1 ($g'Eg = 1$)	Bloco	8	8,93	1,11	71,5	7,94	0,99	63,5
	Tratamento	8	30,02	3,75	240,1	25,94	3,24	207,5
	Resíduo	64	1,00	0,016		1,00	0,016	
2 ($e'Se = 1$)	Bloco	8	571,76	71,47	71,5	507,66	63,46	63,5
	Tratamento	8	1921,06	240,13	240,1	1659,99	207,5	207,5
	Resíduo	64	64,0	1,00		64,0	1,00	

Resultados na tabela estão arredondados.

A Figura 5.2 apresenta a média dos escores por tratamento em função da \widehat{VC}_1 e \widehat{VC}_2 pela padronização 1 (Figura 5.2 A) e 2 (Figura 5.2 B).

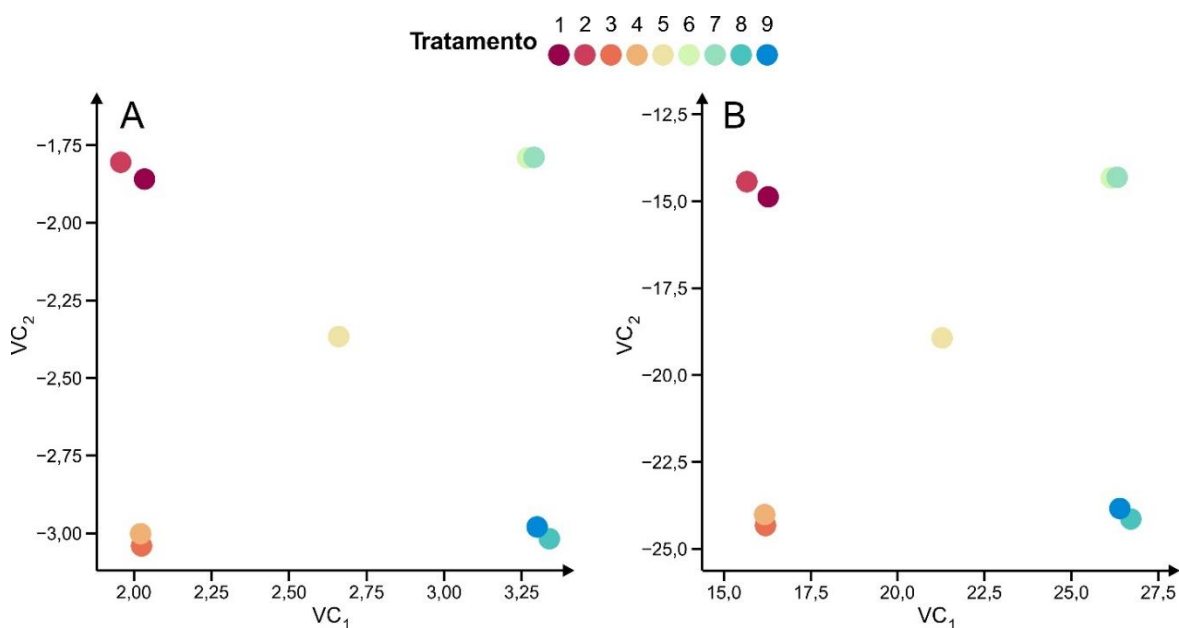


Figura 5.2 - Média dos escores por tratamento em função da \widehat{VC}_1 e \widehat{VC}_2 pela padronização 1 ($g'Eg = 1$) (A) e 2 ($e'Se = 1$) (B) no DBC.

5.3.2. Funções Discriminantes

O conjunto dos valores transformados a partir da retirada dos efeitos de bloco e utilizados para obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ funções discriminantes (FDs) também apresentou multinormalidade (*Assimetria*: $p > 0,05$; *Curtose*: $p > 0,05$) e igualdade das matrizes de covariâncias ($p > 0,05$), respectivamente, de acordo com o teste de Mardia e o teste M-Box. Uma vez que a função *lda* do pacote MASS não informa as matrizes que representam a variação entre e dentro dos grupos, elas foram

obtidas de acordo com o procedimento descrito no item 4.3.2. Dessa forma, tem-se a matriz $\widehat{\mathbf{B}}_{\mu}$, que representa a variação entre os tratamentos:

$$\widehat{\mathbf{B}}_{\mu} = \begin{bmatrix} 2107,5 & 2106,5 & -254,6 & -105,4 \\ 2106,5 & 2106,6 & -274,1 & -124,5 \\ -254,6 & -274,1 & 17717,8 & 17458,9 \\ -105,4 & -124,6 & 17458,9 & 17218,9 \end{bmatrix}$$

E, a matriz \mathbf{S} , que representa a variação dentro dos tratamentos:

$$\mathbf{S} = \begin{bmatrix} 7,82 & 7,68 & -1,01 & -0,61 \\ 7,68 & 9,04 & -1,99 & -1,10 \\ -1,01 & -1,99 & 77,26 & 72,55 \\ -0,61 & -1,10 & 72,55 & 81,37 \end{bmatrix}$$

A matriz $\widehat{\mathbf{B}}_{\mu}$ é idêntica a matriz de quadrados médios e produtos médios dos tratamentos provenientes da MANOVA no DBC, mas a matriz \mathbf{S} é diferente, uma vez que, a função *lda* utiliza apenas os graus de liberdade relacionados as fontes de variação entre e dentro dos tratamentos, ou seja, não há perda de graus de liberdade para o fator bloco.

Os escores das *FDs* padronizadas foram utilizados para encontrar os autovalores e, conseqüentemente, a importância relativa de cada função discriminante também foi encontrada. A Tabela 5.10 apresenta os coeficientes para cada padronização 2 bem como os autovalores e a importância relativa (IR) (que é, em termos absolutos, idêntica ao valor da proporção do traço que é gerada pela função *lda*) associada a cada função discriminante (\widehat{FD}_m , para $m = 1, 2, 3$ e 4).

O teste de dimensionalidade de Bartlett indicou a \widehat{FD}_1 e \widehat{FD}_2 como suficientes para explicar a maior parte da variação total (primeiro valor da estatística do teste não significativo: $D_2^2 = 7,97 < \chi_{[(4-2)(9-1-2)]}^2 = 21,03$); estas duas variáveis explicam juntas quase 100% da variação total (Tabela 5.10).

Tabela 5.10 - Funções discriminantes (*FDs*) padronizadas obtidas de um experimento instalado sob o DBC

\widehat{FD}_m	Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)				$\hat{\lambda}_m$	IR (%)	Proporção do traço
	Y_1	Y_2	Y_3	Y_4			
\widehat{FD}_1	0,324	0,034	0,002	0,005	30,017	53,53	0,5353
\widehat{FD}_2	0,064	-0,059	-0,087	-0,029	25,937	46,25	0,4625
\widehat{FD}_3	-0,821	0,824	0,031	-0,031	0,084	0,15	0,0015
\widehat{FD}_4	0,022	-0,003	0,269	0,273	0,042	0,07	0,0007

As correlações das duas primeiras FDs com as variáveis-resposta originais Y_1 , Y_2 , Y_3 e Y_4 apresentaram o mesmo valor que as correlações obtidas com as duas primeiras VCs (Tabela 5.8). Além disso, as análises de variância dos escores das FDs padronizadas obtidas do experimento instalado sob o DBC apresentaram exatamente o mesmo valor do teste F para os fatores de bloco e tratamento (Tabela 5.11), embora, os $QM_{Res.}$ foram maiores do que aqueles obtidos para as VCs , uma vez que as $SQ_{Res.}$ apresentaram o valor de 72 para ambas as FDs (Tabela 5.11).

Tabela 5.11 - Análise de variância (ANOVA) das funções discriminantes lineares ($\widehat{FD}s$) padronizadas obtidas com os escores dos dados originais de um experimento instalado sob o DBC

Padronização	Fonte de variação	GL	\widehat{FD}_1			\widehat{FD}_2		
			SQ	QM	F	SQ	QM	F
2 ($e'Se = 1$)	Bloco	8	643,23	80,404	71,47	571,12	71,47	63,5
	Tratamento	8	2161,20	270,15	240,1	1867,5	233,44	207,5
	Resíduo	64	72,0	1,125		72,0	1,125	

Finalmente, a disposição gráfica das médias dos escores por tratamento também foi a mesma como apresentado na Figura 5.2, já que os valores médios das duas primeiras FDs foram idênticos aos das duas primeiras VCs na padronização 2.

5.4. Delineamento em Quadrado Latino

5.4.1. Variáveis Canônicas

Os valores observados considerando um experimento multivariado simulado realizado sob o DQL foram obtidos e os resíduos apresentaram multinormalidade (*Assimetria*: $p > 0,05$; *Curtose*: $p > 0,05$) e igualdade das matrizes de covariâncias ($p > 0,05$), respectivamente, de acordo com o teste de Mardia e o teste M-Box. As matrizes H e E (Tabela 5.12) foram obtidas da análise de variância multivariada para posterior obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ VCs associadas às variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4 .

Tabela 5.12 - Resultado parcial da análise de variância multivariada (MANOVA) de um experimento hipoteticamente instalado sob o DQL com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4

FV	GL	SQ e SP				QM e PM			
Lin	8	-				-			
Col	8	-				-			
Trat	8	$H = \begin{bmatrix} 16860,32 & 16851,96 & -2036,69 & -843,22 \\ 16851,96 & 16852,74 & -2192,76 & -996,36 \\ -2036,69 & -2192,76 & 141742,45 & 139671,19 \\ -843,22 & -996,36 & 139671,19 & 137750,94 \end{bmatrix}$				-			
Res	56	$E = \begin{bmatrix} 524,38 & 526,13 & -71,90 & -42,80 \\ 526,13 & 621,15 & -97,99 & -31,44 \\ -71,90 & -97,99 & 5244,96 & 4895,75 \\ -42,80 & -31,44 & 4895,75 & 5495,52 \end{bmatrix}$				$S = \begin{bmatrix} 9,36 & 9,40 & -1,28 & -0,76 \\ 9,40 & 11,09 & -1,75 & -0,56 \\ -1,28 & -1,75 & 93,66 & 87,42 \\ -0,76 & -0,56 & 87,42 & 98,13 \end{bmatrix}$			

Os autovetores (\mathbf{v}) obtidos a partir da aplicação da função *eigen* no produto das matrizes $\mathbf{E}^{-1}\mathbf{H}$ foram reescalados utilizando os valores residuais da MANOVA e as VCs foram obtidas de acordo com as padronizações 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$) e 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$). Os escores das VCs padronizadas foram utilizados para encontrar os autovalores e, conseqüentemente, a importância relativa de cada VC também foi encontrada. A Tabela 5.13 apresenta os coeficientes para cada padronização bem como os autovalores e a importância relativa (IR) associada a cada variável canônica (\widehat{VC}_m , para $m = 1, 2, 3$ e 4).

O teste de dimensionalidade de Bartlett indicou a \widehat{VC}_1 e \widehat{VC}_2 como suficientes para explicar a maior parte da variação total (primeiro valor da estatística do teste não significativo: $D_2^2 = 7,75 < \chi_{[(4-2)(9-1-2)]}^2 = 21,03$); estas duas variáveis explicam juntas quase 100% da variação total (Tabela 5.13).

Tabela 5.13 - Variáveis canônicas padronizadas (VCs) obtidas posteriormente a MANOVA de um experimento instalado sob o DQL com quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4

\widehat{VC}_m	Padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)				Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)				$\hat{\lambda}_m$	IR (%)
	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4		
\widehat{VC}_1	0,0447	-0,0010	0,00002	0,0006	0,3347	-0,0078	0,002	0,0045	32,170	53,84
\widehat{VC}_2	0,0017	-0,0015	-0,0100	-0,004	0,0128	-0,0110	-0,0747	-0,0299	27,445	45,93
\widehat{VC}_3	0,1040	-0,1044	-0,0045	0,0045	0,7782	-0,7810	-0,0388	0,0334	0,097	0,16
\widehat{VC}_4	-0,0035	0,0012	-0,0321	0,0326	-0,0262	-0,0093	-0,2405	0,2438	0,043	0,07

As correlações das duas primeiras VCs com as variáveis-resposta originais Y_1 , Y_2 , Y_3 e Y_4 do experimento instalado sob o DQL também foram calculadas e são demonstradas na Tabela 5.14. Para ambas as padronizações, as correlações foram as mesmas.

Tabela 5.14 - Correlação obtida entre as variáveis canônicas (VCs) padronizadas obtidas posteriormente a MANOVA e as variáveis-resposta de um experimento instalado sob o DQL

\widehat{VC}_m	Padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)				Padronização 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)			
	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
\widehat{VC}_1	1,00	1,00	0,36	0,38	1,00	1,00	0,36	0,38
\widehat{VC}_2	-0,32	-0,32	-1,00	-1,00	-0,32	-0,32	-1,00	-1,00

As análises de variâncias das VCs obtidas do experimento instalado sob o DBC foram também procedidas e a Tabela 5.15 demonstra os resultados para as estimativas das \widehat{VC}_1 e \widehat{VC}_2 . É importante destacar que as VCs na padronização 1 apresentaram $SQ_{Res.} = 1$ e o $QM_{Res.} = 0,018$, porém, na padronização 2, elas apresentaram a $SQ_{Res.} = 56$ e o $QM_{Res.} = 1$. Apesar disso, o valor da estatística F é o mesmo entre as duas padronizações para mesma VC .

Tabela 5.15 - Análise de variância (ANOVA) das variáveis canônicas (VCs) padronizadas obtidas posteriormente a MANOVA de um experimento instalado sob o DQL

Padronização	Fonte de variação	GL	\widehat{VC}_1			\widehat{VC}_2		
			SQ	QM	F	SQ	QM	F
1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$)	Linha	8	9,43	1,18	66,0	8,52	1,07	59,7
	Coluna	8	8,50	1,06	59,5	8,56	1,07	59,9
	Tratamento	8	32,17	4,02	225,2	27,45	3,43	192,1
	Resíduo	56	1,00	0,018		1,00	0,018	
2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$)	Linha	8	528,2	66,02	66,0	477,2	59,65	59,7
	Coluna	8	475,9	59,449	59,5	479,3	59,01	59,0
	Tratamento	8	1801,5	225,19	225,2	1536,9	192,12	192,1
	Resíduo	56	56,00	1,00		56,00	1,00	

Resultados na tabela estão arredondados.

A Figura 5.3 apresenta a média dos escores por tratamento em função da \widehat{VC}_1 e \widehat{VC}_2 pela padronização 1 (Figura 5.3 A) e 2 (Figura 5.3 B).

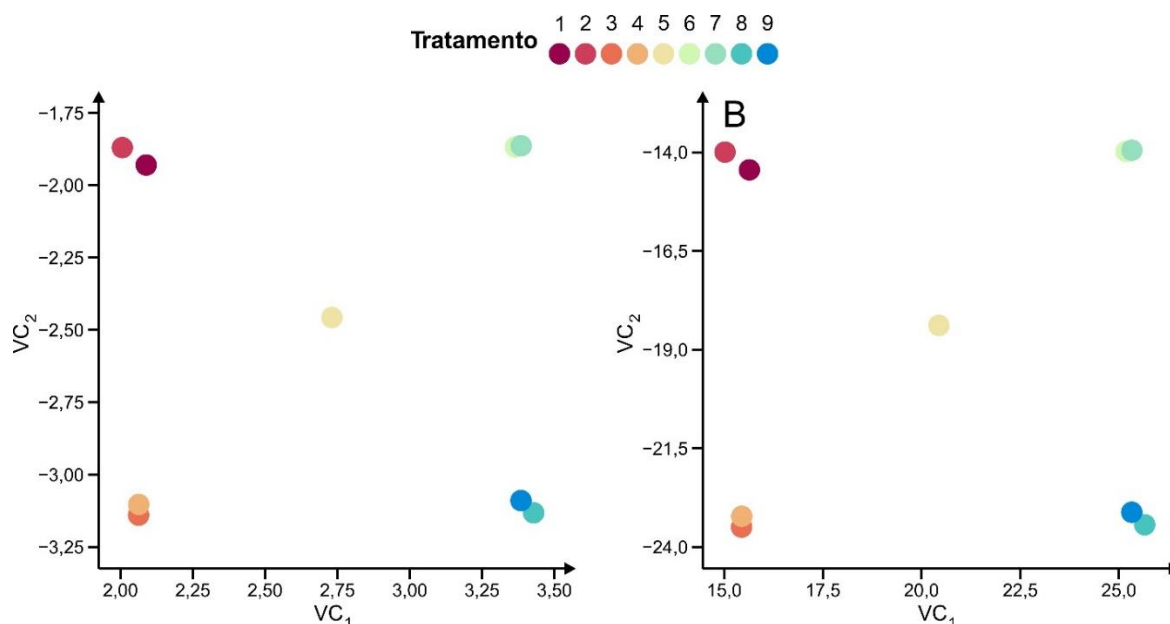


Figura 5.3 - Média dos escores por tratamento em função da \widehat{VC}_1 e \widehat{VC}_2 pela padronização 1 ($\mathbf{g}'\mathbf{E}\mathbf{g} = 1$) (A) e 2 ($\mathbf{e}'\mathbf{S}\mathbf{e} = 1$) (B) no DQL.

5.4.2. Funções Discriminantes

O conjunto dos valores transformados a partir da retirada dos efeitos de linha e coluna, e utilizados para obtenção das $s = \text{mín.}(p = 4, t - 1 = 8) = 4$ funções discriminantes (*FDS*) também apresentou multinormalidade (*Assimetria*: $p > 0,05$; *Curtose*: $p > 0,05$) e igualdade das matrizes de covariâncias ($P > 0,05$), respectivamente, de acordo com o teste de Mardia e o teste M-Box. Uma vez que a função *lda* do pacote *MASS* não informa as matrizes que representam a variação entre e dentro dos grupos, elas foram obtidas de acordo com o procedimento descrito no item 4.4.2. Dessa forma, tem-se a matriz $\widehat{\mathbf{B}}$, que representa a variação entre os tratamentos:

$$\widehat{\mathbf{B}}_{\mu} = \begin{bmatrix} 2107,5 & 2106,5 & -254,6 & -105,4 \\ 2106,5 & 2106,6 & -274,1 & -124,5 \\ -254,6 & -274,1 & 17717,8 & 17458,9 \\ -105,4 & -124,6 & 17458,9 & 17218,9 \end{bmatrix}$$

E, a matriz \mathbf{S} , que representa a variação dentro dos tratamentos:

$$\mathbf{S} = \begin{bmatrix} 7,28 & 7,31 & -1,00 & -0,59 \\ 7,31 & 8,63 & -1,36 & -0,44 \\ -1,00 & -1,36 & 72,85 & 68,00 \\ -0,59 & -0,44 & 68,00 & 76,33 \end{bmatrix}$$

A matriz $\widehat{\mathbf{B}}_{\mu}$ é idêntica a matriz de quadrados médios e produtos médios dos tratamentos provenientes da MANOVA no DQL, mas a matriz \mathbf{S} é diferente, uma vez que, a função *lda* utiliza apenas os graus de liberdade relacionados as fontes de variação

entre e dentro dos tratamentos, ou seja, não há perda de graus de liberdade para os fatores linha e coluna.

Os escores das *FDs* padronizadas foram utilizados para encontrar os autovalores e, conseqüentemente, a importância relativa de cada função discriminante também foi encontrada. A Tabela 5.16 apresenta os coeficientes para cada padronização 2 bem como os autovalores e a importância relativa (IR) (que é, em termos absolutos, idêntica ao valor da proporção do traço que é gerada pela função *lda*) associada a cada função discriminante (\widehat{FD}_m , para $m = 1, 2, 3$ e 4).

O teste de dimensionalidade de Bartlett indicou a \widehat{FD}_1 e \widehat{FD}_2 como suficientes para explicar a maior parte da variação total (primeiro valor da estatística do teste não significativo: $D_2^2 = 7,75 < \chi_{[(4-2)(9-1-2)]}^2 = 21,03$); estas duas funções explicam juntas quase 100% da variação total (Tabela 5.16).

Tabela 5.16 - Funções discriminantes (*FDs*) padronizadas obtidas de um experimento instalado sob o DQL

\widehat{FD}_m	Padronização 2 ($e'Se = 1$)				$\hat{\lambda}_m$	IR (%)	Proporção do traço
	Y_1	Y_2	Y_3	Y_4			
\widehat{FD}_1	0,379	-0,009	0,000	0,005	32,170	53,84	0,5384
\widehat{FD}_2	0,015	-0,012	-0,085	-0,034	27,445	45,93	0,4593
\widehat{FD}_3	-0,882	0,886	0,038	-0,038	0,097	0,16	0,0016
\widehat{FD}_4	0,030	-0,011	0,273	0,276	0,043	0,07	0,0007

As correlações das duas primeiras *FDs* com as variáveis-resposta originais Y_1 , Y_2 , Y_3 e Y_4 apresentaram o mesmo valor que as correlações obtidas com as duas primeiras *VCs* (Tabela 5.14). Além disso, as análises de variância dos escores das *FDs* padronizadas obtidas do experimento instalado sob o DQL apresentaram exatamente o mesmo valor do teste F para os fatores de linha, coluna e tratamento (Tabela 5.17), embora, os $QM_{Res.}$ foram maiores do que aqueles obtidos para as *VCs*, uma vez que as $SQ_{Res.}$ apresentaram o valor de 72 para ambas as *FDs* (Tabela 5.17).

Tabela 5.17 - Análise de variância (ANOVA) das funções discriminantes lineares (FDS) padronizadas obtidas dos dados transformados de um experimento instalado sob o DQL

Padronização	Fonte de variação	GL	\widehat{FD}_1			\widehat{FD}_2		
			SQ	QM	F	SQ	QM	F
2 ($e'Se = 1$)	Linha	8	679,10	84,88	66,0	613,50	76,65	59,7
	Coluna	8	611,84	76,48	59,5	616,21	77,03	59,9
	Tratamento	8	2316,2	289,53	225,2	1976,06	247,01	192,1
	Resíduo	56	72,00	1,286		72,00	1,286	

Resultados na tabela estão arredondados.

Finalmente, a disposição gráfica das médias dos escores por tratamento também foi a mesma como apresentado na Figura 5.3, já que os valores médios das duas primeiras FDS foram idênticos aos das duas primeiras VCs na padronização 2.

6. DISCUSSÃO

A análise das variáveis canônicas é uma técnica multivariada utilizada na identificação e separação de grupos conhecidos *a priori*. Em suma, combinações lineares de variáveis quantitativas são encontradas a fim de maximizar a separação de indivíduos. Estas combinações lineares de variáveis são encontradas a partir da maximização da variação existente entre os grupos e dentro deles, e são chamadas de variáveis canônicas (VCs). No contexto de experimentos executados com tratamentos repetidos sob diferentes delineamentos experimentais, tais como o DIC, DBC e DQL, as VCs podem ser obtidas como sendo as combinações lineares das p variáveis-resposta Y_p avaliadas ($p = 1, 2, \dots, n$) e os tratamentos representarão os grupos conhecidos *a priori* (RENCHE, 2002).

A técnica conhecida atualmente para a obtenção das VCs a partir de variáveis coletadas em experimentos pressupõe a realização da MANOVA para utilização das matrizes de soma de quadrados e produtos de tratamentos (H) e residual (E) como fontes da variação entre e dentro dos tratamentos, respectivamente. Isso é necessário tanto para os experimentos executados sob DIC, mas, especialmente para aqueles com fatores perturbadores, tais como o DBC e DQL, já que para obtenção das VCs , toda a fonte de variação precisa ser decomposta para uso somente daquelas necessárias para o método, ou seja, a variação entre e dentro dos tratamentos. Para experimentos executados sob o DIC, as VCs podem ser também obtidas pela aplicação direta das funções discriminantes lineares de Fisher (FDS), uma vez que, matematicamente, sua derivação é equivalente a

das *VCs*, e sob este delineamento experimental, apenas as fontes de variação entre e dentro dos tratamentos estão presentes (NØRGAARD et al. 2006, RIBA et al. 2020).

Considerando a equivalência matemática na obtenção das *VCs* e das *FDs*, uma transformação linear dos modelos estatísticos dos delineamentos experimentais com fatores perturbadores, tais como DBC e DQL, pode ser executada para fornecer apenas as fontes de variação entre e dentro dos tratamentos necessárias para a obtenção das *FDs*, as quais, semelhantemente, funcionam como as *VCs* obtidas pelo uso das matrizes **H** e **E** da *MANOVA*. Na prática, isso pode ser facilmente realizado pela aplicação da função *lda* (*linear discriminant analysis*) – do pacote *MASS* do software R (VENABLES & RIPLEY, 2002) – nos valores observados corrigidos pela retirada dos efeitos perturbadores de experimentos executados sob o DBC ou DQL; para experimentos realizados em DIC, a função *lda* é aplicada nos valores observados originais.

Neste sentido, este trabalho propôs demonstrar as similaridades das *VCs* obtidas a partir das matrizes **H** e **E** da *MANOVA* com as *FDs* obtidas a partir dos dados transformados pela modificação dos modelos estatísticos (exceto para o DIC) em condições computacionalmente simuladas de experimentos instalados sob o DIC, DBC e DQL. Adicionalmente, duas diferentes padronizações dos autovetores (padronização 1: $\mathbf{g}'\mathbf{E}\mathbf{g} = 1$; padronização 2: $\mathbf{e}'\mathbf{S}\mathbf{e} = 1$), que representam os coeficientes das *VCs*, foram adotadas. Já as *FDs* estimadas pela função *lda* apresentam a padronização 2.

Diversas medidas foram utilizadas para comprovar a similaridade dos métodos tais como os autovalores, autovetores e IR (%) ou proporção do traço, assim como, a correlação existente entre as funções (*VCs* ou *FDs*) e as variáveis-resposta *Ys*, a análise de variância e uma representação gráfica da relação entre as duas primeiras funções (*VCs* ou *FDs*) que acumularam a maior parte da variabilidade dos dados. Além disso, para tornar os resultados comparáveis, todos os procedimentos foram realizados a partir de um único conjunto de erros experimentais simulados, os quais foram utilizados para dar origem aos valores observados das quatro variáveis-resposta Y_1 , Y_2 , Y_3 e Y_4 consideradas nos experimentos simulados e instalados sob o DIC, DBC ou DQL com 9 tratamentos e 9 repetições para cada um deles.

Após a adição do vetor de médias geral, dos efeitos de tratamentos e dos efeitos perturbadores (quando existentes) à matriz de resíduos simulada (81×4), os resíduos da *MANOVA*, seja para o DIC, DBC ou DQL, apresentaram normalidade multivariada e igualdade de matrizes de variâncias de acordo com os testes M-Box e de Mardia, respectivamente. Portanto, em todas as situações, a pressuposição de igualdade das

matrizes de covariâncias entre os tratamentos para obtenção das *VCs* ou *FDs* (FISHER, 1936, 1938) foi atendida; embora, a normalidade *p*-variada não seja um pressuposto para os métodos, ela é requerida quando inferências estatísticas são realizadas, como destacam Owen & Chmielewski (1984) e Manly & Alberto (2016).

É interessante notar que as matrizes que representam a variação entre os tratamentos, ou seja, \mathbf{H} ou \mathbf{B}_μ , são respectivamente as mesmas para obtenção das *VCs* ou das *FDs*, sejam estas matrizes obtidas dos valores observados originais ou daqueles transformados pela retirada dos efeitos perturbadores. Nesse sentido, destaca-se que as diferenças observadas nas estimativas das *VCs* ou *FDs* entre os diferentes delineamentos de casualização (DIC, DBC ou DQL) surgem, primeiramente, devido às diferentes matrizes que representam a variação dentro dos tratamentos, ou seja, às matrizes \mathbf{E} ou \mathbf{S} . Secundariamente, a padronização utilizada para os autovetores, seja ela $\mathbf{g}'\mathbf{E}\mathbf{g} = 1$ ou $\mathbf{e}'\mathbf{S}\mathbf{e} = 1$, interfere apenas na escala numérica com as quais as *VCs* serão trabalhadas. Isso porque, dentro do mesmo delineamento, independentemente da padronização, as *VCs*, terão a mesma interpretabilidade e utilidade para análises *a posteriori*, já que possuirão, como demonstrado neste trabalho, os mesmos autovalores e porcentagem de IR, os mesmos valores de correlação entre as funções e as variáveis-resposta originais, os mesmos valores para o F calculado, e dispersão espacial das observações idêntica quando graficamente representados. Nesse sentido, a escolha da padronização é meramente uma opção por qual escala se deseja trabalhar, embora, seja mais comum na literatura a padronização pela qual as *VCs* ou *FDs* possuam variância unitária, ou seja, $\mathbf{e}'\mathbf{S}\mathbf{e} = 1$ (GITTINGS, 1985), o que pode ser verificado pelo valor do $QM_{Res.}$ da ANOVA procedida com os escores das *VCs*.

Na comparação dos resultados entre as *VCs* e as *FDs* com a padronização 2 para o mesmo delineamento (neste caso, para o DBC ou no DQL), ao se dar origem as *FDs* com as observações transformadas, o $QM_{Res.}$ da ANOVA com os escores das *FDs* obtidos a partir das observações originais possuiu valor maior que 1. Isso foi porque a matriz \mathbf{S} utilizada para obtenção das *FDs* no DBC e no DQL foi diferente daquela utilizada para obtenção das *VCs* nestes delineamentos, pois não houve redução dos graus de liberdade para gerar a matriz \mathbf{S} quando a função *lda* é utilizada. Isso preservou, na ANOVA, a $SQ_{Res.}$ no valor de 72 para as *FDs* no DBC e no DQL (mesmo valor encontrado com a ANOVA dos escores obtidos pela *FD* no DIC), o que resultou nos valores de QM maiores que 1. O resultado final, porém, do teste F para a ANOVA é o mesmo entre as *FDs* e as *VCs* sob o mesmo delineamento experimental, o que demonstra consistência no uso da

função *lda* para obtenção das *FDs* a partir dos valores das observações transformados das variáveis-resposta. Porém, os autovetores que dão origem aos coeficientes das *FDs* podem ser corrigidos de modo que eles sejam os mesmos das *VCs* no DBC ou no DQL. Assim, todo o quadro da ANOVA será idêntico entre as *FDs* e *VCs* de mesma ordem. Essa correção pode ser realizada multiplicando os autovetores (\hat{e}) pela raiz da divisão entre os graus de liberdade do resíduo no DBC ou DQL ($GL_{Res.(DBC\ ou\ DQL)}$) e os graus de liberdade do resíduo do DIC ($GL_{Res.(DIC)}$), ou seja: $\hat{e}_{corrigido} = \hat{e} \sqrt{\frac{GL_{Res.(DBC\ ou\ DQL)}}{GL_{Res.(DIC)}}$.

Os autovetores, frutos da solução do sistema de equações homogêneas para obtenção das *VCs* ou *FDs*, correspondem nada mais nada menos, do que os pesos canônicos ou coeficientes destas funções e podem, em princípio, indicar a presença/importância de certas variáveis, seus efeitos e sua direção (GITTINS, 1985). Os resultados gerados aqui mostraram uma mudança de sinal dos coeficientes das *VCs* para as *FDs* considerando um mesmo delineamento experimental. No DIC, a mudança de sinal dos coeficientes aconteceu para as duas primeiras funções lineares, já no DBC e DQL, sinais contrários acompanharam os coeficientes das duas últimas funções lineares (dados não mostrados). Em termos interpretativos, essa troca de sinal dos coeficientes entre as *VCs* para as *FDs* é irrelevante para separação de tratamentos, já que ela ocorreu para todo o conjunto dos coeficientes relacionados às variáveis-resposta Y_1, Y_2, Y_3, Y_4 . Além disso, os sinais dos coeficientes em relação a todo o conjunto de variáveis nas *VCs* ou *FDs* podem ser alterados por conveniência (para manter, por exemplo, os valores dos escores positivos) sem que haja prejuízo na interpretação dos resultados.

Sob a adição ou remoção de variáveis-resposta, pode ocorrer instabilidade dos pesos canônicos para fins interpretativos e uma solução apresentada e discutida é a utilização das correlações entre as variáveis originais e as canônicas (neste caso, *VCs* ou *FDs*) como um indicativo mais confiável da contribuição de cada variável-resposta para formação das últimas (neste caso, *VCs* ou *FDs*) (BARGMANN, 1962). Os resultados apresentados aqui demonstram que as correlações obtidas entre *VCs* e as variáveis-resposta são as mesmas em termos absolutos do que aquelas obtidas entre os *FDs* e as variáveis-resposta, considerando o mesmo delineamento experimental, portanto, novamente, as interpretações com essas correlações serão as mesmas. Embora o estabelecimento de correlações entre as funções canônicas e as variáveis-resposta constituem um auxílio interpretativo, nem sempre elas são eficazes neste propósito uma vez que elas apenas fornecem informações sobre como as variáveis contribuem

individualmente para a separação dos grupos, desconsiderando a contribuição conjunta das variáveis, ou seja, a forma multivariada de sua contribuição para formação das variáveis canônicas ou funções discriminantes (RENCHEER, 2002).

Um dos benefícios do uso das *VCs* ou *FDs* é a redução da dimensionalidade para representação gráfica e/ou para o emprego de análises univariadas. Pelo menos dois critérios têm sido apresentados para seleção *VCs* ou *FDs*: o primeiro é estatístico com uso do teste de dimensionalidade de Bartlett, o segundo utiliza um ponto de corte (normalmente, 70%) que represente grande parte da variabilidade contida nas p variáveis originais. Este trabalho demonstrou que independentemente do delineamento utilizado e do método para obtenção das *VCs* ou *FDs*, as duas primeiras funções foram igualmente selecionadas tanto pelo critério estatístico quanto pela representatividade de variabilidade, que alcançou cerca de 100%. Para ambos os critérios, é necessário, no entanto, a utilização dos autovalores obtidos da solução do sistema de equação homogêneas para obtenção das *VCs* ou *FDs*.

A função *eigen* aplicada no produto das matrizes $E^{-1}H$, as quais são obtidas da realização da MANOVA (ou seja, no caso da obtenção das *VCs*), resulta nos autovetores normalizados do tipo $v'v = 1$ e nos autovalores. Estes últimos são necessários para execução do teste de dimensionalidade de Bartlett e para cálculo da IR (%) de cada *VC*, cujo valor acumulado é utilizado para seleção das funções pelo critério de representatividade da variação total. Por outro lado, a função *lda* do pacote *MASS* não produz os autovalores, apesar disso, fornece a proporção do traço de cada *FD*, que pode ser compreendida como a IR em termos absolutos. Nesse sentido, a seleção das *FDs* como resultado da aplicação da função *lda* será realizada utilizando, em princípio, somente a proporção do traço de cada *FD*. No entanto, para qualquer situação, o autovalor correspondente a cada *VC* ou *FD* pode ser obtido pela razão entre a $SQ_{Trat.}$ e a $SQ_{Res.}$ para cada combinação linear de variáveis (MANLY & ALBERTO, 2016). Assim, também é possível, no caso da aplicação da função *lda*, estimar os autovalores necessários para a realização do teste de dimensionalidade de Bartlett, possibilitando, portanto, um critério estatístico para escolha das combinações lineares de variáveis que serão utilizadas para inferências posteriores.

Finalmente, diante de todos os indicadores utilizados para comparar as *VCs* (obtidas após a realização da MANOVA) com as *FDs* (obtidas pela aplicação da função *lda* em dados transformados pela retirada dos efeitos perturbadores), é possível atestar o êxito do último método e a sua contribuição para ser utilizado na separação de

tratamentos, uma vez que nenhuma perda na qualidade da interpretação e nenhuma modificação nos resultados das análises *a posteriori* foi detectada nas funções lineares estimadas.

7. CONCLUSÕES

De modo geral, conclui-se, portanto, que:

- A modificação dos modelos estatísticos com fontes de variação perturbadoras para gerar valores observados transformados pode ser recomendada para a obtenção das *FDs* por meio da função *lda*;
- A similaridade teórica entre a obtenção das variáveis canônicas e funções discriminantes lineares de Fisher reflete em resultados idênticos;
- A função *lda* facilita a obtenção das variáveis canônicas por meio das funções discriminantes;
- As padronizações das variáveis canônicas apenas interferem nas escalas dos resultados na qual se quer trabalhar, mas não em suas interpretações.

8. REFERÊNCIAS BIBLIOGRÁFICAS

BARGMANN, R.E. **Representative ordering and selection of variables**. 1. ed. Virginia: Virginia Polytechnic Institute, 1962.

BARTLETT, M. S. The statistical significance of canonical correlations. **Biometrika**, v. 32, n. 1, p. 29-37, 1941.

BOX, G. E. P. A general distribution theory for a class of likelihood criteria. **Biometrika**, v. 36, n. 3-4, p. 317-346, 1949.

CRUZ, C. D. **Programa Genes: análise multivariada e simulação**. 1. ed. Viçosa: Editora UFV, 2006.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. 1. ed. Viçosa: Editora UFV, 2012.

EVERITT, B. S. **An R and S-PLUS® Companion to Multivariate Analysis**. 1. ed. London: Springer London, 2005.

FERREIRA, D. F. **Estatística multivariada**. 3. ed. Lavras: Editora UFLA, 2018.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v. 7, n. 2, p. 179-188, 1 set. 1936.

FISHER, R. A. The statistical utilization of multiple measurements. **Annals of Eugenics**, v. 8, n. 4, p. 376-386, 1938.

GITTINS, R. **Canonical analysis: a review with applications in ecology**. 1. ed. Berlin: Springer Berlin Heidelberg, 1985. (Biomatematics, v. 12)

KASSAMBARA, A. **rstatix: pipe-friendly framework for basic statistical tests**. Disponível em: <<https://cran.r-project.org/package=rstatix>>, 2020.

KORKMAZ, S.; GOKSULUK, D.; ZARARSIZ, G. MVN: An R Package for Assessing Multivariate Normality. **The R Journal**, v. 6, n. 2, p. 151-162, 2014.

MANLY, B. F. J.; ALBERTO, J. A. N. **Multivariate statistical methods: a primer**. 4. ed. Boca Ratón: CRC Press, 2016.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. 6. ed. San Diego: Academic Press, 1997.

MARDIA, K. V. Measures of multivariate skewness and kurtosis with applications. **Biometrika**, v. 57, n. 3, p. 519-530, 1970.

MCLACHLAN, G. J. **Discriminant analysis and statistical pattern recognition**. Hoboken: John Wiley & Sons, 1992. (Wiley Series in Probability and Statistics).

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. 1. ed. Belo Horizonte: UFMG, 2005.

MONTGOMERY, D. C. **Design and analysis of experiments**. 8. ed. New York: John Wiley & Sons, 2012.

MUENCHEN, R. A. **The popularity of data science software | r4stats.com**. Disponível em: <<http://r4stats.com/articles/popularity/>>. Acesso em: 20 nov. 2019.

NØRGAARD, L. et al. A modification of canonical variates analysis to handle highly collinear multivariate data. v. 20. p. 425–435, **Journal of Chemometrics**, 2006.

OWEN, J. G.; CHMIELEWSKI, M. A. On canonical variates analysis and the construction of confidence ellipses in systematic studies. **Systematic Biology**, v. 34, n. 3, p. 366–374, set. 1985.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Disponível em: <<https://www.r-project.org/>>, 2020

RENCHEER, A. C. **Methods of Multivariate Analysis**. 2. ed. Wiley, 2002. (Wiley Series in Probability and Statistics).

RIBA, J.-R. et al. Multivariate identification of extruded PLA samples from the infrared spectrum. **Journal of Materials Science**, v. 55, n. 3, p. 1269–1279, 2020.

TIPPMANN, S. Programming tools: Adventures with R. **Nature**, v. 517, n. 7532, p. 109–110, 1. 2015.

VENABLES, W. N.; RIPLEY, B. D. **Modern applied statistics with S**. 4. ed. New York: Springer, 2002.

WICKHAM, H. et al. Welcome to the {tidyverse}. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.