Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Auxílios à interpretação na Análise Discriminante de Fisher

Autora: Celeste María Díaz Cónsul

Orientadora: Regina C.C.P. Moran

Campinas, 5 de maio de 2000

UNICAMP
BIBLIOTECA CENTRAL
SEÇÃO CIRCULANTE

UNICAMP
BIBLIOTECA CENTRAL

UNIDADE BC
Nº CHAMADA:
T/UNICAMP
D543a
V. Ex.
TOMBO BC/42206
PROC. 16-278/00
C [ ] D [X]
PREÇO R$ 11,00
DATA 21/09/00
Nº CPD

CM-00145109-8

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP

Díaz Cónsul, Celeste María

D543a Auxílios à interpretação na análise discriminante de Fisher / Celeste María Díaz Cónsul -- Campinas, [S.P. :s.n.], 2000.

Orientador : Regina Célia de Carvalho Pinto Moran

Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Análise multivariada. 2. Análise discriminante. I. Moran, Regina Célia de Carvalho Pinto. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Auxílios à interpretação na Análise Discriminante de Fisher

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Celeste María Díaz Cónsul, e aprovada pela Comisão Julgadora.

Campinas, 5 de maio de 2000

Profa. Dra. Regina C.C.P. Moran
Orientadora

Banca Examinadora:

Profa. Dra. Regina C.C.P. Moran
Prof. Dr. Reinaldo Charnet
Prof. Dr. George John Shepherd

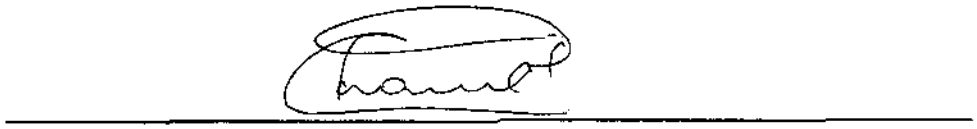
Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica, UNICAMP, como requisito parcial para a obtenção do Título de MESTRE em Estatística.

UNICAMP
BIBLIOTECA CENTRAL
SEÇÃO CIRCULANTE

Dissertação de Mestrado defendida e aprovada em 05 de maio de 2000

pela Banca Examinadora composta pelos Profs. Drs.

Prof (a). Dr (a). REGINA CÉLIA CARVALHO P. MORAN



Prof (a). Dr (a). REINALDO CHARNET

Prof (a). Dr (a). GEORGE JOHN SHEPHERD

UNICAMP
BIBLIOTECA CENTRAL
SEÇÃO CIRCULANTE

Dedico esta tese a meus queridíssimos pais, minhas inseparáveis irmãs e meus adorados sobrinhos.

UNICAMP
BIBLIOTECA CENTRA
SEÇÃO CIRCULANTE

AGRADECIMENTOS

Muitas pessoas tem contribuído de uma maneira ou de outra à realização deste trabalho. É por isto que não ficaria tranquila sem escrever esta nota de agradecimento, lamentavelmente ela não pode ser tão extensa quanto eu gostaria...

É à Prof. Regina quem quero destacar primeiro. Sua orientação dedicadíssima e entusiasmo infinito me deram confiança e segurança sempre que precisei. Além disso ela é um exemplo de professora, pesquisadora e acima de tudo uma pessoa maravilhosa pela qual sempre guardarei um imenso respeito e carinho. Tive muita sorte de ela ter me aceitado como orientada.

O inicio dos estudos foi muito difícil para mim, demorei um pouco para me adaptar ao Brasil, cheio de coisas novas e tendo pela frente o curso mais exigente que já fiz. Neste momento difícil e estressante tive a honra de ter do meu lado pessoas como Maristela (Mari), Sirlei (Sisi), Leonardo (Leo) e Marcelo (Barriginha), meus companheiros de turma que junto com Daniela (Daaaani), Rui (magrinho) e Fernando (cubano) deram-me desinteressadamente sua valiosa amizade. Eles me ensinaram desde estatística e SAS até história e música brasileiras, além de preocuparse e ocuparse de melhorar o mais possível meu português .... lamentavelmente eu ainda os decepciono com meu "portunhol".

Falando de portunhol, agradeço muitíssimo a minha amiga Sheila por ter feito a revisão da maior parte deste trabalho, com a rapidez que precisei, só ela para não se queixar da extensão desta tese...!

O pessoal do NEA, Instituto de Economia, no pode ficar de fora. Foi um milagre encontrar trabalho, com a necessária remuneração, no momento em que mais precisava. Se só tivesse sido isso já estaria bom, mais não, não foi somente isso, tinha mais detrás do grupo de trabalho. São pessoas tão queridas como meus amigos Tuca (Antônio Márcio), ele é único; o Marcelo Magalhães, exigeeeente; o Marcelo (Melo) com quem adoro conversar e as meninas Carol e Daniela.

Também quero agradecer ao Prof. George Shepherd e a Fabio Vitta, orientador de doutorado e orientando, respectivamente, do Instituto de Biologia da UNICAMP, por emprestar os dados que serviram para desenvolver o capítulo de aplicação, além das valiosos sugestões e comentários.

Agradeço também aos professores do Departamento de Estatística do IMECC que souberam transmitir sua experiência de uma forma didática e efetiva. Lamentavelmente nem todos seguem este padrão. Mesmo assim tive a sorte de contar com excelentes professores na maioria das matérias.

À CAPES pela ajuda financeira.

Também quero mencionar a Anderson, Marcelo, Olga, Fernando (Cavalinho), Mariana, Carlos (Muchacho), Luzia, entre outros amigos de diferentes turmas do mestrado em estatística que tem contribuído a fzer a estadia no laboratório y no predinho mais prazerosa.

Quero agradecer também ao Gustavo Fávaro, companheiro de turma e atualmente trabalhador do CENAPAD, pela ajuda com minhas dúvidas de SAS para as quais ele sempre teve uma resposta rápida e eficaz.

O pessoal da minha casa, tanto as meninas Ana, Helô, Janete (quem me levou aos dados), Sandra, Denise e Déborah, quanto a Dona Rosa e o Seu Agenor, tem sido muito gentis e amigos, a eles também meu agradecimento.

O Richard ocupa um lugar importante nesta enumeração de agradecimentos, ele tem me dado amor, ajuda e compreensão, além de infinita paciência, coisas das que tenho precisado muito nestes meses de intenso trabalho e que precisarei sempre.

Por último minha maravilhosa família em Cuba ... sempre tão perto mesmo estando longe. Eles tem sido minha inspiração, do amor deles tenho tirado forças nos momentos mais difíceis. É por isso que este trabalho está especialmente dedicado a minha mãe, meu pai, minhas três irmãs e meus quatro sobrinhos. Além deles também estão minha avó Celeste, meu avô Matías e minha tia Julia, para eles meu agradecimento "in memoriam" desde o fundo do meu coração pelo amor que me entregaram.

RESUMO

Cada vez mais a análise multivariada é empregada para resolver problemas nas ciências biológicas, a indústria, estudos demográficos, etc. A análise discriminante de Fisher é uma das técnicas mais usadas quando se trata de dados com estrutura de grupo.

A interpretação dos coeficientes nas funções lineares discriminantes de Fisher apresenta especificidades adicionais em relação, por exemplo, à técnica de componentes principais. A razão para estas é a presença das fontes de variação entre grupos e dentro dos grupos, levando a mais de uma métrica e portanto mais de uma padronização.

Neste trabalho os coeficientes da função linear discriminante de Fisher são estudados através de diferentes configurações para estruturas de grupo. Desta forma, através de estruturas conhecidas, foi possível focalizar as relações das estruturas com os coeficientes; consequentemente mais clareza para a partir dos coeficientes inferir possíveis estruturas de grupo.

Também construímos, como auxílio à interpretação, os círculos de correlação para a análise discriminante de Fisher. Estes permitem inferir a direção e magnitude das contribuições das variáveis na separação dos grupos, bem como recuperar as direções e magnitudes relativamente à estrutura de variabilidade dentro dos grupos.

Foi desenvolvido o uso de pontos suplementares na análise discriminante de Fisher, recurso que se tem mostrado muito importante quando se quer estudar indivíduos que não participem dos critérios de agrupamento iniciais dos dados.

Finalmente, uma aplicação ilustra todo o desenvolvimento metodológico deste trabalho e um roteiro para auxiliar na interpretação foi produzido.

Palavras chave: análise discriminante de Fisher, círculo de correlações, ponto suplementar

ABSTRACT

Multivariable analysis is increasingly being used to solve problems in biological science, in industry, demographic studies, etc. The Fisher´s discriminant analysis is one of the more frequently used techniques for analysis of data with group structure.

The interpretation of the coefficients in the Fisher´s discriminant linear functions present additional specifications in relation, for example, to the technique of principal components. The reason for these is the presence of sources of variation between and inside the groups. Resulting in more than one metric and therefore more than one standardization.

In this work the coefficients of Fisher´s discriminant linear function are studied through the different configurations for the structures of the group. In this way, through known structures, it was possible to focus on the relations of the structures with the coefficients; consequently making it easier to infer possible structures of the group from the coefficients.

As an aid to this interpretation the correlation circles for Fisher´s discriminant analysis were constructed. These allow us to infer the direction and the magnitude of the contributions of variables in group separation, as well as to recover the directions and magnitudes relatively to the variability structure inside the groups.

The use of supplementary points in the Fisher´s discriminant analysis was developed, a resource that has been shown to be very important when we want to study individuals that do not participate of the initial grouping criteria of the data.

Finally, an application illustrates all the methodological development of this work and a guide to help with the interpretation was produced.

Key words: Fisher's discriminant analysis, correlation circles, suplementary point.

SUMÁRIO

LISTA DE FIGURAS v

LISTA DE TABELAS vii

LISTA DE EQUAÇÕES xii

CAPÍTULO 1: INTRODUÇÃO 1

1.1 Análise Discriminante de Fisher 1

1.1.1 Ilustrando geometricamente o primeiro eixo discriminante 1

1.1.2 Motivação para o estudo dos coeficientes da Função Linear Discriminante de Fisher 2

1.2 Representação gráfica 8

CAPÍTULO 2: ANÁLISE DISCRIMINANTE DE FISHER 12

2.1 Análise Discriminante de Fisher: fundamentos e método 12

2.1.1 Discriminação de Fisher para o caso de vários grupos 14

2.2 Algumas propriedades interessantes das variáveis discriminantes 19

2.3 Cálculo de coeficientes para a função discriminante de Fisher segundo o tipo de padronização utilizada. 20

2.3.1 Cálculo do Coeficiente Bruto 20

2.3.2 Cálculo do Coeficiente Dentro 21

2.3.3 Cálculo do Coeficiente Total 21

2.4 Cálculo das correlações entre as variáveis originais e as variáveis discriminantes, usando diferentes padronizações 22

2.4.1 Cálculo da estrutura de correlação Total 22

2.4.2 Cálculo da estrutura de correlação Dentro dos grupos 24

2.4.3 Cálculo da estrutura de correlação Entre os grupos 25

CAPÍTULO 3: ASPECTOS IMPORTANTES A SEREM RESSALTADOS NA ANÁLISE DISCRIMINANTE DE FISHER 27

3.1 Consequências de algumas transformações nos dados 27

3.2 Decomposição da variação total quando os dados estão padronizados 29

3.2.1 Tratando o problema de mudanças na escala, no caso de Análise Discriminante de Fisher 30

3.3 Dificuldades com a interpretação dos coeficientes 33

3.4 Distância de Mahalanobis 34

CAPÍTULO 4: REPRESENTAÇÃO GRÁFICA NO ESPAÇO DAS VARIÁVEIS 38

4.1 Distância entre indivíduos 38

4.2 Distância entre variáveis 39

4.2.1 Significação algébrica da distância entre variáveis 39

4.2.2 Distância à origem 45

4.2.3 Correlações entre as variáveis originais e as discriminantes 47

4.3 Construção dos Círculos das Correlações para um exemplo. 48

4.4 Indivíduos suplementares no caso da Análise Discriminante de Fisher 54

CAPÍTULO 5: ESTUDO ATRAVÉS DE SIMULAÇÃO 56

5.1 Simulação de amostras bivariadas 57

5.1.1 Esquema de simulação para o caso bivariado 60

5.1.2 Recortes para as análises 71

5.2 Simulação de amostras trivariadas 73

CAPÍTULO 6: RESULTADOS 79

6.1 Comportamento dos coeficientes da Função Linear Discriminante de Fisher e das correlações entre as variáveis originais e discriminantes variando as correlações "DENTRO" e "ENTRE"os grupos, para duas variáveis originais 81

6.1.1 Comparando a influência de mudanças na correlação "DENTRO" dos grupos, mantendo fixa a correlação "ENTRE" 81

6.1.2 Comparando a influência de mudanças na correlação "ENTRE" os grupos, mantendo fixa a correlação "DENTRO" 109

6.1.3 Para comparar a influência da variância das variáveis originais, mantendo fixas as correlações, "DENTRO" e "ENTRE", entre as variáveis originais 127

6.2 Comportamento dos coeficientes da Função Linear Discriminante de Fisher e das correlações entre as variáveis originais e discriminantes no caso de três variáveis 149

6.2.1 Para comparar a influência da correlação "DENTRO" dos grupos, entre as variáveis originais e as discriminantes 149

6.2.2 Para comparar a influência de diferentes matrizes de correlação "ENTRE". 161

6.3 Conclusões derivadas dos resultados das simulações 172

6.3.1 Conclusões sobre o comportamento dos coeficientes 172

6.3.2 Conclusões sobre o comportamento das correlações entre as variáveis originais e as discriminantes 174

6.3.3 Procedimento para analise dos Círculos das Correlações 175

6.3.4 Conclusões sobre o significado dos Círculos das Correlações 177

CAPÍTULO 7: APLICAÇÃO EM UM EXEMPLO COM DADOS REAIS 178

CAPÍTULO 8: ROTEIRO PARA ANÁLISE 240

CAPÍTULO 9 :CONCLUSÕES 242

CAPÍTULO 10: REFERÊNCIAS BIBLIOGRÁFICAS 243

CAPÍTULO 11: BIBLIOGRAFIA 245

ANEXO 1: Saída completa da Análise Discriminante de Fisher para os dados dos Insetos A1-1

ANEXO 2: Exemplos dos programas para os casos 1 e 4 bivariados A2-1

ANEXO 3: Exemplo do programa para o caso 1 trivariado A3-1

LISTA DE FIGURAS

Figura 1.1.1: Representação hipotética das direções aproximadas para a primeira Componente Principal y, quando interessa a variabilidade total, a primeira Componente Principal w, quando interessa a variabilidade dentro dos grupos e z a primeira Função Linear Discriminante de Fisher quando interessa a variabilidade entre os grupos.....2

Figura 1.2.1: Gráfico dos escores dos indivíduos, diferenciados pela espécie qual pertencem, no plano formado pelas duas primeiras variáveis discriminantes. (NOTA: 1 obs. escondida).....9

Figura 4.2.1: Esboço da relação entre a correlação e a distância entre variáveis, no espaço das variáveis.....41

Figura 4.2.2: Exemplo hipotético que mostra o significado geométrico da distância entre variáveis. d é a distância entre as duas variáveis; c_{jj} a projeção da variável j na j' e $h=1$ a distância da origem ao circulo.....41

Figura 4.2.3: Esboço geométrico da média e o desvio padrão para o exemplo da variável X_i.....43

Figura 4.2.4: Esboço que mostra a geometria da amostra para o caso de duas variáveis.....44

Figura 4.2.5: Exemplo hipotético de Círculo das Correlações.....46

Figura 4.3.1: Círculo das Correlações para os valores das correlações Entre os grupos, na Tabela 4.2.1.....50

Figura 4.3.2: Círculo das Correlações para os valores das correlações dentro dos grupos , na Tabela 4.2.1.....51

Figura 4.3.3: Círculo das Correlações para os valores das correlações Totais, na Tabela 4.2.1.....52

Figura 5.1.1: Contorno de densidade constante para uma distribuição normal bivariada com $\sigma_{11}=\sigma_{22}$ e $\sigma_{12}>0$. Os semi-eixos da elipse têm os valores $a=c\sqrt{\lambda_1^*}$ e $b=c\sqrt{\lambda_2^*}$ 58

Figura 5.1.2: Esquema das simulações para o caso bivariado, a direita os números servirão para identificar cada caso. 61

Figura 5.1.3: Exemplo das configurações correspondentes ao esquema de simulações para os casos bivariados com $\sigma_{11}=\sigma_{22}$. Por linha (por coluna) as correlações dentro (entre) dos grupos são iguais. As diagonais correspondem aos casos com correlações entre e dentro iguais. 63

Figura 5.1.4: Exemplo do cálculo das posições dos grupos para conseguir a configuração desejada 65

Figura 5.1.5 Cálculos para alocar as médias dos três grupos no caso da configuração dentro alta positiva e entre zero 66

Figura 5.1.6 Gráfico das três amostras geradas e das posições na qual foram colocados os grupos. 67

Figura 5.1.7: Fragmento do programa em SAS que mostra o procedimento de perturbação das médias para evitar o problema de colinearidade. 69

Figura 5.2.1: Esquema das simulações para o caso de três variáveis 75

Figura 5.2.2: Gráficos de dispersão para os casos 1, 2 e 3, das simulações com de três variáveis ... 76

Figura 5.2.3: Gráficos de dispersão para os casos 4, 5 e 6, das simulações com de três variáveis ... 77

Figura 5.2.4: Gráficos de dispersão para os casos 7, 8 e 9, das simulações com de três variáveis ... 78

Figura 6.1.1: Gráfico das variáveis discriminantes para o caso 1 83

Figura 6.1.2: Gráfico das variáveis discriminantes para o caso 2 92

Figura 6.1.3: Gráfico das variáveis discriminantes para o caso 8 95

LISTA DE TABELAS

Tabela 1.1.1: Coeficientes Discriminantes de Fisher 5

Tabela 3.2.1: Coeficientes Discriminantes de Fisher para diferentes transformações nos dados. 32

Tabela 4.2.1. Correlações entre as variáveis originais e as discriminantes para os dados dos insetos com e sem transformação; relativas à estrutura de correlação: Dentro, Entre e Total.. 47

Tabela 4.3.1: Magnitude dos autovalores para os dados dos insetos 49

Tabela 6.1.1: Coeficientes quando a correlação entre é fixada para ser alta positiva (Casos 1-7-13) 82

Tabela 6.1.2: CÍRCULOS DAS CORRELAÇÕES para os casos 1-7-13 85

Tabela 6.1.3: Correlações entre originais e discriminantes (Casos 1-7-13)..... 86

Tabela 6.1.4: Coeficientes (Casos 4-10-16)..... 88

Tabela 6.1.5: Correlações entre originais e discriminantes (Casos 4-10-16)..... 89

Tabela 6.1.6: CÍRCULOS DAS CORRELAÇÕES para os casos 4-10-16 90

Tabela 6.1.7: Coeficientes (Casos 2-8-14) 93

Tabela 6.1.8: CÍRCULOS DAS CORRELAÇÕES para os casos 2-8-14 94

Tabela 6.1.9: Correlações entre originais e discriminantes (Casos 2-8-14)..... 96

Tabela 6.1.10: Coeficientes (Casos 5-11-17)..... 97

Tabela 6.1.11: Correlações entre originais e discriminantes (Casos 5-11-17)..... 98

Tabela 6.1.12: CÍRCULOS DAS CORRELAÇÕES para os casos 5-11-17 99

Tabela 6.1.13: Coeficientes (Casos 3-9-15) 100

Tabela 6.1.14: Correlações entre originais e discriminantes (Casos 3-9-15) 101

Tabela 6.1.15: CÍRCULOS DAS CORRELAÇÕES para os casos 3-9-15 104

Tabela 6.1.16: Coeficientes (Casos 6-12-18) 105

Tabela 6.1.17: Correlações entre originais e discriminantes (Casos 6-12-18) 106

Tabela 6.1.18: CÍRCULOS DAS CORRELAÇÕES para os casos 6-12-18 108

Tabela 6.1.19: Coeficientes (Casos 1-2-3) 109

Tabela 6.1.20: Correlações entre originais e discriminantes (Casos 1-2-3) 110

Tabela 6.1.21: CÍRCULOS DAS CORRELAÇÕES para os casos 1-2-3 112

Tabela 6.1.22: Coeficientes (Casos 4-5-6) 113

Tabela 6.1.23: Correlações entre originais e discriminantes (Casos 4-5-6) 114

Tabela 6.1.24: CÍRCULOS DAS CORRELAÇÕES para os casos 4-5-6 115

Tabela 6.1.25: Coeficientes (Casos 7-8-9) 116

Tabela 6.1.26: Correlações entre originais e discriminantes (Casos 7-8-9) 117

Tabela 6.1.27: CÍRCULOS DAS CORRELAÇÕES para os casos 7-8-9 118

Tabela 6.1.28: Coeficientes (Casos 10-11-12) 119

Tabela 6.1.29: Correlações entre originais e discriminantes (Casos 10-11-12) 120

Tabela 6.1.30: CÍRCULOS DAS CORRELAÇÕES para os casos 10-11-12 121

Tabela 6.1.31: Coeficientes (Casos 13-14-15) 122

Tabela 6.1.32: Correlações entre originais e discriminantes (Casos 13-14-15) 123

Tabela 6.1.33: CÍRCULOS DAS CORRELAÇÕES para os casos 13-14-15 124

Tabela 6.1.34: Coeficientes (Casos 16-17-18) 125

Tabela 6.1.35: Correlações entre originais e discriminantes (Casos 16-17-18) 126

Tabela 6.1.36: CÍRCULOS DAS CORRELAÇÕES para os casos 16-17-18 127

Tabela 6.1.37: Coeficientes (Casos 1 e 4) 128

Tabela 6.1.38: Correlações entre originais e discriminantes (casos 1 e 4) 129

Tabela 6.1.39: CÍRCULOS DAS CORRELAÇÕES para os casos 1 e 4. 130

Tabela 6.1.40: Coeficientes (Casos 2 e 5) 131

Tabela 6.1.41: Correlações entre originais e discriminantes (casos 2 e 5) 131

Tabela 6.1.42: CÍRCULOS DAS CORRELAÇÕES para os casos 2 e 5. 132

Tabela 6.1.43: Coeficientes (Casos 3 e 6) 133

Tabela 6.1.44: CÍRCULOS DAS CORRELAÇÕES para os casos 3 e 6. 134

Tabela 6.1.45: Correlações entre originais e discriminantes (casos 3 e 6) 135

Tabela 6.1.46: Coeficientes (Casos 7 e 10) 135

Tabela 6.1.47: Correlações entre originais e discriminantes (casos 7 e 10) 136

Tabela 6.1.48: CÍRCULOS DAS CORRELAÇÕES para os casos 7 e 10. 137

Tabela 6.1.49: Coeficientes (Casos 8 e 11) 138

Tabela 6.1.50: CÍRCULOS DAS CORRELAÇÕES para os casos 8 e 11. 139

Tabela 6.1.51: Correlações entre originais e discriminantes (casos 8 e 11) 140

Tabela 6.1.52: Coeficientes (Casos 9 e 12) 140

Tabela 6.1.53: Correlações entre originais e discriminantes (casos 9 e 12) 141

Tabela 6.1.54: CÍRCULOS DAS CORRELAÇÕES para os casos 9 e 12. 141

Tabela 6.1.55: Coeficientes (Casos 13 e 16) 143

Tabela 6.1.56: CÍRCULOS DAS CORRELAÇÕES para os casos 13 e 16. 143

Tabela 6.1.57: Correlações entre originais e discriminantes (casos 13 e 16) 144

Tabela 6.1.58: Coeficientes (Casos 14 e 17) 144

Tabela 6.1.59: CÍRCULOS DAS CORRELAÇÕES para os casos 14 e 17. 145

Tabela 6.1.60: Correlações entre originais e discriminantes (casos 14 e 17) 146

Tabela 6.1.61: Coeficientes (Casos 15 e 18) 146

Tabela 6.1.62: Correlações entre originais e discriminantes (casos 15 e 18) 147

Tabela 6.1.63: CÍRCULOS DAS CORRELAÇÕES para os casos 15 e 18. 147

Tabela 6.2.1: Coeficientes (Casos 1-4-7) 151

Tabela 6.2.2: Correlações entre originais e discriminantes (Casos 1-4-7) 151

Tabela 6.2.3: CÍRCULOS DAS CORRELAÇÕES para os casos 1-4-7 153

Tabela 6.2.4: Coeficientes (Casos 2-5-8) 155

Tabela 6.2.5: Correlações entre originais e discriminantes (Casos 2-5-8) 156

Tabela 6.2.6: CÍRCULOS DAS CORRELAÇÕES para os casos 2-5-8 157

Tabela 6.2.7: Coeficientes (Casos 3-6-9) 159

Tabela 6.2.8: CÍRCULOS DAS CORRELAÇÕES para os casos 3-6-9 160

Tabela 6.2.9: Correlações entre originais e discriminantes (Casos 3-6-9) 161

Tabela 6.2.10: Coeficientes (Casos 1-2-3) 162

Tabela 6.2.11: Correlações entre originais e discriminantes (Casos 1-2-3) 163

Tabela 6.2.12: CÍRCULOS DAS CORRELAÇÕES para os casos 1-2-3 164

Tabela 6.2.13: Coeficientes (Casos 4-5-6) 165

Tabela 6.2.14: Correlações entre originais e discriminantes (Casos 4-5-6) 166

Tabela 6.2.15: CÍRCULOS DAS CORRELAÇÕES para os casos 4-5-6 167

Tabela 6.2.16: Coeficientes (Casos 7-8-9) 168

Tabela 6.2.17: Correlações entre originais e discriminantes (Casos 7-8-9) 169

Tabela 6.2.18: CÍRCULOS DAS CORRELAÇÕES para os casos 7-8-9 170

LISTA DE EQUAÇÕES

(2.1.1) 13

(2.1.2) 13

(2.1.3) 13

(2.1.4) 13

(2.1.5) 14

(2.1.6) 14

(2.1.7) 14

(2.1.8) 14

(2.1.9) 15

(2.1.10) 15

(2.1.11) 15

(2.1.12) 15

(2.1.13) 15

(2.1.14) 16

(2.1.15) 16

(2.1.16) 16

(2.1.17) 16

(2.1.18) 16

(2.1.19) 17

(2.1.20) 17

(2.1.21) 17

(2.1.22) 17

(2.1.23) 18

(2.2.1) 19

(2.2.2) 19

(2.2.3) 19

(2.2.4) 19

(2.2.5) 19

(2.2.6) 19

(2.2.7) 19

(2.2.8) 20

(2.2.9) 20

(2.3.1) 21

(2.3.2) 21

(2.3.3) 21

(2.3.4) 22

(2.4.1) 23

(2.4.2) 23

(2.4.3) 24

(2.4.4) 24

(2.4.5) 25

(2.4.6) 25

(3.1.1) 28

(3.1.2) 28

(3.2.1) 29

(3.2.2) 29

(3.2.3) 29

(3.2.4) 29

(3.2.5) 30

(3.2.7) 30

(3.2.8) 31

(3.4.1) 36

(3.4.2) 36

(3.4.3) 36

(4.1.1) 39

(4.2.1) 39

(4.2.2) 40

(4.2.3) 40

(4.2.4) 40

(4.2.5) 40

(4.2.6) 42

(4.2.7) 42

(4.2.8) 42

(4.2.9) 43

(4.2.10) 43

(4.2.11) 43

(4.2.12) 44

(4.2.13) 45

(4.2.14) 45

(4.2.15) 45

(5.1.1) 57

(5.1.2) 59

(5.1.3) 59

Capítulo 1: Introdução

1.1 Análise Discriminante de Fisher

A Análise Discriminante de Fisher é uma técnica multivariada para trabalhar dados que apresentam uma estrutura de grupo conhecida a *priori*. Exploratória por natureza, esta técnica têm como objetivos principais encontrar funções discriminantes, ou novos eixos, que descrevam gráfica e algebricamente a separação entre os grupos assim como derivar regras que permitam classificar um novo indivíduo em um dos grupos conhecidos, minimizando o erro de má classificação.

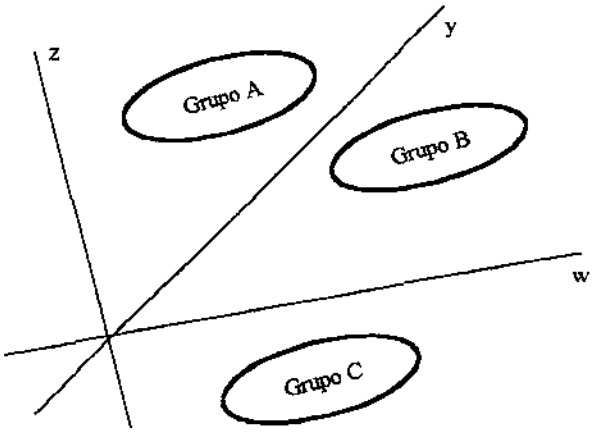
Este trabalho está dirigido ao leitor que, detendo já conhecimentos de Análise Discriminante de Fisher e Análise em Componentes Principais, tenha se defrontado com dificuldades na interpretação das Funções Lineares Discriminantes de Fisher.

1.1.1 Ilustrando geometricamente o primeiro eixo discriminante

Consideremos três grupos hipotéticos de dados bivariados cujas representações gráficas estão na Figura 1.1.1. As três direções y, w e z aí representadas são as seguintes:

Quando ignorada a estrutura de grupos, e considerando o conjunto de dados como um todo, y é a primeira Componente Principal, aquela que descreve a direção de máxima variabilidade do conjunto completo dos dados. Quando o que nos interessa é a variabilidade dentro dos grupos, que neste exemplo é igual para todos, esta pode ser vista na direção da Componente Principal w. Por

último, quando é a separação entre os grupos o mais importante, z, que é a primeira Função Linear Discriminante de Fisher, é o melhor índice para representar a direção onde melhor se enxerga a separação dos grupos.



Fonte: KRZANOWSKI, W.J., 1988. p. 306.

Figura 1.1.1: Representação hipotética das direções aproximadas para a primeira Componente Principal y, quando interessa a variabilidade total, a primeira Componente Principal w, quando interessa a variabilidade dentro dos grupos e z, a primeira Função Linear Discriminante de Fisher, quando interessa a variabilidade entre os grupos.

1.1.2 Motivação para o estudo dos coeficientes da Função Linear Discriminante de Fisher

Graças aos avanços da informática atualmente existem pacotes estatísticos que fazem todos os cálculos necessários para a Análise Discriminante de Fisher. Estes cálculos, dependendo do tamanho da matriz de dados e do número de grupos, seriam complicados de desenvolver sem a ajuda da computação.

Nosso interesse principal será o estudo da Análise Discriminante de Fisher, aprofundando, em particular, na compreensão da Função Linear Discriminante de Fisher e na influência que têm neste tipo de dados a presença não só de estrutura de variâncias e covariâncias total senão de uma estrutura de variâncias e covariâncias dentro e entre os grupos.

1.1.2.1 Um exemplo da Análise Discriminante de Fisher com dados reais.

Quando nos enfrentamos com um problema real muitas são as questões que nos inquietam. A título de ilustração vejamos um exemplo baseado em dados que foram trabalhados em um artigo de LINDSEY, J.C., HERZBERG, A.M. e WATTS, D.G.[1], 1987. Neste artigo é proposto um método de *cluster* que usa projeções multidimensionais e uma variante menor dos gráficos de quartil-quartil, para ser usado na análise exploratória de dados, na detecção de *outliers* e para fins de classificação. Eles assumem que não sabem a que grupo pertence cada inseto e o objetivo foi conseguir agrupar insetos similares. Eles chegam à conclusão da existência de três *clusters* para estes dados, um outlier, que é o inseto 14 e quando partem para classificação chegam a só um inseto mal classificado (o 13 que fica localizado como integrante da espécie 1 e não da 2 que é a qual realmente ele pertence); os dados podem ser vistos a seguir. LUBISCHEW, 1962, trabalhou o mesmo exemplo com muito mais variáveis; os dados trabalhados no artigo de LINDSEY são uma versão abreviada deles.

Espécie 1				**Espécie 2**				**Espécie 3**			
Inseto	X_1	X_2	X_3	**Inseto**	X_1	X_2	X_3	**Inseto**	X_1	X_2	X_3
1	191	131	53	**11**	186	107	49	**21**	158	141	58
2	185	134	50	**12**	211	122	49	**22**	146	119	51
3	200	137	52	**13**	201	144	47	**23**	151	130	51
4	173	127	50	**14**	242	131	54	**24**	122	113	45
5	171	128	49	**15**	184	108	43	**25**	138	121	53
6	160	118	47	**16**	211	118	51	**26**	132	115	49
7	188	134	54	**17**	217	122	49	**27**	131	127	51
8	186	129	51	**18**	223	127	51	**28**	135	123	50

[1] Esta referência foi tomada de HAND D.J., et. al., 1994

Espécie 1				Espécie 2				Espécie 3			
Inseto	X_1	X_2	X_3	**Inseto**	X_1	X_2	X_3	**Inseto**	X_1	X_2	X_3
9	174	131	52	**19**	208	125	50	**29**	125	119	51
10	163	115	47	**20**	199	124	46	**30**	130	120	48

Fonte: Dados originais tomados de HAND, D.J. et al, 1994. p. 254-255

No exemplo três variáveis foram medidas em cada um de 10 insetos para três espécies diferentes de um tipo de inseto cujo nome científico é *Chaetocnema*[2]. A variável X_1 mede a largura da primeira junção do primeiro tarso, X_2 a largura da primeira junção do segundo tarso e X_3 a largura máxima do *aedegus*[3]; as três variáveis em microns. A seguir, na Tabela 1.1.1, mostram-se os coeficientes para a FLDF para estes dados, saído de um dos softwares disponíveis, neste caso o SAS.

Na Análise Discriminante de Fisher, as Funções Lineares Discriminantes são autovetores associados a autovalores em ordem decrescente da matriz $\Sigma_D^{-1}\Sigma_E$, onde Σ_D é a matriz de variâncias dentro dos grupos e Σ_E a matriz de variâncias e covariâncias entre os grupos. Estas funções aparecem como soluções de maximização de formas quadráticas segundo o critério postulado em cada técnica, neste caso a direção que maximiza a variabilidade entre sujeita à padronização da variabilidade dentro. Assim temos duas matrizes de variâncias e covariâncias resultantes da decomposição da matriz de variâncias e covariâncias total, portanto, mais de uma norma ou distância estatística. Daí surgem diferentes padronizações e os distintos coeficientes calculados na Tabela 1.1.1.

Os coeficientes Brutos, na Tabela 1.1.1, veremos que são os autovetores da matriz $\mathbf{S}_D^{-1}\mathbf{S}_E$, sendo $\mathbf{S}_D$, a estimativa amostral da matriz de variâncias e

[2] O *Chaetocnema* é um tipo de besouro, conhecido também como "besouro da batata doce" (*sweetpotato flea*)
[3] Aedegus está se referindo ao órgão copulativo masculino.

covariâncias populacional dentro dos grupos[4] $\Sigma_D = \frac{1}{N-1}\sum_{i=1}^{s}\sum_{j=1}^{N_i}(x_{ij}-\mu_i)(x_{ij}-\mu_i)'$ [5], e S_E a estimativa amostral da matriz de variâncias e covariâncias populacional entre os grupos $\Sigma_E = \frac{1}{N-s}\sum_{i=1}^{s}(\mu_i-\bar{\mu})(\mu_i-\bar{\mu})'$, sendo $\bar{\mu} = \frac{1}{s}\sum_{i=1}^{s}\bar{\mu}_i$ a média geral para as s populações. Por outro lado os coeficientes total e dentro serão aqueles padronizados pela variância total e a variância dentro, respectivamente. Mais detalhes do cálculo destes coeficientes serão vistos na seção 2.3.

Tabela 1.1.1: Coeficientes Discriminantes de Fisher

Variação[6]	Variáveis	Brutos		Padronizados	
		Can1	Can2	Can1	Can2
Total	X_1	4,1350	-0,0883	4,1350	-0,0883
	X_2	-0,5071	1,1769	-0,5071	1,1769
	X_3	-1,1701	-0,2819	-1,1701	-0,2819
Dentro	X_1	1,7784	-0,0380	1,7784	-0,0380
	X_2	-0,5009	1,1624	-0,5009	1,1624
	X_3	-1,1655	-0,2808	-1,1655	-0,2808
Bruto	X_1	0,1261	-0,0026	4,1350	-0,0883
	X_2	-0,0571	0,1327	-0,5071	1,1769
	X_3	-0,3946	-0,0950	-1,1701	-0,2819

[4] Sendo N_i e μ_i o tamanho da população i e a média para a população i respectivamente, com $i = 1,2,...,s$, e $N = \sum_{i=1}^{s} N_i$ o tamanho da população total; s é a quantidade de populações.

[5] Na equação (2.1.3) podemos ver explicitamente esta matriz para os dados amostrais.

[6] No software SAS, os coeficientes Total, Dentro e Bruto são chamados de *Total-Sample Standardized Canonical Coefficients*, *Pooled Within Canonical Coefficient* e *Raw Coefficients*, respectivamente.

1.1.2.2 Qual é o coeficiente usado para construir as Funções Discriminantes de Fisher?

Diante dos distintos coeficientes na Tabela 1.1.1 surgem as perguntas usuais: Qual a contribuição das variáveis na separação dos grupos? É X_3 quem têm mais influência por ter o coeficiente Bruto de maior magnitude ou é X_1 por ter o coeficiente Total e Dentro maiores? Qual a interpretação correta dos coeficientes? Qual o impacto da variação dentro e entre os grupos? Qual o efeito da correlação dentro e entre os grupos? Porque muda o coeficiente Bruto no caso dos dados padronizados? Devemos padronizar? Qual a influência nos coeficientes das mudanças de escala nos dados?[7]

Sabe-se, da teoria da Análise Discriminante de Fisher, que para construir as Funções Lineares Discriminantes de Fisher, são usados os autovetores de $S_D^{-1}S_E$. JONHSON, R.A., WICHERN, D.W., 1992 e MARDIA, K.V., KENT, J.T., BIBBY, J.M., 1979, entre outros, explicam isto detalhadamente. Desta forma podemos resumir dizendo que eles coincidem com o que é o vetor de coeficientes *Bruto*, usado para a construção das Funções Lineares Discriminantes de Fisher.

Atenção especial merece o processo de interpretação. A influência de uma variável na separação dos grupos poderá simplesmente ser explicada pela magnitude do valor absoluto do coeficiente *Bruto*? Este é um ponto que não fica claro pois nem sempre na prática aquela variável cujo coeficiente *Bruto* é maior é a que mais influencia na separação dos grupos.

Não aparece com freqüência na literatura referência aos coeficientes *Dentro* e *Total*. Qual a importância ou a contribuição deles na hora de interpretar as Funções Lineares Discriminantes de Fisher?

KRZANOWSKI, W.J., 1988, propõe calcular um novo coeficiente, resultado da multiplicação do coeficiente *Bruto* pela diagonal da matriz S_{pooled}. Como a

[7] Para ficar mais claro estabeleceremos que quando falarmos de correlações entre originais e discriminantes; Dentro, Entre ou Total; ou dos coeficientes; Bruto, Dentro ou Total; escreveremos com *letra itálica*, i.e., *Dentro*, *Entre* ou *Total* e *Bruto*, *Dentro* ou *Total*, para o caso dos coeficientes e as correlações entre as variáveis originais e as discriminantes, respectivamente. Por outro lado, se estivermos falando das variâncias ou correlações das variáveis originais as identificaremos como "ENTRE", "DENTRO" ou "TOTAL".

variabilidade entre grupos, que é de fato o que interessa nesta técnica, esta sendo analisada relativamente à variabilidade dentro dos grupos (ver equação (2.1.18)), a magnitude do coeficiente *Bruto* pode depender não de variabilidade grande entre os grupos senão de variabilidade muito pequena dentro dos grupos para uma determinada variável. A padronização proposta por KRZANOWSKI poria as variáveis em um patamar de comparabilidade com respeito à variabilidade dentro dos grupos. Este coeficiente ao qual o autor está se referindo não é outro senão aquele que chamamos de *Dentro*, cuja magnitude, agora desprovida do efeito da variabilidade dentro, seria proporcional à importância ou influência da variável na separação dos grupos.

Por outro lado STEVENS,.J., 1992, expõe que, para fins de interpretação, existem dois métodos: (1) examinar os coeficientes padronizados, referindo-se aos valores resultantes da multiplicação do coeficiente *Bruto* de cada variável pelo desvio padrão respectivo; e (2) examinar as correlações entre cada variável discriminante e as originais.

Ainda que STEVENS apresente uma pequena seção dedicada a como interpretar as Funções Discriminantes de Fisher, algo difícil de encontrar nas referências consultadas, os métodos aos quais ele se refere não ficam claros e muitas dúvidas aparecem logo após uma leitura cuidadosa. Sua orientação "padronizar" os coeficientes não diz respeito a qual fonte de variação: *Dentro*, *Entre*, *Total*. Igualmente no caso das correlações, qual estrutura de correlação deve ser analisada: *Dentro*, *Entre*, Total? Além disso ele reconhece que a magnitude destes coeficientes ou correlações será proporcional à ordem de importância das variáveis, no entanto, no parágrafo acima vimos que nem sempre um coeficiente grande será sinônimo de que esta variável em particular tenha uma importância grande na separação dos grupos, e então poderíamos estar errados se cegamente olhamos a magnitude sem analisar mais nada.

Quanto à importância do uso das correlações entre originais e discriminantes, para fins de interpretação, STEVENS dá numerosas referências de autores que argumentaram a favor do seu uso pelas vantagens observadas

quanto à estabilidade das correlações para amostras de tamanho grande, e a indicação direta de quais variáveis originais estão perto das variáveis discriminantes. Também diz que estes coeficientes de correlação são coeficientes parciais, pelo que, para cada um, têm sido tirado o efeito do resto das variáveis.

Por outro lado, LUBISCHEW, 1962, fala de calcular primeiramente um Coeficiente de Discriminação, *K*, proposto por ele neste artigo, para decidir quais variáveis devem ser escolhidas dentro de um conjunto, de tal forma a garantir a escolha daquelas que mais vão contribuir na discriminação dos grupos. Mas ele diz que isto não é suficiente pois devem ser consideradas também as correlações entre as variáveis. Para isto ele, assumindo variáveis com variâncias iguais, propõe a construção de elipses e círculos trabalhando com as correlações entre grupos e dentro deles. Nestes diagramas, que como ele mesmo diz, podem ser considerados frouxamente como a representação de grupos de pontos observados, ou mais precisamente como sendo os contornos dentro dos quais esperamos encontrar uma proporção fixa de observações, ele consegue representar como as correlações "ENTRE" e "DENTRO" influenciam na separação dos grupos, embora o que ele esteja representando aqui não seja o plano discriminante.

1.2 Representação gráfica

A saída numérica dos resultados nem sempre consegue por si mesma dar uma idéia clara do que está acontecendo com os dados, a menos que o usuário do método seja experiente e saiba traduzir a saída digital, e mesmo assim restaria a dificuldade em transmitir a informação.

Tendo-se ampliado muito o uso da Análise Discriminante nas ciências sociais e econômicas, nos estudos demográficos; na engenharia, no reconhecimento de padrões; na indústria, com o controle multivariado de processos, o estudo das redes neurais, etc., resulta importante para os usuários desta técnica conseguir um bom entendimento dos resultados.

É por isto que outro dos objetivos deste trabalho é acrescentar saída gráfica à Análise Discriminante de Fisher usando as estruturas de correlação, *Dentro*, *Entre* e *Total* entre as variáveis originais e discriminantes.

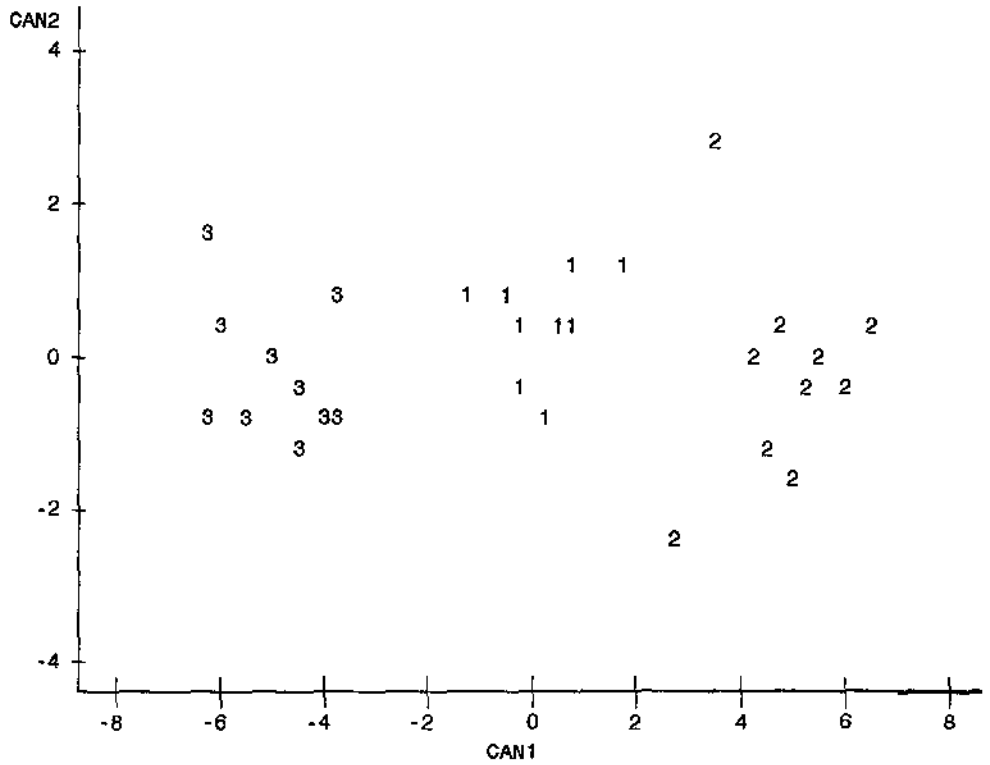


Figura 1.2.1: Gráfico dos escores dos indivíduos, diferenciados pela espécie a qual pertencem, no plano formado pelas duas primeiras variáveis discriminantes. (NOTA: 1 obs. escondida).

Na Figura 1.2.1 pode ser vista a representação gráfica usualmente disponível nos pacotes estatísticos, no caso da Análise Discriminante de Fisher.. para o exemplo dos insetos. Os escores dos indivíduos são colocados no plano formado pelas duas primeiras variáveis discriminantes. Este gráfico dá uma idéia da maior separação possível entre os grupos, conseguida pela Análise Discriminante de Fisher.

Fica muitas vezes a dúvida sobre qual dos coeficientes foi usado para produzir o gráfico, e se as variáveis foram tomadas padronizadas ou não. Além disso, a visibilidade das comparações entre indivíduos por grupo, sendo

necessária, não é suficiente; interessa descrevê-las ao longo do comportamento das variáveis analisadas.

Ao construir o gráfico das funções discriminantes, ver Figura 1.2.1, é conveniente centralizar as variáveis originais para conseguir um gráfico mais claro e simples. Porém, quando trabalhamos com os dados brutos, é aconselhável centrar as variáveis na hora de fazer o gráfico. Quando os dados já estão centrados ou padronizados o gráfico pode ser feito diretamente. Em todos os casos, i.e., sem interessar se os dados sofreram ou não algum tipo de transformação, o coeficiente usado para calcular os escores do indivíduos para gerar o gráfico é o *Bruto*.

O trabalho de LEBART, L., MORINEAU, A. e PIRON, M., 1995 agrega à saída gráfica das Componentes Principais os chamados Círculos das Correlações, colocando as correlações entre as variáveis originais e as Componentes Principais em um círculo de raio um, permitindo extrair informações importantes do comportamento das variáveis. Também aqui, eles tratam o referente a inserir na análise indivíduos que não participam do desenvolvimento algébrico do método por terem caraterísticas peculiares, eles são chamados de pontos suplementares.

Desta forma tratar-se-á de desenvolver neste trabalho as mesmas representações, usando o Círculo das Correlações, para as correlações entre as variáveis originais e as discriminantes, quando calculadas a partir de cada uma das estruturas de variâncias e covariâncias: *Total*, *Dentro* e *Entre*. O objetivo é enriquecer a saída da Análise Discriminante, permitindo extrair mais informações das variáveis originais e ajudando na interpretação dos coeficientes. Além disto usar a definição de ponto suplementar para introduzir na Análise Discriminante de Fisher o tratamento de indivíduos que não participam da derivação dos eixos canônicos. As razões para que indivíduos sejam tratados como suplementares são variadas e de natureza do conhecimento substantivo de cada problema. De forma geral pretende-se examinar o comportamento dos pontos suplementares no cenário derivado pelos indivíduos ativos.

Por exemplo na análise comparativa de regiões distintas: uma pode ser tomada como ativa e outra como suplementar. Em comparações longitudinais no estudo de perfis evolutivos: utiliza-se a técnica em um período e projeta-se os resultados de outro período como suplementar. Neste trabalho o estudo de híbridos no capítulo 7 ilustrará a aplicação e importância dos pontos suplementares.

Estas questões serviram de motivação para este trabalho. Nos propusemos como objetivo específico construir um roteiro que sirva de guia à interpretação da Análise Discriminante de Fisher, usando para isto não só a saída digital derivada dos cálculos como também uma nova proposta gráfica aqui construída.

Em resumo, os objetivos deste trabalho são: aprofundar na interpretação dos resultados da Análise Discriminante de Fisher, levando em conta a influência das diferentes estruturas de variâncias e covariâncias entre e dentro dos grupos; melhorar a visualização gráfica dos resultados e construir um roteiro com indicações precisas de como interpretar os resultados desta técnica.

Capítulo 2: Análise Discriminante de Fisher

Na prática muitos são os casos de dados com uma estrutura de grupo conhecida. Características medidas em casos tais como: espécies diferentes de animais ou plantas; turmas de estudantes diferenciadas por sucesso ou não nos estudos; características da voz ou da escrita para diferentes grupos de pessoas; sobre condições sociais ou econômicas das distintas regiões geográficas em um país; etc. são alguns exemplos onde a Análise Discriminante de Fisher ajuda, entre outras coisas, a estudar que variáveis medidas contribuem na diferenciação dos grupos e a construir funções lineares que permitem exibir a sua máxima separação.

2.1 Análise Discriminante de Fisher: fundamentos e método

Estamos na presença de um problema de Análise Discriminante quando temos n *indivíduos* ou observações, descritos por um conjunto de p *variáveis* quantitativas, $X_1, X_2, \cdots, X_p$, separados em s *grupos*, definidos a *priori* por uma variável indicadora. A matriz de dados pode ser escrita segundo a seguinte notação:

$$\mathbf{X}_{nxp} = \begin{bmatrix} X'_{11} \\ \vdots \\ X'_{1n_1} \\ \vdots \\ X'_{s1} \\ \vdots \\ X'_{sn_s} \end{bmatrix} \tag{2.1.1}$$

onde cada $X_{ij} = \begin{bmatrix} x_{ij1} \\ \vdots \\ x_{ijp} \end{bmatrix}$ é tal que i=1, 2, ..., s; j=1, 2, ..., n_s. O sub-índice i refere-se ao número de grupos e j identifica o número de indivíduos ou observações dentro de cada grupo. A matriz de dados pode ser escrita detalhadamente como:

$$\mathbf{X}_{nxp} = \begin{bmatrix} x_{111} & \cdots & x_{11p} \\ \vdots & & \vdots \\ x_{1n_1 1} & \cdots & x_{1n_1 p} \\ \vdots & & \vdots \\ x_{s11} & \cdots & x_{s1p} \\ \vdots & & \vdots \\ x_{sn_s 1} & \cdots & x_{sn_s p} \end{bmatrix} \tag{2.1.2}$$

Fisher converteu as observações multivariadas X_{ij}, i=1, 2, ..., s; j =1, 2, ..., n_s, em observações univariadas Y tal que as Y derivadas de uma população estão separadas o máximo possível das derivadas das outras populações. Um ponto importante é que Fisher não trabalha com o pressuposto de normalidade das populações, mas assume matrizes de variâncias e covariâncias dentro dos grupos iguais, $\Sigma_1 = \Sigma_2 = ... = \Sigma_s = \Sigma_D$, ao usar $\mathbf{S}_{pooled} = \mathbf{S}_D$, como a estimativa de Σ_D. A expressão desta matriz pode ser vista a seguir:

$$\left(\mathbf{S}_{pooled}\right)_{pxp} = \sum_{l=1}^{s} \frac{(n_l - 1)}{n - s} \mathbf{S}_l = \frac{1}{n-s} \sum_{l=1}^{s} \left(X - \bar{X}_l\right)\left(X - \bar{X}_l\right)' \tag{2.1.3}$$

$$s_{ij}^{D} = \frac{1}{n-s} \sum_{l=1}^{s} \sum_{t=1}^{n_l} \left(x_t - \bar{x}_{lt}\right)\left(x_t - \bar{x}_{lt}\right)' \tag{2.1.4}$$

Com $i, j = 1,2,...,p$ e vetores X e $\bar{X}$, definidos como na seção 2.1.1.

A Análise Discriminante de Fisher é uma técnica de muita utilidade na prática[1] pois ajuda a visualizar melhor a separação entre populações usando umas poucas combinações lineares das variáveis, reduzindo assim a dimensionalidade do problema.

2.1.1 Discriminação de Fisher para o caso de vários grupos

O interesse será mostrar o desempenho da Análise Discriminante de Fisher para uma amostra[2] de s grupos, por isto a notação será trabalhada em nível amostral e a seguir definimos as expressões mais usadas.

O vetor de médias para a amostra total será representado na forma:

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}_{px1}, \text{ com } \bar{x}_i = \sum_{i=1}^{s}\sum_{t=1}^{n_i} \frac{x_{iti}}{n_t} \tag{2.1.5}$$

com i=1, 2, ..., s; j =1, 2, ..., n_s. O vetor de variáveis e a média por grupo, respectivamente, como:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}_{pxi}, \quad \bar{X}_i = \begin{bmatrix} \bar{x}_{i1} \\ \vdots \\ \bar{x}_{ip} \end{bmatrix}_{px1}, \text{ com } \bar{x}_{ij} = \sum_{t=1}^{n_i} \frac{x_{tij}}{n_t} \tag{2.1.6}$$

Por outro lado (2.1.7) é a soma de quadrados e produtos cruzados entre os grupos e $SQ_D = (n-s)\mathbf{S}_{pooled}$ é a soma de quadrados e produtos cruzados dentro dos grupos.

$$SQ_E = \sum_{i=1}^{s} (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' \tag{2.1.7}$$

Consideremos a combinação linear (2.1.8)

$$Y_{lx1} = \mathbf{E}_{lxp} X_{px1} \tag{2.1.8}$$

A matriz de coeficientes $\mathbf{E}$ pode ser escrita como:

---

[1] Na prática é comum que não se cumpra a hipótese $S_1=S_2=....=S_s$, no entanto a Análise Discriminante de Fisher é robusta e pode ser aplicada e funcionar perfeitamente nestes casos. Ver GILBERT, E.S. 1969.

[2] O tratamento populacional para a Análise Discriminante de Fisher pode ser visto, por exemplo, em JOHNSON, R.A. e WICHERN, D.W., 1992; MARDIA, K.V., KENT, J.T. e BIBBY, J.M., 1979 ou LACHENBRUCH, P.A, 1975

$$\mathbf{E}_{kxp} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & & \vdots \\ e_{i1} & e_{i2} & \cdots & e_{ip} \\ \vdots & \vdots & & \vdots \\ e_{k1} & e_{k2} & \cdots & e_{kp} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1' \\ \mathbf{e}_2' \\ \vdots \\ \mathbf{e}_i' \\ \vdots \\ \mathbf{e}_k' \end{bmatrix} \tag{2.1.9}$$

Explicitamente as combinações lineares ficam da seguinte forma:

$$\begin{aligned} y_1 &= e_{11}X_1 + e_{12}X_2 + \ldots + e_{1p}X_p \\ y_2 &= e_{21}X_1 + e_{22}X_2 + \ldots + e_{2p}X_p \\ \vdots \quad & \quad \vdots \quad \quad \vdots \quad \quad \quad \vdots \\ y_k &= e_{k1}X_1 + e_{k2}X_2 + \ldots + e_{kp}X_p \end{aligned} \tag{2.1.10}$$

E cada uma destas combinações lineares pode ser escrita como

$$y_i = \mathbf{e}_i' X \tag{2.1.11}$$

com $i = 1,2,..,k$ e onde o vetor de coeficientes $\mathbf{e}_i'$ é a i-ésima linha da matriz (2.1.9). A combinação linear, da forma vista em (2.1.11) e condicionada à população de interesse, têm esperança igual a:

$$E(y_i) = \mathbf{e}_i' E(X \mid \pi_l) = \mathbf{e}_i' \overline{X}_l \tag{2.1.12}$$

para a combinação i e o grupo $l = 1,2,..,s$, e variância igual a

$$Var(y_i) = \mathbf{e}_i' \, Cov(X) \mathbf{e}_i = \mathbf{e}_i' \mathbf{S}_{pooled} \mathbf{e}_i \tag{2.1.13}$$

$\mathbf{S}_{pooled}$ como definida na expressão 2.1.3, e X e $\overline{X}$ vetores definidos como em (2.1.5) e (2.1.6).

Por outro lado, a média geral para as combinações lineares y_i, $i = 1,2,..,k$ é uma constante igual a $\overline{y} = \sum_{i=1}^{k} \overline{y}_i$, sendo $\overline{Y} = \begin{bmatrix} \overline{y}_1 \\ \vdots \\ \overline{y}_p \end{bmatrix}_{px1}$ o vetor de médias para (2.1.8).

No entanto, desde que as n observações estão divididas em s grupos, será conveniente descrever esta estrutura nas novas variáveis y. Para isto chamemos de y_{ij} o valor de y para a observação j no grupo i, com $i = 1,2,..,s$ e $j = 1,2,...,n_i$.

Desta forma a média do grupo i será igual a $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ e a média geral $\bar{y} = \frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{s} n_i \bar{y}_i$.

Para pesquisar se os s grupos estão bem diferenciados teríamos que particionar a soma de quadrados total, SQ_T , dos y_{ij} na soma de quadrados entre grupos, SQ_E , e soma de quadrados dentro dos grupos, SQ_D .

$$SQ_T = \sum_{i=1}^{s}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2 = \mathbf{Y'HY} = \mathbf{e'X'HXe} = \mathbf{e'S}_T\mathbf{e} \qquad (2.1.14)$$

Com $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{11'}$ sendo a matriz de centralização, cuja ajuda possibilita representações matriciais convenientes para as somas de quadrados e produtos cruzados,

$$SQ_E = \sum_{i=1}^{s} n_i\left(\bar{y}_i - \bar{y}\right)^2 = \sum_{i=1}^{s} n_i\left\{\left(\mathbf{e'}\overline{\mathbf{X}}_i - \overline{\mathbf{X}}\right)\right\}^2 = \mathbf{e'S}_E\mathbf{e} \qquad (2.1.15)$$

$$SQ_D = \sum_{i=1}^{s}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2 = \sum_{i=1}^{s}\mathbf{e'X}_i'\mathbf{H}_i\mathbf{X}_i\mathbf{e} = \mathbf{e'S}_D\mathbf{e} \qquad (2.1.16)$$

onde $\mathbf{H}_i = \mathbf{I}_{n_i x n_i} - \frac{1}{n_i}\mathbf{11'}$. Como é sabido, a relação entre estas somas de quadrados é a seguinte:

$$SQ_T = SQ_E + SQ_D \qquad (2.1.17)$$

O critério de Fisher é particularmente atrativo porque resulta da utilização da soma de quadrados e produtos cruzados entre grupos e a soma de quadrados dentro dos grupos. Isto significa que o interesse será maximizar a razão dada na expressão (2.1.18), para que a nova variável tenha maior variabilidade entre relativamente a dentro:

$$\frac{SQ_E}{SQ_D} = \frac{\mathbf{e'S}_E\mathbf{e}}{\mathbf{e'S}_D\mathbf{e}} \qquad (2.1.18)$$

onde $\mathbf{S}_E$ e $\mathbf{S}_D$ são as matrizes de variâncias e covariâncias entre e dentro dos grupos, respectivamente. A forma da matriz $\mathbf{S}_E$ pode ser vista em (2.1.19).

$$\mathbf{S}_E = \begin{bmatrix} s_{11}^E & s_{12}^E & \cdots & s_{1p}^E \\ s_{21}^E & s_{22}^E & \cdots & s_{2p}^E \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1}^E & s_{p2}^E & \cdots & s_{pp}^E \end{bmatrix}_{pxp} = \frac{1}{s-1}\sum_{l=1}^{s}(\overline{X}_l - \overline{X})(\overline{X}_l - \overline{X})' \qquad (2.1.19)$$

E um elemento genérico desta matriz é:

$$s_{ii}^E = \frac{1}{s-1}\sum_{l=1}^{s}(\overline{x}_{il} - \overline{x}_i)^2 \qquad (2.1.20)$$

$$s_{ij}^E = \frac{1}{s-1}\sum_{l=1}^{s}(\overline{x}_{il} - \overline{x}_i)(\overline{x}_{jl} - \overline{x}_i)' \qquad (2.1.21)$$

com $i, j = 1,2,...,p$.

Quanto maior for a razão (2.1.18), mais forte a indicação de variabilidade maior entre grupos do que dentro dos grupos. A mudança nos coeficientes das combinações lineares, $\mathbf{e}_i' = (e_{i1},\ e_{i2},\ \ldots,\ e_{ip})$, com $i = 1,2,....,k$, muda os valores de y_{ij} e, portanto, produzirá diferentes valores para (2.1.18). Maximizar o valor desta razão permitirá enxergar da melhor forma possível a diferença entre os grupos, e é por isto que o problema da Análise Discriminante de Fisher será achar os coeficientes que maximizem (2.1.18), o qual é um problema de otimização, puramente matemático. Para resolvê-lo devemos selecionar e tal que (2.1.18) seja máxima, portanto trabalharemos com esta expressão, apresentando a otimização na forma dos autovetores de $\mathbf{S}_D^{-1}\mathbf{S}_E$.

A matriz $\mathbf{S}_D^{-1}\mathbf{S}_E$ têm k autovalores diferentes de zero, onde $k = \min(s-1,\ p)$; esta constante k determina o número de funções discriminantes. Os autovetores associados, $l_1, l_2, \ldots, l_k$, padronizados tal que:

$$l_i'\mathbf{S}_D l_i = 1\ ,\ \forall i = 1,2,..,k \qquad (2.1.22)$$

imporão uma condição que permitirá obter uma solução única para este problema. Pode-se provar analiticamente que o vetor de coeficientes que maximiza a equação (2.1.18) é o primeiro autovetor de $\mathbf{S}_D^{-1}\mathbf{S}_E$. Isto quer dizer que

a melhor combinação linear $y_i = \mathbf{e}_i' X$ das variáveis originais que exalta a diferença entre os grupos têm como coeficientes os autovetores de $\mathbf{S}_D^{-1}\mathbf{S}_E$.

Simplificação da discriminação discriminante de Fisher para o caso de duas populações

O objetivo de Fisher, de maximizar a separação entre os grupos, reduz-se, no caso de dois grupos, a maximizar:

$$\phi = \frac{|\bar{y}_1 - \bar{y}_2|}{e'\mathbf{S}_{pooled}e} \qquad (2.1.23)$$

Os coeficientes da combinação linear serão escolhidos para que maximizem a razão entre o quadrado da distância entre as médias nas novas variáveis y e a estimativa ponderada da matriz de variâncias e covariâncias dentro dos grupos. O desenvolvimento da expressão (2.1.23) fica:

$$\phi^2 = \frac{\left(|\bar{y}_1 - \bar{y}_2|\right)^2}{e'\mathbf{S}_{pooled}^{-1}e} = \frac{(e'\bar{\mathbf{x}}_1 - e'\bar{\mathbf{x}}_2)^2}{e'\mathbf{S}_{pooled}^{-1}e} = \frac{(e'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{e'\mathbf{S}_{pooled}^{-1}e}$$

Usando o Lema de Maximização[2] de formas quadráticas[3] chegamos que

$$\max_e \frac{(e'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{e'\mathbf{S}_{pooled}^{-1}e} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D^2,$$

ou seja, o máximo é atingido quando $e = \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, sendo D^2 o quadrado da distância entre as duas médias amostrais, medidas em unidades de desvio padrão. A Função Linear Discriminante de Fisher fica então algebricamente explicitada por $Y = e'X = (\bar{x}_1 - \bar{x}_2)'\mathbf{S}_{pooled}^{-1}X$, para o caso de duas variáveis.

---

[2] Lema de Maximização: Seja $\mathbf{B}_{pxp}$ positiva definida e $\mathbf{d}_{pxp}$ um vetor dado, então para um vetor $\mathbf{x}_{px1}$

$\max_{\mathbf{x}\neq 0} \frac{(\mathbf{x}'\mathbf{d})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{d}'\mathbf{B}^{-1}\mathbf{d}$, atingindo o máximo quando $\mathbf{x}_{px1} = c\left(\mathbf{B}^{-1}\right)_{pxp}\mathbf{d}_{px1}$, $\forall c \neq o$.

[3] A prova do Lema de Maximização de formas quadráticas pode ser vista em JOHNSON, R.A. AND WICHERN, D.W., 1992. p.66.

2.2 Algumas propriedades interessantes das variáveis discriminantes

Ao maximizar (2.1.18), temos que diferenciá-la com respeito a e, igualando a zero obtemos $\mathbf{S}_E\mathbf{e}-\left(\frac{\mathbf{e}'\mathbf{S}_E\mathbf{e}}{\mathbf{e}'\mathbf{S}_D\mathbf{e}}\right)\mathbf{S}_D\mathbf{e}=0$. Quer dizer que, chamando $\lambda=\frac{\mathbf{e}'\mathbf{S}_E\mathbf{e}}{\mathbf{e}'\mathbf{S}_D\mathbf{e}}$, o referente valor de e deve satisfazer a equação

$$(\mathbf{S}_E-\lambda\mathbf{S}_D)\mathbf{e}=0 \tag{2.2.1}$$

Se escrevemos esta equação como

$$(\mathbf{S}_D^{-1}\mathbf{S}_E-\lambda\mathbf{I})\mathbf{e}=0 \tag{2.2.2}$$

podemos ver que ela toma a forma de uma equação característica que pode ser interpretada como λ sendo um autovalor do autovetor associado à matriz $\mathbf{S}_D^{-1}\mathbf{S}_E$.

No entanto, como λ é o máximo de (2.1.18), e deve ser o autovetor correspondente ao maior autovalor de $\mathbf{S}_D^{-1}\mathbf{S}_E$[4].

De (2.2.2) vemos que quaisquer dois pares autovalor-autovetor $(\lambda_i,\mathbf{e}_i)$ e $(\lambda_j,\mathbf{e}_j)$ satisfazem

$$\mathbf{S}_E\mathbf{e}_i=\lambda_i\mathbf{S}_D\mathbf{e}_i \tag{2.2.3}$$
$$\mathbf{S}_E\mathbf{e}_j=\lambda_j\mathbf{S}_D\mathbf{e}_j \tag{2.2.4}$$

Pré-multiplicando (2.2.3) por $\mathbf{e}_j'$ e (2.2.4) por $\mathbf{e}_i'$ temos

$$\mathbf{e}_j'\mathbf{S}_E\mathbf{e}_i=\lambda_i\mathbf{e}_j'\mathbf{S}_D\mathbf{e}_i \tag{2.2.5}$$
$$\mathbf{e}_i'\mathbf{S}_E\mathbf{e}_j=\lambda_j\mathbf{e}_i'\mathbf{S}_D\mathbf{e}_j \tag{2.2.6}$$

Mas $\mathbf{e}_j'\mathbf{S}_E\mathbf{e}_i'=\mathbf{e}_i'\mathbf{S}_E\mathbf{e}_j'$, pois ambas expressões são constantes e $\mathbf{S}_E$ é uma matriz simétrica. Disto segue que

$$\lambda_i\mathbf{e}_j'\mathbf{S}_D\mathbf{e}_i=\lambda_j\mathbf{e}_i'\mathbf{S}_D\mathbf{e}_j \tag{2.2.7}$$

No entanto $\mathbf{e}_j'\mathbf{S}_D\mathbf{e}_i=\mathbf{e}_i'\mathbf{S}_E\mathbf{e}_j$, pelos mesmos argumentos acima, tal que se $\lambda_i\neq\lambda_j$, a única possibilidade de que (2.2.7) se cumpra é que $\mathbf{e}_j'\mathbf{S}_D\mathbf{e}_i=\mathbf{e}_i'\mathbf{S}_E\mathbf{e}_j=0,\forall i\neq j$. Agora, $\mathbf{E}'S_D\mathbf{E}$ é uma matriz diagonal e, considerando a

[4] Ver KRZANOWSKI, 1988. p.295

restrição imposta para ter autovetores de norma um, sabemos que se cumpre que $\mathbf{E}'\mathbf{S}_D\mathbf{E}=\mathbf{I}$ [5]. Isto se traduz na esferização da dispersão dos indivíduos no plano discriminante. Com esta normalização, as variáveis discriminantes são construídas para ser não correlacionadas e de igual variância dentro dos grupos. Além disso, de (2.2.1) temos que $\mathbf{S}_E\mathbf{E}=\mathbf{S}_D\mathbf{E}\Lambda$, onde $\Lambda=\text{diag}(\lambda_i)$, então

$$\mathbf{E}'\mathbf{S}_E\mathbf{E}=\mathbf{E}'\mathbf{S}_D\mathbf{E}\Lambda=\Lambda \qquad (2.2.8)$$

Isto significa que as variáveis discriminantes também são não correlacionadas entre os grupos, só que agora isto ocorre na ordem decrescente da variância entre grupos. Ou seja, (2.2.8) se traduz por projeções dos indivíduos com variâncias decrescentes ao longo dos eixos.

Finalmente, desde que cumpre-se a relação (2.1.17), podemos escrever

$$\mathbf{E}'\mathbf{S}_T\mathbf{E}=\mathbf{E}'\mathbf{S}_E\mathbf{E}+\mathbf{E}'\mathbf{S}_D\mathbf{E}=\Lambda+\mathbf{I} \qquad (2.2.9)$$

Daqui concluímos que as variáveis discriminantes são não correlacionadas sobre a amostra total. A consequência disto é que mesmo na presença de correlações altas na estrutura de variâncias e covariâncias entre grupos, perde-se a visualização destas tendências no plano das funções lineares discriminantes de Fisher. Este padrão no comportamento entre grupos pode ser de interesse nas interpretações e veremos adiante (seção 4.3) que pode ser percebido pelas correlações entre as variáveis discriminantes de Fisher e as originais.

2.3 Cálculo de coeficientes para a função discriminante de Fisher segundo o tipo de padronização utilizada.

2.3.1 Cálculo do Coeficiente *Bruto*

Estes coeficientes, chamados de *Brutos*, não são mais que os autovetores e de $\mathbf{S}_E^{-1}\mathbf{S}_D$, derivados anteriormente; logo $\mathbf{E}'\mathbf{S}_D\mathbf{E}=\mathbf{I}$.

[5] Ver KRZANOWSKI, 1988. p.299

2.3.2 Cálculo do Coeficiente *Dentro*

O vetor de coeficientes chamado *Dentro*, $\mathbf{e}_D$, surge de uma padronização com respeito à variância dentro dos grupos. O cálculo pode ser feito a partir dos coeficientes brutos.

$$(\mathbf{e}_D)_{pxl} = [diag(\mathbf{S}_D)]^{1/2}\mathbf{e} \tag{2.3.1}$$

Exemplo: Usando os dados dos insetos, a matriz de variâncias e covariâncias "DENTRO" para estes dados é:

$$S_D = \begin{bmatrix} 198.670 & 83.133 & 32.818 \\ 83.133 & 76.726 & 15.415 \\ 32.818 & 15.415 & 8.722 \end{bmatrix} \tag{2.3.2}$$

O cálculo do vetor de coeficientes dentro fica:

$$\mathbf{e}_D = \begin{bmatrix} 14,095 & 0 & 0 \\ 0 & 8,759 & 0 \\ 0 & 0 & 2,953 \end{bmatrix} \begin{bmatrix} 0,126 & -0,003 \\ -0,057 & 0,133 \\ -0,395 & -0,095 \end{bmatrix} = \begin{bmatrix} 1,776 & -0,042 \\ -0,499 & 1,165 \\ -1,166 & -0,280 \end{bmatrix}.$$

Se compararmos com o coeficiente dentro, para o caso dos dados brutos, na Tabela 1.1.1, veremos que, a menos de algumas pequenas diferenças devido à aproximação dos valores utilizados[6], os coeficientes Dentro coincidem com os calculados anteriormente.

2.3.3 Cálculo do Coeficiente *Total*

O vetor de coeficientes chamado *Total*, $\mathbf{e}_T$, é obtido através de uma padronização usando a variação total

$$(\mathbf{e}_T)_{pxl} = [diag(\mathbf{S}_T)]^{1/2}\mathbf{e} \tag{2.3.3}$$

[6] As matrizes de variâncias e covariâncias, assim como as saídas dos coeficientes da Função Linear Discriminante de Fisher, com valores mais aproximados dos verdadeiros, podem ser vistas no ANEXO 1, caso seja de interesse reproduzir os cálculos.

Exemplo: Usando novamente os dados dos insetos, e sabendo que a matriz de variâncias e covariâncias "TOTAL" para estes dados é

$$S_T = \begin{bmatrix} 1073.941 & 85.896 & 9.458 \\ 85.896 & 78.644 & 15.253 \\ 9.458 & 15.253 & 8.792 \end{bmatrix}, \qquad (2.3.4)$$

o cálculo do vetor de coeficientes *Total*, seguindo a fórmula (2.3.3), fica:

$$\mathbf{e}_T = \begin{bmatrix} 32,771 & 0 & 0 \\ 0 & 8,868 & 0 \\ 0 & 0 & 2,965 \end{bmatrix} \begin{bmatrix} 0,126 & -0,003 \\ -0,057 & 0,133 \\ -0,395 & -0,095 \end{bmatrix} = \begin{bmatrix} 4,129 & -0,098 \\ -0,505 & 1,179 \\ -1,171 & -0,282 \end{bmatrix}$$

Se compararmos com o coeficiente *Total*, para o caso dos dados brutos, na Tabela 1.1.1, veremos que os coeficientes coincidem com os calculados anteriormente, a menos de pequenas diferenças devido à aproximação.

2.4 Cálculo das correlações entre as variáveis originais e as variáveis discriminantes, usando diferentes padronizações.

Temos também os valores das correlações entre as variáveis originais e discriminantes; eles dão uma informação, que agrega valor à interpretação dos eixos, necessária na descrição substantiva da separação dos indivíduos no plano discriminante.

2.4.1 Cálculo da estrutura de correlação *Total*.

Seja um vetor $\mathbf{b}' = (0,\ldots,0,1,0,\ldots,0)$ canônico. Sabendo que as variáveis discriminantes de Fisher são dadas por uma expressão do tipo da equação (2.1.11), o cálculo da estrutura de correlação *Total* entre as variáveis originais e as discriminantes é feito segundo a fórmula de correlação de Pearson.

$$corr_T(X_i, can_j) = corr_T(\mathbf{b}'\mathbf{X}, \mathbf{e}_j'\mathbf{X}) = \frac{\text{cov}_T(\mathbf{b}'\mathbf{X}, \mathbf{e}_j'\mathbf{X})}{\sqrt{\text{var}(\mathbf{b}'\mathbf{X})}\sqrt{\text{var}(\mathbf{e}_j'\mathbf{X})}} \quad (2.4.1)$$

onde $\text{cov}_T(\mathbf{b}'\mathbf{X}, \mathbf{e}_j'\mathbf{X}) = \mathbf{b}'S_T\mathbf{e}_j$, $\text{var}(\mathbf{b}'\mathbf{X}) = \mathbf{b}'\mathbf{S}_T\mathbf{b} = s_{ii}^T$ e $\text{var}(\mathbf{e}_j'\mathbf{X}) = \mathbf{e}_j'\mathbf{S}_T\mathbf{e}_j$, e portanto a expressão para a correlação *Total* entre as variáveis originais e as discriminantes fica:

$$corr_T(X_i, can_j) = \frac{\mathbf{b}'\mathbf{S}_T\mathbf{e}_j}{\sqrt{s_{ii}^T}\sqrt{\mathbf{e}_j'\mathbf{S}_T\mathbf{e}_j}} \quad (2.4.2)$$

Exemplo: Usando os dados dos insetos, e sabendo que a matriz de variâncias e covariâncias total é da forma de (2.3.4), podemos calcular o valor da correlação *Total* entre cada variável original e cada discriminante usando a equação (2.4.2). Sabendo que, por exemplo, para a correlação entre X_1 e $\mathbf{can}_1$, o numerador é igual a:

$$\mathbf{b}'\mathbf{S}_T\mathbf{e}_1 = (1,0,0)\begin{bmatrix} 1073{,}941 & 85{,}896 & 9{,}458 \\ 85{,}896 & 78{,}644 & 15{,}253 \\ 9{,}458 & 15{,}253 & 8{,}792 \end{bmatrix}\begin{pmatrix} 0{,}126179 \\ -0{,}057184 \\ -0{,}394650 \end{pmatrix},$$

e o denominador é

$$\sqrt{s_{11}^T}\sqrt{\mathbf{e}_1'\mathbf{S}_T\mathbf{e}_1} = \sqrt{1073{,}941}\sqrt{(0{,}126179, -0{,}057184, -0{,}394650)\begin{bmatrix} 1073{,}941 & 85{,}896 & 9{,}458 \\ 85{,}896 & 78{,}644 & 15{,}253 \\ 9{,}458 & 15{,}253 & 8{,}792 \end{bmatrix}\begin{pmatrix} 0{,}126179 \\ -0{,}057184 \\ -0{,}394650 \end{pmatrix}},$$

o valor desta correlação fica:

$$corr_T(X_1, can_1) = 0.9325749$$

Se compararmos com o coeficiente de correlação *Total*, para o caso dos dados brutos, na Tabela 4.2.1, veremos que o valor calculado anteriormente é efetivamente o mesmo, a menos de pequenas diferenças devido à aproximação.

Fazendo este mesmo cálculo pode-se obter os outros valores, que apresentamos a seguir:

	Can1	Can2
X_1	0.9325749	0.2285619
X_2	0.0087333	0.972375
X_3	-0.255804	0.3862524

2.4.2 Cálculo da estrutura de correlação *Dentro* dos grupos

O cálculo da estrutura de correlação *Dentro* entre as variáveis originais e as discriminantes é feito da seguinte forma:

$$corr_D(X_i, can_j) = corr_D(\mathbf{b'X}, \mathbf{e}'_j\mathbf{X}) = \frac{\text{cov}_D(\mathbf{b'X}, \mathbf{e}'_j\mathbf{X})}{\sqrt{\text{var}(\mathbf{b'X})}\sqrt{\text{var}(\mathbf{e}'_j\mathbf{X})}} \qquad (2.4.3)$$

onde $\text{cov}_D(\mathbf{b'X}, \mathbf{e}'_j\mathbf{X}) = \mathbf{b'}S_D\mathbf{e}_j$ e $\text{var}(\mathbf{b'X}) = \mathbf{b'S}_D\mathbf{b} = s_{ii}^D$, neste caso. Além disso, e lembrando da padronização vista em (2.1.22), $\text{var}(\mathbf{e}'_j\mathbf{X}) = \mathbf{e}'_j\mathbf{S}_D\mathbf{e}_j = 1$, e portanto a expressão para a correlação *Dentro* entre as variáveis originais e as discriminantes fica:

$$corr_D(X_i, can_j) = \frac{\mathbf{b'S}_D\mathbf{e}_j}{\sqrt{s_{ii}^D}} \qquad (2.4.4)$$

Exemplo: Usando os dados dos insetos, e sabendo que a matriz de variâncias e covariâncias "DENTRO" é da forma de (2.3.2), podemos calcular o valor da correlação *Dentro* entre cada variável original e cada discriminante, usando a equação (2.4.4). Sabendo que, por exemplo, para a correlação entre X_3 e $\mathbf{can}_2$, o numerador é igual a $\mathbf{b'S}_D\mathbf{e}_2 = (1,0,0)\begin{bmatrix} 198{,}670 & 83{,}133 & 32{,}818 \\ 83{,}133 & 76{,}726 & 15{,}415 \\ 32{,}818 & 15{,}415 & 8{,}722 \end{bmatrix}\begin{pmatrix} -0{,}002696 \\ 0{,}132715 \\ -0{,}095081 \end{pmatrix}$

e o denominador $\sqrt{s_{33}^D} = \sqrt{8{,}722}$. O valor desta correlação fica:

$$corr_D(X_3, can_2) = 0{,}3819561$$

Este valor é muito próximo ao calculado pelo software, para o caso dos dados brutos, na Tabela 4.2.1. Da mesma forma podemos ver que os outros valores são:

	Can1	Can2
X_1	0.5223339	0.5233787
X_2	0.0020813	0.9695859
X_3	-0.061857	0.3819561

2.4.3 Cálculo da estrutura de correlação *Entre* os grupos

O cálculo da estrutura de correlação *Entre* os grupos, para as variáveis originais e as discriminantes, é feito seguindo o mesmo padrão, isto é:

$$corr_E(X_i, can_j) = corr_E(\mathbf{b}'\mathbf{X}, \mathbf{e}_j'\mathbf{X}) = \frac{\text{cov}_E(\mathbf{b}'\mathbf{X}, \mathbf{e}_j'\mathbf{X})}{\sqrt{\text{var}(\mathbf{b}'\mathbf{X})}\sqrt{\text{var}(\mathbf{e}_j'\mathbf{X})}} \qquad (2.4.5)$$

onde $\text{cov}_E(\mathbf{b}'\mathbf{X}, \mathbf{e}_j'\mathbf{X}) = \mathbf{b}'\mathbf{S}_E\mathbf{e}_j$, $\text{var}(\mathbf{b}'\mathbf{X}) = \mathbf{b}'\mathbf{S}_E\mathbf{b} = s_{ii}^E$ e $\text{var}(\mathbf{e}_j'\mathbf{X}) = \mathbf{e}_j'\mathbf{S}_E\mathbf{e}_j$, neste caso. Portanto a expressão para a correlação *Entre*, entre originais e discriminantes, fica:

$$corr_E(X_i, can_j) = \frac{\mathbf{b}'\mathbf{S}_E\mathbf{e}_j}{\sqrt{s_{ii}^E}\sqrt{\mathbf{e}_j'\mathbf{S}_E\mathbf{e}_j}} \qquad (2.4.6)$$

Exemplo: Da mesma forma, nos casos da *Total* e da *Dentro* são calculadas as correlações *Entre*, sabendo que a matriz de correlações "ENTRE" para este exemplo dos insetos é $S_E = \begin{bmatrix} 1289{,}0100 & 12{,}3200 & -30{,}5900 \\ 12{,}3200 & 10{,}4533 & 1{,}3066 \\ -30{,}5900 & 1{,}3066 & 0{,}9733 \end{bmatrix}$.

Fazendo os cálculos pela equação (2.4.6) pode-se comprovar que elas ficam:

	Can1	Can2
X_1	0,9969387	0,0782001
X_2	0,0280632	0,9996069
X_3	-0,900403	0,4350629

Capítulo 3: Aspectos importantes a serem ressaltados na Análise Discriminante de Fisher.

Dependendo da técnica multivariada, transformações feitas nos dados podem ou não ter consequências que afetam em maior ou menor grau os resultados obtidos. Neste capítulo examinaremos a influência de algumas das transformações dos dados mais usadas e de como fica a decomposição da variação total quando trabalhamos com a matriz de correlações **R** e não com a matriz de variâncias e covariâncias **S**[1]. Assuntos referentes às dificuldades com a interpretação dos coeficientes e à distância que vemos quando referentes ao plano discriminante, também serão tratados.

3.1 Consequências de algumas transformações nos dados

O problema da invariância de escala é sempre uma discussão de interesse encontrada na literatura de análise multivariada. Dependendo da técnica multivariada a mudança de escala pode ou não afetar os resultados.

Uma transformação de escala pode ser do tipo $y_j = \mathbf{D}_T^{-1/2}(\mathrm{x_j} - \overline{\mathrm{x}})$, onde $\mathbf{D}_T = \mathrm{diag}(s_{jj}^2)$, com $s_{jj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ e *j=1,2,...,p,*, segundo Mardia[2]. Esta transformação leva cada variável a ter média zero e variância unitária e não é outra coisa que padronizar os dados tirando a influência das unidades de medida. As matrizes de variâncias e covariâncias **S** e de correlações **R**, para a amostra total, podem ser vistas explicitamente nas expressões (3.1.1) e (3.1.2).

[1] Trabalharemos todos os resultados em nível amostral.
[2] Ver MARDIA, K.V., KENT, J.T. e BIBBY, J.M, 1979. p.14.

$$\mathbf{S}_{pxp} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \quad (3.1.1)$$

$$\mathbf{R}_{pxp} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \quad (3.1.2)$$

No caso da Análise de Componentes Principais, quando feita nas variáveis padronizadas, os cálculos são desenvolvidos na matriz **R** da mesma forma que no caso da matriz **S**. No entanto os autovetores derivados de **S** não são em geral as mesmos de **R**[3]. Quando trabalharmos com a matriz de variâncias e covariâncias **S**, a variância das variáveis influencia na magnitude dos coeficientes, dominando aquelas variáveis com maior variância. A padronização afeta o tamanho dos coeficientes e portanto a importância relativa das variáveis na hora da interpretação; o poder explicativo de cada componente muda, pois muda também a magnitude dos autovalores.

O que acontece é que nas Componentes Principais a solução de um caso não se expressa como função algebricamente fechada da solução em outro. Isto quer dizer que a padronização traz consigo conseqüências que devem ser tomadas em conta, i.é, as Componentes Principais não são invariantes quanto à escala. Neste caso a padronização é recomendada em alguns casos[4], sendo mais usada na presença de variâncias distintas e escalas de medição diferentes nas variáveis originais.

Já a Análise Fatorial, diferentemente da Análise de Componentes Principais, "não é afetada" pelo re-escalamento das variáveis[5], pois são conhecidas as expressões fechadas que permitem relacionar o resultado obtido para as novas variáveis transformadas com aquele calculado para as variáveis originais.

[3] Ver MARDIA, K.V., KENT, J.T. e BIBBY, J.M., 1979. p.219. e JOHNSON, R.A. e WICHERN, D.W., 1992. p.365.
[4] Ver JOHNSON, R.A. e WICHERN, D.W., 1992. p.365
[5] Ver MARDIA, K.V., KENT, J.T. e BIBBY, J.M., 1979. p.257.

3.2 Decomposição da variação total quando os dados estão padronizados

A partir da matriz de variâncias e covariâncias total é conhecida a decomposição da soma de quadrados e produtos cruzados total SQ_T na soma de quadrados e produtos cruzados entre os grupos SQ_E e na soma de quadrados e produtos cruzados dentro dos grupos SQ_D, como foi vista na equação (2.1.17).

As matrizes SQ_T, SQ_E e SQ_D se diferenciam das respectivas matrizes de variâncias e covariâncias $\mathbf{S}_T$, $\mathbf{S}_E$ e $\mathbf{S}_D$ por uma constante que é precisamente o grau de liberdade respectivo para cada caso. Expressando as somas de quadrados de produtos cruzados em relação as matrizes $\mathbf{S}_T$, $\mathbf{S}_E$ e $\mathbf{S}_D$, a equação (2.1.17) se converte em:

$$(n-1)\mathbf{S}_T = (s-1)\mathbf{S}_E + (n-s)\mathbf{S}_D \tag{3.2.1}$$

Mas, o que acontece com a decomposição da variância total quando ela é dada pela matriz de correlações total $\mathbf{R}$?

Se padronizamos os dados, ou seja, se fazemos a transformação a seguir na matriz de dados,

$$\mathbf{XP} = \mathbf{D}_T^{-1/2}\mathbf{X}, \tag{3.2.2}$$

a equação (3.2.1) se converte em:

$$(n-1)\mathbf{D}_T^{-1/2}\mathbf{S}_T\mathbf{D}_T^{-1/2} = (s-1)\mathbf{D}_T^{-1/2}\mathbf{S}_E\mathbf{D}_T^{-1/2} + (n-s)\mathbf{D}_T^{-1/2}\mathbf{S}_D\mathbf{D}_T^{-1/2} \tag{3.2.3}$$

onde $\mathbf{D}_T^{1/2} = \text{diag}\left(1/\sqrt{s_{ii}^T}\right)$, com $i = 1,2,...,p$, pode ser representada explicitamente como:

$$\mathbf{D}_T^{-1/2} = \begin{bmatrix} \frac{1}{\sqrt{s_{11}^T}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{s_{22}^T}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{s_{pp}^T}} \end{bmatrix}_{pxp} \tag{3.2.4}$$

Do lado esquerdo da expressão 3.2.3 temos a matriz de correlações total $\mathbf{R}$, a menos de uma constante, decomposta em duas matrizes que chamaremos de $\mathbf{R}_E^*$ e $\mathbf{R}_D^*$.

$$(n-1)\mathbf{R}_T = (s-1)\mathbf{D}_T^{-1/2}\mathbf{S}_E\mathbf{D}_T^{-1/2} + (n-s)\mathbf{D}_T^{-1/2}\mathbf{S}_D\mathbf{D}_T^{-1/2} \qquad (3.2.5)$$

Esta decomposição não é equivalente a (3.2.1) uma vez que as somas de quadrados e produtos cruzados não estão padronizadas pelos desvios padrões que levariam às correlações entre e dentro, senão que, em todos os casos, a padronização é feita usando o desvio padrão total.

3.2.1 Tratando o problema de mudanças na escala, no caso de Análise Discriminante de Fisher

Na técnica de Correlações Canônicas é conhecido que a partir da matriz de variâncias e covariâncias ou a partir da matriz de correlações obtem-se as mesmas correlações canônicas. Pode-se passar dos coeficientes de um caso para outro por simples proporcionalidade, como aponta GITTINS R., 1985. Apesar da Análise Discriminante poder ser vista como caso particular da Análise canônica, este efeito será formalizado para a técnica da Análise Discriminante de Fisher.

As distintas padronizações das variáveis e o interesse na repercussão disto no cálculo dos coeficientes da Função Linear Discriminante de Fisher será explorado a seguir. De fato, se trabalharmos com os dados transformados segundo a padronização mostrada na equação 3.2.2, com o conseqüente efeito mostrado na equação 3.2.3, a formulação das Funções Discriminantes de Fisher muda um pouco pois neste caso a equação (2.1.18) se transformaria na maximização de:

$$F = \frac{\mathbf{v}'\mathbf{D}_T^{-1/2}SQ_E\mathbf{D}_T^{-1/2}\mathbf{v}/(s-1)}{\mathbf{v}'\mathbf{D}_T^{-1/2}SQ_D\mathbf{D}_T^{-1/2}\mathbf{v}/(n-s)} = \frac{\mathbf{v}'\mathbf{D}_T^{-1/2}\mathbf{S}_E\mathbf{D}_T^{-1/2}\mathbf{v}}{\mathbf{v}'\mathbf{D}_T^{-1/2}\mathbf{S}_D\mathbf{D}_T^{-1/2}\mathbf{v}} \qquad (3.2.7)$$

Portanto, trata-se agora de escolher os coeficientes $\mathbf{v}$ tal que maximizem a razão 3.2.7. Usando os mesmos critérios já vistos na seção 2.1.1, estes coeficientes serão os autovetores da matriz $\left(\mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{D}_T^{1/2}\right)\left(\mathbf{D}_T^{-1/2}\mathbf{S}_E\mathbf{D}_T^{-1/2}\right) = \mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{S}_E\mathbf{D}_T^{-1/2}$

Revisando o conceito de similaridade entre matrizes vemos que, segundo WATKINS, 1991, duas matrizes **A** e **B** são similares se existe uma matriz não singular **P** tal que $\mathbf{B}=\mathbf{P}^{-1}\mathbf{AP}$. Tomando $\mathbf{A}=\mathbf{S}_D^{-1}\mathbf{S}_E$, $\mathbf{B}=\mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{S}_E\mathbf{D}_T^{-1/2}$ e $\mathbf{P}=\mathbf{D}_T^{-1/2}$, podemos ver facilmente que as nossas matrizes são similares.

Resultados conhecidos demonstram que matrizes similares têm os mesmos autovalores e que **v** é autovetor de $\mathbf{S}_D^{-1}\mathbf{S}_E$ (i. é, $\mathbf{S}_D^{-1}\mathbf{S}_E(\mathbf{v})=\tilde{\lambda}(\mathbf{v})$) se e somente se $\mathbf{D}_T^{1/2}\mathbf{v}$ é autovetor de $\mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{S}_E\mathbf{D}_T^{-1/2}$ (i.é, $\mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{S}_E\mathbf{D}_T^{-1/2}\left(\mathbf{D}_T^{1/2}\mathbf{v}\right)=\tilde{\lambda}\left(\mathbf{D}_T^{1/2}\mathbf{v}\right)$). De fato, pois

$$\mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{S}_E\mathbf{D}_T^{-1/2}\left(\mathbf{D}_T^{1/2}\mathbf{v}\right)=\mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{S}_E\mathbf{v}=\mathbf{D}_T^{1/2}(\tilde{\lambda}\mathbf{v})=\tilde{\lambda}(\mathbf{D}_T^{1/2}\mathbf{v}) \qquad (3.2.8)$$

Em outras palavras, se **v** é o vetor de coeficientes *Brutos* da Função Linear Discriminante de Fisher para os dados originais (sem nenhuma transformação), então $\mathbf{D}_T^{1/2}\mathbf{v}$ é o vetor de coeficientes *Brutos* da Função Linear Discriminante de Fisher para estes dados, padronizados, e vice-versa. Isto indica que os autovetores da matriz $\mathbf{S}_D^{-1}\mathbf{S}_E$, para os dados brutos, têm uma relação fechada com os autovetores da matriz $\mathbf{D}_T^{1/2}\mathbf{S}_D^{-1}\mathbf{S}_E\mathbf{D}_T^{-1/2}$, para os dados padronizados.

As implicações deste resultado são muitas, por exemplo, pode-se passar de um resultado para o outro usando a transformação $\mathbf{D}_T^{1/2}\mathbf{v}$, como explicada anteriormente e, portanto, será indistinto trabalhar com a matriz **S** para os dados brutos ou com **R** para os dados padronizados, dado que os coeficientes da Função Linear Discriminante de Fisher para os dados brutos e padronizados estão relacionados por uma expressão conhecida.

Para ilustrar isto numericamente a seguir é continuado o exemplo dos dados dos insetos, citados anteriormente na seção 1.1.2. Na Tabela 3.2.1 é apresentada uma ampliação da Tabela 1.1.1, adicionando os resultados da Análise Discriminante de Fisher para a transformação de centralização dos dados. São apresentados os coeficientes discriminantes para os dados brutos,

centralizados e padronizados, derivados respectivamente a partir dos dados originais, ou seja, foi feita Análise Discriminante de Fisher a partir dos dados brutos, posteriormente estes foram centralizados e ainda padronizados, fazendo Análise Discriminante de Fisher da mesma forma que como feita para o primeiro caso dos dados brutos.

Tabela 3.2.1: Coeficientes Discriminantes de Fisher para diferentes transformações nos dados.

Variação	Variáveis	Tipo de transformação aplicada nos dados					
		Brutos		Centralizados		Padronizados	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	4,1350	-0,0883	4,1350	-0,0883	4,1350	-0,0883
	X_2	-0,5071	1,1769	-0,5071	1,1769	-0,5071	1,1769
	X_3	-1,1701	-0,2819	-1,1701	-0,2819	-1,1701	-0,2819
Dentro	X_1	1,7784	-0,0380	1,7784	-0,0380	1,7784	-0,0380
	X_2	-0,5009	1,1624	-0,5009	1,1624	-0,5009	1,1624
	X_3	-1,1655	-0,2808	-1,1655	-0,2808	-1,1655	-0,2808
Bruto	X_1	0,1261	-0,0026	0,1261	-0,0026	4,1350	-0,0883
	X_2	-0,0571	0,1327	-0,0571	0,1327	-0,5071	1,1769
	X_3	-0,3946	-0,0950	-0,3946	-0,0950	-1,1701	-0,2819

Fonte: Dados originais tomados para esta análise de HAND, D.J. et al, 1994. p. 254-255

A existência desta relação fechada entre os coeficientes da Função Linear Discriminante de Fisher quando trabalhamos com S ou com tem como consequência imediata, no nosso caso, a perda de interesse em trabalhar, nas simulações, com os casos de variâncias diferentes.

Além disso, as distâncias entre os vetores de médias ficam preservadas na passagem pelos coeficientes brutos para os dados centralizados ou padronizados, conforme desenvolvimento que será visto na seção 3.4, quando falarmos da distância de Mahalanobis.

3.3 Dificuldades com a interpretação dos coeficientes

As dificuldades na interpretação dos coeficientes na Análise Discriminante de Fisher decorrem da existência de dois produtos internos, duas variabilidades, presentes no caso de dados com estrutura de grupos: a dentro e a entre. Estes três coeficientes *Bruto*, *Dentro* e *Total* são calculados tomando em conta a decomposição da soma de quadrados e produtos cruzados total na soma de quadrados e produtos cruzados entre e a soma de quadrados e produtos cruzados dentro, decomposição que já foi vista na equação (2.1.17).

Na Tabela 3.2.1 foram mostrados os coeficientes da Função Linear Discriminante de Fisher: o *Bruto*, que são os autovetores de $\mathbf{S}_D^{-1}\mathbf{S}_E$; o *Total*, que são os coeficientes *Brutos* multiplicados pelo desvio padrão total, $\sqrt{s_{ii}^T}$, referentes às variáveis padronizadas segundo a variação total; e o *Dentro*, que são os *Brutos* multiplicados desta vez pelo desvio padrão dentro, $\sqrt{s_{ii}^D}$.

Como tratado na seção 2.1.1, a idéia da Análise Discriminante de Fisher baseia-se na maximização da razão entre a variância entre grupos e a variância dentro dos grupos, tendo como objetivo fundamental achar a direção no espaço multivariado que exiba o maior alongamento da nuvem de pontos constituída pelas médias dos grupos, ou seja, que preserva ao máximo as distâncias entre as médias dos grupos. O primeiro autovetor, $\mathbf{e}_1$, de $\mathbf{S}_D^{-1}\mathbf{S}_E$, exibe a dita direção, os outros autovetores, $\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_s$, nesta ordem, dão as direções nas quais a variação entre grupos, relativa à variação dentro, vai decrescendo conforme a ordem de importância do autovetor. Por outro lado os autovalores $\lambda_1, \lambda_2, \ldots, \lambda_s$ medem quanta variabilidade entre, relativa a dentro, é explicada por cada variável discriminante $y_i = \mathbf{e}_i' X$, i=1,...,s.

A interpretação de uma variável discriminante particular, y_i, pode ser feita similarmente à interpretação de uma componente principal, considerando o caso das componentes principais construídas sobre $\mathbf{R}$, de tal forma que coeficientes de maior magnitude serão os que determinaram a importância das variáveis.

No entanto devemos chamar a atenção ao fato de que no caso da Análise Discriminante de Fisher, a variabilidade entre grupos está sendo acessada relativamente à variabilidade dentro dos grupos, i.é, coeficientes grandes podem ser conseqüência tanto de variabilidade grande entre os grupos, neste caso a interpretação estaria sendo correta, quanto de pequena variabilidade dentro dos grupos para aquela variável, caso em que concluir sobre a importância das variáveis segundo o tamanho dos coeficientes seria enganoso.

Uma sugestão para lidar com este problema é dada em KRZANOWSKI, W.J., 1988; ele sugere considerar o coeficiente modificado $\mathbf{e}_i^* = (\mathbf{e}_{i1}^*, \mathbf{e}_{i2}^*, \ldots, \mathbf{e}_{ip}^*)$ tal que cada $\mathbf{e}_{ij}^* = \mathbf{e}_{ij}\sqrt{s_{jj}^D}$, onde s_{jj}^D é o i-ésimo elemento da diagonal de $\mathbf{S}_D$. Observa-semos com atenção este não é outro que o coeficiente *Dentro*, explicado na equação (2.3.1). Esta seria uma padronização a *posteriori* que leva todas as variáveis a ter variância unitária dentro dos grupos e as coloca em um patamar de comparabilidade que permite interpretar corretamente, pois agora a causa de mudança na magnitude dos coeficientes só seria devido à variabilidade entre os grupos.

3.4 Distância de Mahalanobis

Quando se viu a representação gráfica da Análise Discriminante de Fisher no exemplo dos insetos foi observado que os indivíduos ficam representados no plano discriminante com uma distância entre indivíduos e entre grupos de forma a conseguirmos enxergar o comportamento dos dados. Agora, qual é a métrica utilizada para calcular estas distâncias e representá-las neste plano? Como elas são calculadas?

Primeiramente notemos que a condição $\mathbf{E}'\mathbf{S}_D\mathbf{E} = \mathbf{I}$ mostra que a transformação da variável original $\mathbf{x}$ à variável discriminante $\mathbf{y}$ não é ortogonal. Isto significa em termos geométricos que os eixos das variáveis discriminantes, em geral, não formam ângulo reto entre eles, e que o espaço das variáveis discriminantes envolve alguma deformação com respeito ao espaço de referência

das variáveis originais, com alguns eixos sendo "pressionados" para ficarem mais juntos e outros sendo "afastados" para ficarem mais separados, conforme aponta KRZANOWSKI[6].

Não entanto, quando construímos um gráfico para as variáveis discriminantes, tal como o mostrado no Capítulo 1 referente aos dados dos insetos, construímos eixos ortogonais para as variáveis discriminantes. Uma justificativa para tal procedimento, segundo KRZANOWSKI, é que o objetivo é representar as distâncias entre grupos o mais próximo possível da representação exata.

Para entender mais claramente estas idéias consideremos dois grupos que diferem nas suas médias mas têm matrizes de variâncias e covariâncias dentro dos grupos iguais neste contexto estamos em condições de aplicar a Análise Discriminante de Fisher. Se todas as variáveis têm igual variância e são mutuamente não correlacionadas dentro dos grupos então a distância euclidiana entre os grupos será uma representação exata das diferenças entre eles ou, equivalentemente, da distância entre os dois grupos.

Entretanto, se algumas variáveis têm variância dentro dos grupos muito maior do que o resto, a influência delas deverá ser diminuída na hora de calcular a distância, desde que os indivíduos dentro do mesmo grupo difiram apreciavelmente nestas variáveis. Similarmente, se duas variáveis estão altamente correlacionadas dentro dos grupos, incluir ambas igualmente nos cálculos distorcerá a distância calculada entre os dois grupos, porque os valores da média do grupo tenderão a ser pequenos ou grandes juntos para as duas variáveis. *(KRZNOWSKI, W.J., 1988).*

Desta forma, melhor que usar a distância euclidiana $\left\{(\bar{x}_i - \bar{x}_j)'(\bar{x}_i - \bar{x}_j)\right\}^{1/2}$ entre as médias dos grupos, a forma mais apropriada de medir estas distâncias é $\left\{(\bar{x}_i - \bar{x}_j)'\mathbf{M}(\bar{x}_i - \bar{x}_j)\right\}^{1/2}$, sendo **M** a matriz dos pesos, a qual não é mais que uma

[6] Ver KRZANOWSKI, W.J., 1988. p.299

matriz apropriada que modifica a influência de cada variável nos cálculos. *(KRZNOWSKI, 1988).*

Mahalanobis, em 1936, propôs esta matriz **M** como a inversa da matriz de variâncias e covariâncias dentro dos grupos. Esta matriz leva em conta as covariâncias entre as variáveis quanto às variâncias diferenciadas. Agora podemos dizer que isto é uma medida adequada da distância entre duas populações quantitativas multivariadas.

Suponha que o número de discriminantes coincide com o número de variáveis originais ($k = p$), i. é, que a dimensionalidade do espaço das variáveis discriminantes coincide com a do espaço das originais. Se construímos eixos ortogonais das variáveis discriminantes para formar o espaço discriminante e o usamos para representar diferenças entre os grupos, então estamos tacitamente aceitando a distância euclidiana como a medida apropriada neste espaço. As médias dos grupos neste espaço podem ser escritas como $\bar{y}_i = \mathbf{E}\bar{X}$, com $i=1,2,...,s$. Então podemos escrever a distância euclidiana entre o grupo i e o grupo j no espaço discriminante como

$$d^2 = (\bar{y}_i - \bar{y}_j)'(\bar{y}_i - \bar{y}_j) = (\bar{x}_i - \bar{x}_j)'\mathbf{E}\mathbf{E}'(\bar{x}_i - \bar{x}_j) \qquad (3.4.1)$$

Do capítulo anterior sabemos que se cumpre que $\mathbf{E}'\mathbf{S}_D\mathbf{E} = \mathbf{I}$, pré-multiplicando por $\mathbf{E}$ e pós-multiplicando por $\mathbf{E}'$ chegamos a

$$\mathbf{E}\mathbf{E}' = (\mathbf{E}\mathbf{E}')\mathbf{S}_D(\mathbf{E}\mathbf{E}') \qquad (3.4.2)$$

Daqui vemos facilmente que $\mathbf{S}_D^{-1} = \mathbf{E}\mathbf{E}'$. Substituindo em (3.4.1) temos que

$$d^2 = (\bar{x}_i - \bar{x}_j)'\mathbf{S}_D^{-1}(\bar{x}_i - \bar{x}_j) \qquad (3.4.3)$$

Isto quer dizer que construindo o gráfico das variáveis discriminantes na forma descrita anteriormente, na verdade estamos produzindo um gráfico no qual a distância euclidiana ordinária entre os pontos que representam as médias dos grupos é igual à distância estatística ou de Mahalanobis entre os grupos

correspondentes. É claro que isto é verdadeiro só se a dimensão do gráfico é igual a $k = p$. Quando isto não acontece, i. é, a dimensão do gráfico é menor que o número de discriminantes, que neste caso é igual ao número de variáveis originais, o gráfico que se obtém dará a melhor aproximação da configuração real.

Capítulo 4: Representação geométrica no espaço das variáveis

O desempenho gráfico da Análise Discriminante de Fisher foi vista no Capítulo 1 com um exemplo da representação dos escores dos indivíduos no plano discriminante. Novas idéias para enriquecer esta representação gráfica surgiram de referências à saída de Componentes Principais, encontradas em LEBART, L., MORINEAU, A. e PIRON, M., 1995. A busca de outras alternativas para melhorar a saída gráfica da Análise Discriminante de Fisher é muito importante já que comumente representações geométricas facilitam a compreensão e enriquecem as conclusões.

A seguir exporemos algumas destas idéias para justificar como surgiram as propostas gráficas que serão mostradas neste capítulo para a Análise Discriminante. Trabalharemos referindo-nos a dois espaços, dos indivíduos e das variáveis.

4.1 Distância entre indivíduos

Dependendo da literatura, tanto o espaço dos indivíduos como o espaço das variáveis são definidos de diferentes formas. Neste trabalho o espaço dos indivíduos será definido da seguinte forma:

Espaço dos indivíduos: Temos a nuvem de n pontos, que representa os indivíduos, em $\mathbf{R}^p$, com as variáveis como os eixos. Os n vetores correspondem às linhas da matriz de dados $\mathbf{X}_{nxp}$. O vetor $\lfloor x_{i1},\ x_{i2},\ \ldots,\ x_{ip} \rfloor$, por exemplo, representa as coordenadas do i-ésimo ponto da nuvem, que nada mais é que as medições das p variáveis para o indivíduo i.

A distância entre os indivíduos i e i' é a distância euclidiana usual:

$$d^2(i,i') = (\mathbf{P}_i - \mathbf{P}_{i'})'(\mathbf{P}_i - \mathbf{P}_{i'}) \tag{4.1.1}$$

onde $\mathbf{P}_i = (x_{i1}\, x_{i2}, \ldots, x_{ip})$ é o vetor de coordenadas do indivíduo i no espaço dos indivíduos.

4.2 Distância entre variáveis

Por outro lado o espaço das variáveis será definido como a seguir:

Espaço das variáveis: As p variáveis também podem ser vistas como p pontos em um espaço n dimensional, $\mathbf{R}^n$. Cada coluna da matriz de dados $\mathbf{X}_{nxp}$ determina cada um dos pontos. A j-ésima coluna da matriz $\mathbf{X}$, consistente das n medições da variável j, $\begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$, determina o j-ésimo ponto.

Na próxima seção veremos o que acontece se calculamos a distância euclidiana usual entre duas variáveis, j e j', considerando a notação para as variáveis de z, referidas, neste caso, às variáveis padronizadas, i. é, $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{s_j\sqrt{n}}$.

4.2.1 Significação algébrica da distância entre variáveis

A distância euclidiana entre as variáveis j e j' fica

$$d^2(j,j') = \sum_{i=1}^{n}(z_{ij} - z_{ij'})^2 = \sum_{i=1}^{n} z_{ij}^{\,2} + \sum_{i=1}^{n} z_{ij'}^{\,2} - 2\sum_{i=1}^{n} z_{ij}\, z_{ij'} \tag{4.2.1}$$

como

$$s_{ii}^2 = \frac{1}{n-1}\sum_{l=1}^{n}(x_{il} - \bar{x}_i)^2 \qquad (4.2.2)$$

$$s_{ij} = \frac{1}{n-1}\sum_{l=1}^{n}(x_{il} - \bar{x}_l)(x_{jl} - \bar{x}_j) \qquad (4.2.3)$$

obtemos que $\sum_{i=1}^{n} z_{ij}{}^2 = \sum_{i=1}^{n} z_{ij'}{}^2 = 1$ e $\sum_{i=1}^{n} z_{ij} z_{ij'} = c_{jj'}$, sendo $c_{jj'}$ o coeficiente de correlação entre estas variáveis[1]. Segue daqui que a equação (4.2.1) se converte em

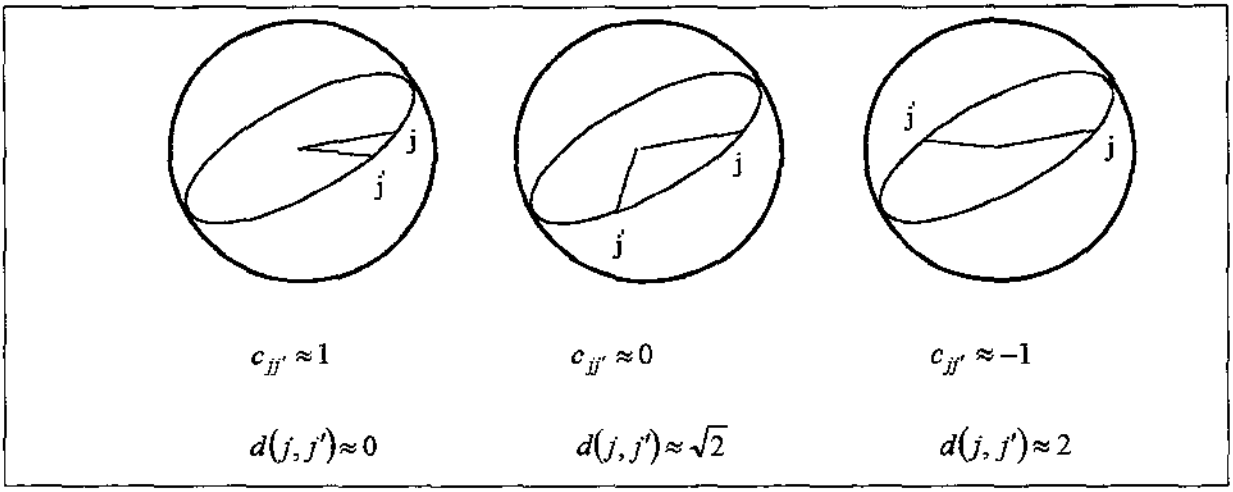
$$d^2(j, j') = 2(1 - c_{jj'}) \qquad (4.2.4)$$

A equação (4.2.4) significa que a relação entre as distâncias em $\mathbf{R}^n$ de duas variáveis j e j' é

$$d(j, j') = \sqrt{2(1 - c_{jj'})} \qquad (4.2.5)$$

onde $c_{jj'}$ é o coeficiente de correlação entre estas variáveis. Também é interessante observar que $0 \le d(j, j') \le 2$.

LEBART, L., MORINEAU, A. e PIRON, M., 1995[2] explicam com clareza a proximidade entre variáveis em termos de correlação. Achamos interessante incluir a Figura 4.2.1, tomada do livro destes autores, para um maior entendimento do que significa este resultado graficamente.



[1] Ver LEBART, L., MORINEAU, A. e PIRON, M., 1995. p. 38

Figura 4.2.1: Esboço da relação entre a correlação e a distância entre variáveis, no espaço das variáveis.

A equação (4.2.4) nos resulta familiar do ponto de vista estatístico pois duas variáveis fortemente correlacionadas positivamente estão uma perto da outra o que implica, como efetivamente podemos ver em (4.2.4), que $c_{jj'} \approx 1$; se pelo contrário, a correlação fosse forte, porém negativa, $c_{jj'} \approx -1$; por último, se as variáveis são ortogonais, $c_{jj'} \approx 0$, i.é, são não correlacionadas.

Uma outra forma de enxergar o resultado dado em (4.2.4) é geometricamente. Suponhamos que temos estas duas variáveis j e j' dentro de um círculo de raio um, o que é possível por elas terem variância um. Aproximadamente estamos nos referindo a um caso parecido ao representado na Figura 4.2.2.

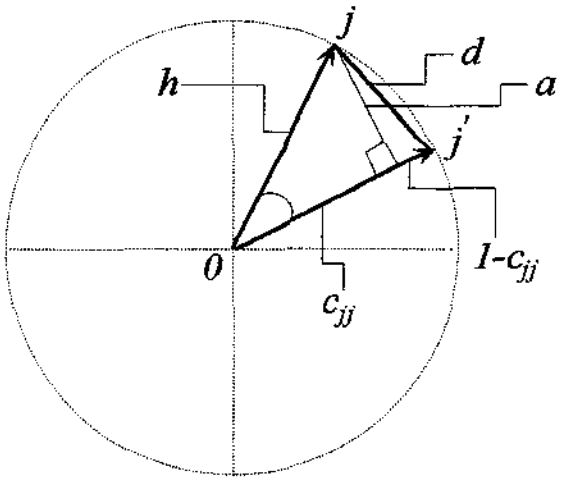


Figura 4.2.2: Exemplo hipotético que mostra o significado geométrico da distância entre variáveis. d é a distância entre as duas variáveis; $c_{jj'}$ a projeção da variável j na j' e $h=1$ a distância da origem ao círculo.

A altura a divide o triângulo isósceles, formado pelas distâncias da origem ao ponto j e ao ponto j' e um terceiro lado correspondente à distância entre as

[2] Ver LEBART, L., MORINEAU, A. e PIRON, M., 1995. p. 39

duas variáveis, em dois triângulos retângulos menores. Aplicando Pitágoras no triângulo retângulo maior temos que $1=c_{jj}^2+a^2$, daqui que $a^2=1-c_{jj}^2$. Usando este resultado calculamos a distância d entre as duas variáveis como

$$d^2=(1-c_{jj})^2+a^2=(1-c_{jj})^2+(1-c_{jj}^2)=1-2c_{jj}+c_{jj}^2+1-c_{jj}^2=2-2c_{jj}=2(1-c_{jj}).$$

e chegamos ao mesmo resultado da equação (4.2.4).

Por outro lado sabe-se que o cosseno do ângulo entre as duas variáveis não é outro que o próprio produto escalar, i.é, $\cos(j,j')=\langle j,j'\rangle$. Isto é devido a que o produto escalar entre dois vetores variáveis, j e j', n-dimensionais, pode ser definido como

$$\langle j,j'\rangle=X_j'X_{j'}=x_{1j}x_{1j'}+x_{2j}x_{2j'}+\ldots+x_{nj}x_{nj'} \qquad (4.2.6)$$

Sendo, por definição, o cosseno do ângulo entre elas, como dado na equação (4.2.7), sendo $L_{X_j}, L_{X_{j'}}$ o comprimento dos vetores X_j e $X_{j'}$, respectivamente.

$$\cos(\theta)=\frac{X_j'X_{j'}}{L_{X_j}L_{X_{j'}}} \qquad (4.2.7)$$

Como neste caso as duas variáveis são de variância unitária, a distância à origem é 1, i.e. $L_{X_j}=L_{X_{j'}}=1$, assim que (4.2.7) fica:

$$\cos(\theta)=X_j'X_{j'} \qquad (4.2.8)$$

A relação entre a correlação e o ângulo entre as variáveis pode ser obtida a partir da geometria da amostra. Isto é possível dado que muitas das expressões algébricas encontradas na análise multivariada podem ser relacionadas às noções geométricas de longitude, ângulo e volume.

Para explicar isto vamos supor duas das p variáveis, que seguiremos chamando de X_j e $X_{j'}$, no espaço n-dimensional e, pela limitação natural que possuímos para visualizar objetos em mais de três dimensões, suponhamos também que temos selecionado uma porção apropriada do espaço n-dimensional,

que contenha, em somente três dimensões, os dois vetores variáveis de interesse preservando as longitudes e ângulos.

Primeiramente, e para definir as médias destas variáveis, construamos o vetor auxiliar $1' = [1,1,\ldots,1]_{1xn}$. Este vetor faz o mesmo ângulo com cada uma das n coordenadas, sendo $(1/\sqrt{n})1$ o vetor unitário, norma um, na direção de 1; esta direção pode ser vista na Figura 4.2.4.

Segundo o conceito de projeção conhecido[3], a projeção do vetor X_i no vetor unitário $(1/\sqrt{n})1$ será

$$X_i'\left(\frac{1}{\sqrt{n}}1\right)\left(\frac{1}{\sqrt{n}}1\right) = \frac{1}{n}[(x_{1i},\ x_{2i},\ \ldots,\ x_{ni})1]1 = \frac{x_{1i}+x_{2i}+\ldots+x_{ni}}{n}1 = \bar{x}_i 1 \qquad (4.2.9)$$

A média amostral para a variável i, equação (4.2.10), diz sobre o múltiplo do vetor 1 necessário para levar a projeção da variável X_i ao vetor 1.

$$\bar{x}_i = \frac{X_i'}{n}1 \qquad (4.2.10)$$

Na Figura 4.2.3 podemos ver $\bar{x}_i$ e também o vetor, $X_i - \bar{x}_i 1$, perpendicular a esta projeção. Explicitamente ele têm a forma seguinte

$$X_i - \bar{x}_i 1 = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix} \qquad (4.2.11)$$

A componente j , $j = 1,2,\ldots,n$, em (4.2.11), representa o desvio da medição da variável i, no indivíduo j, com respeito à média amostral desta variável. Este vetor de desvio será chamado de d_i.

[3] (Projeção de $\mathbf{x}$ em $\mathbf{y}$) $= \frac{(\mathbf{x}'\mathbf{y})}{\mathbf{y}\mathbf{y}'}\mathbf{y} = \frac{(\mathbf{x}'\mathbf{y})}{L_{\mathbf{y}}}\frac{1}{L_{\mathbf{y}}}\mathbf{y}$, para $\mathbf{x}$ e $\mathbf{y}$ dois vetores quaisquer. Ver JOHNSON, R.A. e WICHERN, D.W., 1992. p.42.

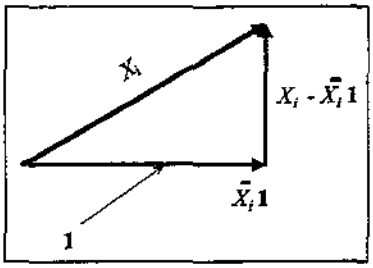


Figura 4.2.3: Esboço geométrico da média e o desvio padrão para o exemplo da variável X_i.

A Figura 4.2.4 mostra aproximadamente, para o caso das variáveis X_i e X_j, a posição das médias e os desvios padrão em três dimensões.

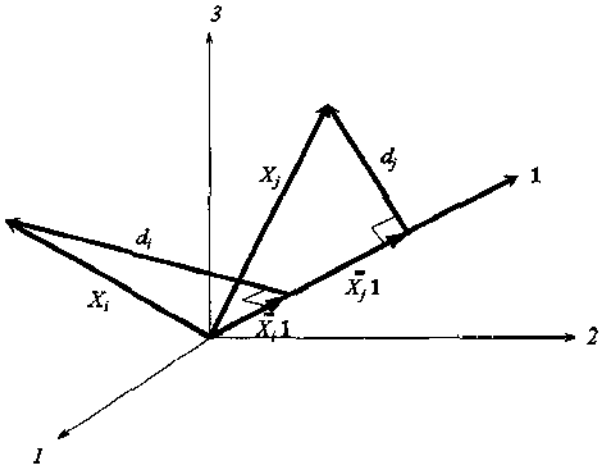


Figura 4.2.4: Esboço que mostra a geometria da amostra para o caso de duas variáveis.

O comprimento do vetor de desvio pode ser calculado como

$$L_{d_i}^2 = d_i' d_i = \sum_{l=1}^{n} (x_{il} - \bar{x}_i)^2 \qquad (4.2.12)$$

Ou seja, a soma ao quadrado dos desvios ou, como nos é mais familiar, a variância da variável X_i. No entanto cruzando os desvios das variáveis X_i e X_j, e somando, obtemos:

$$d_i' d_j = \sum_{l=1}^{n} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j) \tag{4.2.13}$$

Usando novamente Pitágoras é fácil chegar[4] a que o cosseno do ângulo entre as duas variáveis pode ser escrito como:

$$\cos(\theta_{ij}) = \frac{d_i' d_j}{\sqrt{d_i' d_i}\sqrt{d_j' d_j}} = \frac{\sum_{l=1}^{n} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{n} (x_{il} - \bar{x}_i)^2}\sqrt{\sum_{l=1}^{n} (x_{jl} - \bar{x}_j)^2}} \tag{4.2.14}$$

Esta expressão, se comparada com (4.2.2) e (4.2.3), pode ser escrita como

$$\cos(\theta_{ij}) = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} = r_{ij} \tag{4.2.15}$$

Tomando a conhecida forma do coeficiente de correlação amostral entre as variáveis X_i e X_j, r_{ij}. Portanto temos provado que o cosseno do ângulo entre as variáveis i e j coincide com o coeficiente de correlação amostral entre elas.

Estes resultados são válidos para quaisquer duas variáveis X_i, X_j, com $i, j = 1,2,\ldots,p$ e $i \neq j$.

4.2.2 Distância à origem

A distância da variável j à origem pode ser calculada como $d^2(0, j) = \sum_{i=1}^{n} z_{ij}^2 = 1$. Como todas as variáveis têm norma 1[5], i.e., variância unitária, elas podem ser vistas como vetores de uma hiperesfera de raio unitário centrada na origem, a Esfera das Correlações.

Representar as variáveis no plano, onde dois indivíduos são os eixos, significa fazer um corte na hiperesfera; qualquer corte deste tipo gera círculos de

[4] Ver JOHNSON, R.A. AND WICHERN, D.W., 1992. p.40-41

[5] Lembrando que resultados estão sendo mostrados a partir das variáveis padronizadas Z, porém, mais à frente, será explicado o que acontece no caso geral.

raio 1, que são os chamados Círculos das Correlações. Isto, que significa algebricamente a projeção das p variáveis em duas dimensões, distorce um pouco a projeção correta das variáveis; assim, dependendo do plano escolhido, elas ficarão melhor ou pior representadas. Além disso, neste caso, procurar direções que preservem a melhor representação das variáveis implica trabalhar com combinações lineares dos indivíduos, o que não têm sentido. Por isto as representações ficam restritas à tomada das variáveis discriminantes como eixos de referência conforme desenvolvimento na seção a seguir.

As conseqüências gráficas disto observa-se facilmente na distância à origem. Variáveis melhor representadas ficarão muito próximas à borda do círculo ou na própria borda enquanto as mal representadas ficarão longe da borda e mais próximas da origem. Um exemplo hipotético disto pode ser visto na Figura 4.2.5; neste caso as variáveis 1 e 2 estariam bem representadas e a 3 e 4 mal representadas.

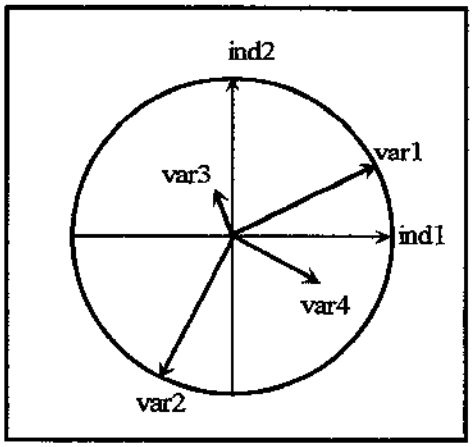


Figura 4.2.5: Exemplo hipotético de Círculo das Correlações.

Queremos observar que, diferentemente da Análise de Componentes Principais, na Análise Discriminante de Fisher o subespaço gerado pelas funções discriminantes com dimensão menor ou igual ao mínimo de (s, p-1), não tem razão para representar bem as variáveis.

Os Círculos de Correlação são construídos a partir das correlações *Total*, *Dentro* e *Entre* (calculadas das expressões 2.4.2, 2.4.4 e 2.4.6, respectivamente) das variáveis originas com as discriminantes. A estrutura de correlações das variáveis originas DENTRO dos grupos pode ser totalmente diferente da estrutura das mesmas variáveis com as variáveis discriminantes. No capítulo 6 serão examinandos casos em que essa estrutura é mantida e casos em que a mesma não é mantida.

4.2.3 Correlações entre as variáveis originais e as discriminantes

Ao falar da interpretação da distância entre variáveis, fizemos tudo para variáveis padronizadas, mas, como a padronização dos dados afeta as correlações entre as variáveis originais e as discriminantes? Será que as interpretações anteriores poderiam ser extrapoladas para o caso destas correlações, mesmo sem nenhuma transformação das variáveis?

É obvio que não afetam as correlações, como visto na análise exploratória dos dados dos insetos (Tabela 4.2.1), as correlações entre as variáveis originais e as discriminantes independem das transformações de centralização ou padronização nos dados; elas não mudam e portanto seria possível construir os Círculos das Correlações para tirar do gráfico mais informação da influência das variáveis.

Como trabalhamos com dados com estrutura de grupos e duas fontes de decomposição da variabilidade total: a variabilidade dentro dos grupos e a variabilidade entre os grupos, foram analisadas as correlações das variáveis originais com as discriminantes em cada um dos casos. Na Tabela 4.2.1 observa-se estas correlações para os dados dos insetos, e elas variam dependendo da estrutura de covariância referenciada. Qual a influência das variáveis quando analisadas estas correlações observando separadamente as estruturas de covariância dentro dos grupos, entre eles e total?

Tabela 4.2.1. Correlações entre as variáveis originais e as discriminantes para os dados dos insetos com e sem transformação, relativas à estrutura de correlação: *Dentro*, *Entre* e *Total*.

Estrutura de Correlação	Variáveis	Tipo de transformação aplicada nos dados					
		Brutos		Centralizados		Padronizados	
		Can1	Can2	Can1	Can2	Can1	Can2
Total[6]	X_1	0,9325	0,2285	0,9325	0,2285	0,9325	0,2285
	X_2	0,0087	0,9723	0,0087	0,9723	0,0087	0,9723
	X_3	-0,2557	0,3862	-0,2557	0,3862	-0,2557	0,3862
Dentro[7]	X_1	0,5223	0,5233	0,5223	0,5233	0,5223	0,5233
	X_2	0,0021	0,9695	0,0021	0,9695	0,0021	0,9695
	X_3	-0,0618	0,3819	-0,0618	0,3819	-0,0618	0,3819
Entre[8]	X_1	0,9969	0,0781	0,9969	0,0781	0,9969	0,0781
	X_2	0,0280	0,9996	0,0280	0,9996	0,0280	0,9996
	X_3	-0,9003	0,4350	-0,9003	0,4350	-0,9003	0,4350

Fonte: Dados originais tomados para esta análise de HAND, D.J. et al, 1994. p. 254-255

4.3 Construção dos Círculos das Correlações para um exemplo.

Conforme a proposta deste trabalho, e como exemplo dos resultados que serão vistos no capítulo 6, usaremos os resultados das correlações *Total, Dentro* e *Entre* dos dados dos insetos, Tabela 4.2.1, para construir os Círculos das Correlações para este exemplo. Estes gráficos podem ser vistos nas figuras a seguir. Em todos os casos os eixos representam as variáveis discriminantes, **Can1** e **Can2** , na ordem habitual, i.é, **Can1** na posição do eixo das abcissas e **Can2** no eixo das ordenadas.

No caso destes dados, os cálculos mostraram que a primeira variável discriminante explica 99,39% da variância entre os grupos. Os valores dos autovalores, na Tabela 4.3.1, exibem claramente que **Can1** é suficiente para explicar a separação entre os grupos, com um valor de λ_1 muito grande com

[6] Chamado no SAS de *Total Canonical Structure.*
[7] Chamado no SAS de *Pooled Within Canonical Structure*
[8] Chamado no SAS de *Between Canonical Structure.*

respeito a λ_2 e que a variável **Can2** não têm importância para discriminar os grupos.

Tabela 4.3.1: Magnitude dos autovalores para os dados dos insetos

Autovalores	Diferença	Proporção	Proporção Acumulada
17.5082	17.4009	0.9939	0.9939
0.1073	-	0.0061	1.0000

Esta excelente discriminação conseguida por **Can1** foi vista no gráfico[9] dos insetos no plano discriminante (Figura 1.2.1). Neste gráfico observa-se claramente que se projetamos cada indivíduo, distinguindo o grupo ao qual corresponde, no eixo correspondente a **Can1**, obteremos uma separação perfeita dos três grupos. Se projetássemos estes mesmos indivíduos em **Can2** só conseguiremos uma mistura desordenada dos 30 pontos.

Trata-se então de interpretar **Can1**, e o pesquisador interessado em descrever a diferença encontrada entre os grupos na Figura 1.2.1, retomando os coeficientes da Tabela 3.2.1[10], têm a considerar o seguinte:

No caso do coeficiente Bruto aplicado nos dados brutos, a indicação da ordem de importância das variáveis é $X_3, X_1 \, e \, X_2$. No caso do coeficiente *Bruto* derivado e aplicado aos dados padronizados é $X_1, X_3 \, e \, X_2$. A inversão das magnitudes relativas de importância se deve à pouca variabilidade de X_3; de fato a variância de X_3 é $s_{33} = 8{,}79$ enquanto que a de X_1 é $s_{11} = 1073{,}94$

O exame dos coeficientes *Dentro* e *Total* leva à ordem $X_1, X_3 \, e \, X_2$, tanto para os dados brutos quanto para os padronizados. No entanto, as magnitudes relativas dos coeficientes *Dentro* levam a julgar uma importância maior de X_1,

[9] O gráfico foi obtido aplicando correspondentemente o coeficiente *Bruto* nos dados brutos ou padronizados.
[10] Só serão de interesse os coeficientes para os dados brutos e padronizados pois os centralizados repetem o comportamento dos brutos, pelo que não serão considerados.

seguido de perto por X_3; o exame do *Total* amplifica a importância de X_1 relativamente a de X_3. Para os três coeficientes existe contraste entre X_1, com valores positivos, e X_2 *e* X_3, com valores negativos.

A pergunta de como X_1, X_2 *e* X_3 contribuem para a separação das espécies teria portanto respostas distintas a partir de cada um dos coeficientes. Vamos então examinar os Círculos das Correlações em cada caso, focalizando **can1**, por ser suficiente para ilustrar a separação entre espécies.

Um primeiro ponto importante e fácil de ser notado é que, tomando em consideração as boas representações de X_2 nos Círculos das Correlações (Figuras 4.3.1, 4.3.2 e 4.3.3) e as baixas correlações com **can1** (Tabela 4.2.1), podemos dizer que X_2 não contribui à separação dos três grupos. Isto quer dizer que a largura da primeira junção do segundo tarso comporta-se de forma semelhante entre espécies.

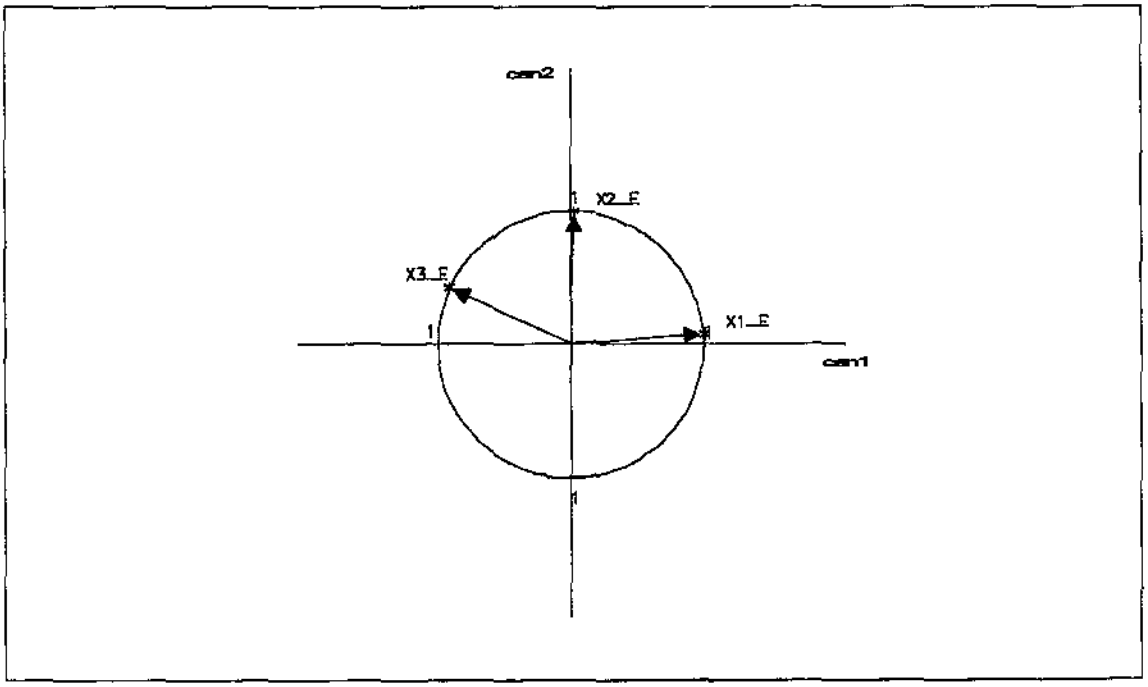


Figura 4.3.1: Círculo das Correlações para os valores das correlações Entre os grupos, na Tabela 4.2.1.

Quando tomamos como referência as correlações *Entre* grupos, Figura 4.3.1, observamos que o valor das correlações entre as variáveis originais e a **can1**, Tabela 4.2.1, é maior para X_1, que diz respeito à largura da primeira junção do primeiro tarso, com $corr_E(\mathbf{can}_1, X_1) = 0{,}99$ seguido de X_3, largura máxima do órgão copulativo, com $corr_E(\mathbf{can}_1, X_3) = -0{,}90$. Podemos dizer que elas contribuem notavelmente à separação dos grupos.

O contraste no sinal significa que a forma da inter-relação entre X_1 e X_3 que contribue à separação entre as espécies é da seguinte maneira: um aumento na largura da primeira junção do primeiro tarso corresponde uma diminuição na largura máxima do órgão copulativo e vice-versa, quando diminui a largura da primeira junção do primeiro tarso aumenta a largura máxima do órgão copulativo.

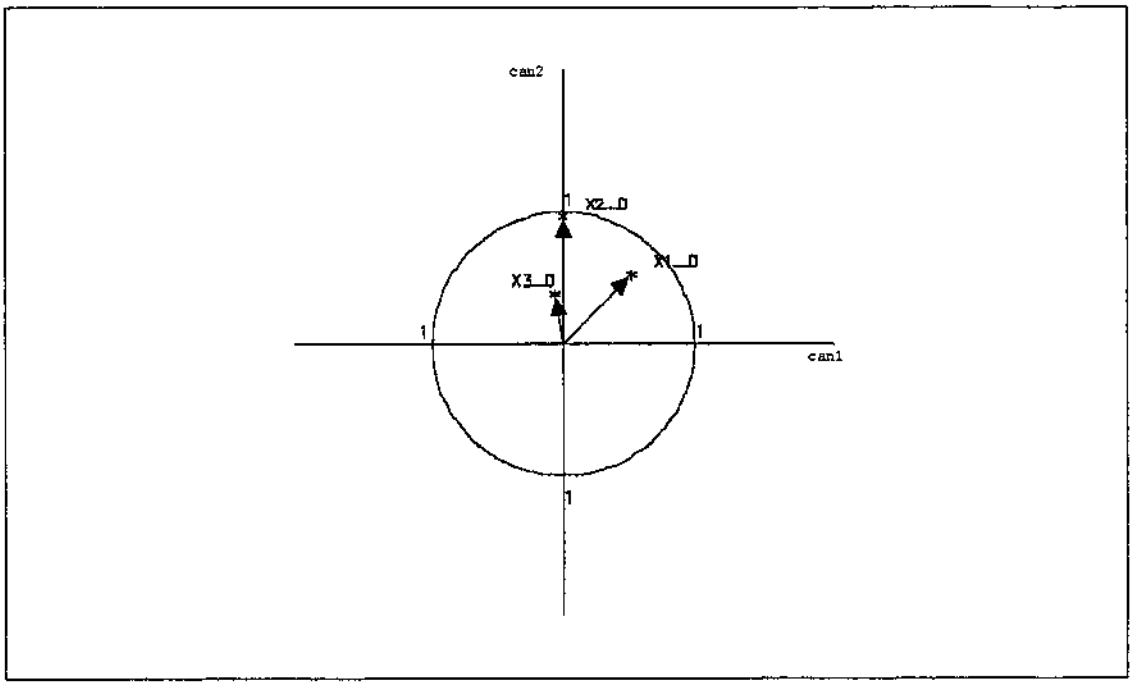


Figura 4.3.2: Círculo das Correlações para os valores das correlações dentro dos grupos , na Tabela 4.2.1.

Se observarmos o Círculo das Correlações *Dentro,* na Figura 4.3.2, vemos que dentro de cada espécie a inter-relação de X_1 não acompanha o padrão da

Figura 4.3.1 com a mesma intensidade e não exibe correlações diferenciadas com as discriminantes pois os valores são iguais, i.é, $corr_D(\mathbf{can}_1, X_1) = 0{,}52$ e $corr_D(\mathbf{can}_2, X_1) = 0{,}52$. O exame da matriz de correlações "DENTRO" revela que as correlações entre X_1 e as demais variáveis são: $corr_D(X_1, X_2) = 0{,}67$, $corr_D(X_1, X_3) = 0{,}79$.

Como X_2, muito bem representada neste gráfico, têm correlações altas com **can2**, $corr_D(\mathbf{can}_2, X_2) = 0{,}97$ e X_1 têm uma alta correlação com ela, nos inclinamos a pensar que X_1 não acompanhará a direção de **can1** dentro das espécies, isto reflete uma mudança da orientação nas inter-relações dentro de cada grupo, relativamente à orientação entre médias dos grupos, sendo também este o comportamento esperado para X_3 dado que a má representação desta variável não permite julgar, na Figura 4.3.2, a natureza da correlação dela com as variáveis discriminantes.

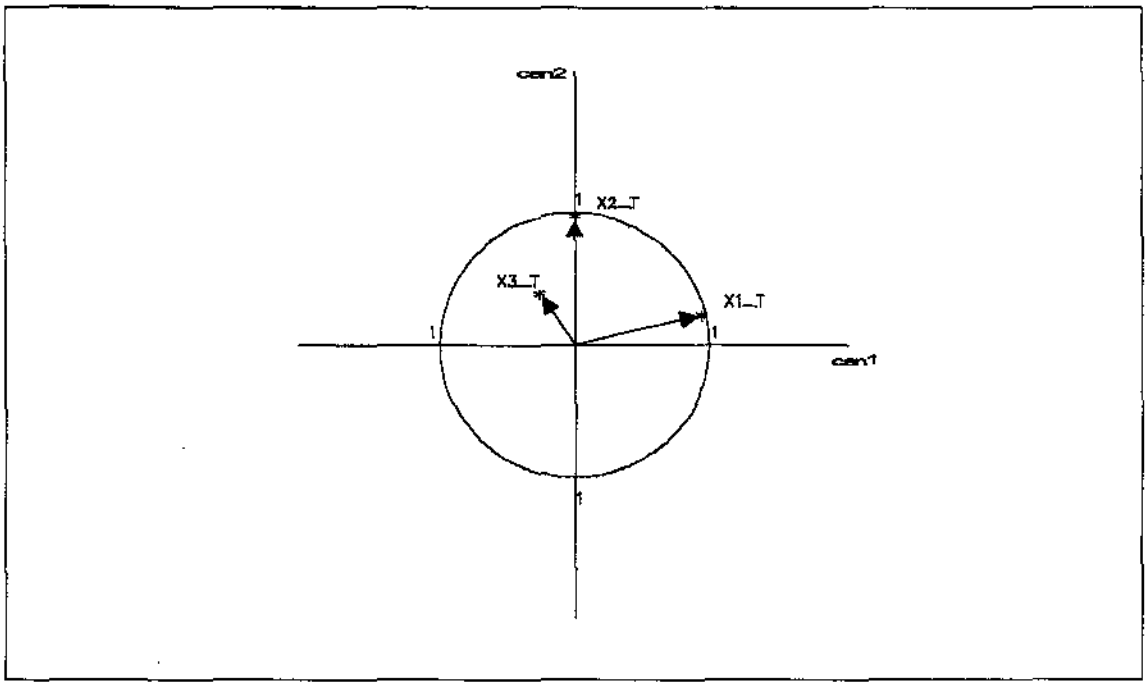


Figura 4.3.3: Círculo das Correlações para os valores das correlações *Totais*, na Tabela 4.2.1.

A observação do Círculo das Correlações para os valores das correlações *Totais*, Figura 4.3.3, dá informação sobre o comportamento da nuvem composta pelos 30 insetos, sem distinção de grupo. Vemos que X_3 vai ficar mal representada, com correlações baixas com **can1** e **can2**, mantendo sempre o contraste com X_1, sendo esta última variável quem carrega o peso fundamental na separação dos grupos, i.é, a largura da primeira junção do primeiro tarso é quem vai acompanhar a direção de **can1** (que é a direção que maximiza a variância entre grupos, por construção). A largura da primeira junção do segundo tarso, X_2, se comportará de forma similar para as três espécies, como tínhamos falado anteriormente, pois seguirá fielmente a direção da segunda discriminante.

Voltando agora nos coeficientes da Tabela 3.2.1, vemos que a ordem de importância das variáveis que ficou mais perto do padrão dado pelas correlações foi aquela dada pelos coeficientes extraídos dos dados padronizados, pois efetivamente, a ordem de importância das variáveis com respeito à estrutura *Entre* os grupos vista na Figura 4.3.1 coincide com a ordem dada pelo coeficiente Bruto, no caso dos dados padronizados, i.é, X_1, X_3, X_2. Para a estrutura *Dentro* e *Total* também, pois fica X_1, X_3, X_2.

Em todo caso podemos dizer que a ordem de importância dada às variáveis através da informação extraída das correlações, parece complementar a dada pelos coeficientes discriminantes, sobrando ainda algumas dúvidas quando as variáveis não ficam bem representadas nos Círculos das Correlações, como por exemplo no caso de X_3.

No entanto a análise feita anteriormente deixa claro que as correlações fornecem muito mais informação que os coeficientes pois, além de contrastes, eles só conseguem estabelecer uma ordem de importância das variáveis que não ajuda a distinguir qual é a medida da influência certa de cada uma delas. Ou seja, como vimos no caso de X_2, na Figura 4.3.2, ao nos referirmos à estrutura da fonte de variação *Dentro*, não é suficiente dizer que ela fica como terceira da ordem de importância na separação dos grupos, como foi dito pelos coeficientes,

pois na verdade X_2 não contribui nada a esta separação *Dentro* dos grupos, como foi constatado analisando às correlações. Inclusive no caso de X_2 este mesmo problema aconteceu quando a referência era a *Entre* e a *Total*.

Assim os Círculos das Correlações[11] permitem entender as diferenças entre os coeficientes, bem como, dão um rumo norte ao exame das correlações originais e suas componentes. Não é criado nenhum fato novo, apenas um instrumento que permite utilizar as semelhanças e/ou as diferenças nos coeficientes para melhor interpretá-los.

No capítulo 6, apresentaremos detalhadamente as interpretações dos Círculos das Correlações para os dados resultantes das simulações realizadas. Estes resultados foram de suma importância para a construção do roteiro para a interpretação dos resultados da Análise Discriminante de Fisher para quaisquer conjuntos de dados. Neste roteiro cruzamos os valores dos três coeficientes discriminantes e das três correlações a serem consideradas, entre as variáveis originais e discriminantes, para chegar a conclusões sobre como interpretar corretamente as saídas desta análise.

4.4 Indivíduos suplementares no caso da Análise Discriminante de Fisher

A importância geral do tratamento de indivíduos suplementares já foi tratada na literatura, algumas referências podem ser encontradas em LEBART, L., MORINEAU, A. e PIRON, M., 1995 e CRIVISQUI, 1999. De fato situações variadas geram o interesse de analisar o comportamento de indivíduos menos definidos ou mais diferenciados em um cenário construido a partir de indivíduos já melhor caracterizados. Por exemplo o estudo dos híbridos na presença de espécies conhecidas; o perfil de um sexo no cenário gerado pelo perfil do outro. Em estudos demográficos tratar casos "extremos" como suplementares têm sido útil, grandes capitais e demais municípios podem gerar distorções que conduzam à necessidade de tratar um ou outro como ponto suplementar. Enfim, o ponto

[11] Para construir estes gráficos foi modificada convenientemente uma macro do SAS (esta pode ser vista nos programas exemplos no Anexo), com este programa obtivemos os Círculos das Correlações para cada um dos casos que serão mostrados nos esquemas das simulações no capítulo 5.

suplementar não participa da definição do cenário e é projetado a posteriori para estudos comparativos.

Na aplicação prática que será feita no Capítulo 7 trataremos um grupo de indivíduos que apresentam características diferentes como indivíduos suplementares.

Teoricamente, dadas p variáveis, identificaremos um indivíduo suplementar qualquer como $w_0 = \{w_{01}, \ldots, w_{0p}\}$. Suas coordenadas na i-ésima variável discriminante de Fisher serão calculadas como:

$$\mathbf{can}_i(w_0) = \sum_{j=1}^{p} e_{ij} w_{0j}$$

Isto significa usar os coeficientes Discriminantes calculados inicialmente (com os indivíduos não suplementares) para substituir nas combinações lineares os valores observados para os indivíduos suplementares de tal forma que possamos usar essa coordenada calculada para localizá-los nos Planos Discriminantes e obter sua posição, relativa aos outros indivíduos originais. Isto será utilizado no exemplo com dados reais no Capítulo 7.

Capítulo 5: Estudo através de simulação

Para estudar a influência das variáveis na discriminação dos grupos assim como construir uma proposta que articule as representações gráficas com os coeficientes Discriminantes, foram trabalhados dados simulados, bivariados e trivariados, com uma estrutura de variâncias e covariâncias e correlação convenientemente selecionadas.

Partindo dos resultados obtidos no caso bivariado e do conceito de similaridade de matrizes, discutido anteriormente na seção 3.2.1, não é necessário diferenciar os casos com variâncias iguais e diferentes pois sabendo quais os coeficientes para variâncias iguais, que seria o caso dos dados padronizados, teríamos também os coeficientes para os dados brutos ou centrados fazendo um cálculo simples. No caso bivariado os resultados para variâncias distintas serão examinados por completitude.

Isto nos permitiu diminuir o número de casos a serem estudados, pois podemos conhecer o comportamento dos coeficientes para variâncias diferentes quando trabalhadas as variâncias iguais e vice-versa.

O trabalho com dados simulados permitiu analisar diferentes casos e configurações, escolhidas com o propósito de tratar situações padrão e que possibilitem, por semelhanças e diferenças, entender o comportamento da Função linear discriminante de Fisher com as extensões a casos reais. A seguir explicaremos as bases teóricas, os estudos de casos e as configurações usadas nas simulações.

5.1 Simulação de amostras bivariadas

♦ Variâncias iguais.

Primeiramente foram simulados dados com distribuição normal bivariada, i.e., temos o vetor $X=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ que se distribui normalmente com vetor de médias $\mu=\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ e matriz de variâncias e covariâncias $\Sigma=\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$.

Como é sabido, os contornos de densidade de probabilidade constantes para uma distribuição normal bivariada são elipses[1], dadas pela equação:

$$(X-\mu)'\Sigma^{-1}(X-\mu)=c^2 \qquad (5.1.1)$$

com $c^2=\chi_2^2(\alpha)$, onde $\chi_2^2(\alpha)$ é o (100α)*ésimo* percentil de uma distribuição qui quadrado com dois graus de liberdade.

Usando resultados da distribuição normal multivariada, no caso de $\sigma_{11}=\sigma_{22}$, e chamando de λ_1^*, λ_2^* e e_1^*, e_2^* aos autovalores e autovetores da matriz de variâncias e covariâncias Σ respectivamente, os autovalores podem ser calculados como $\lambda_1^*=\sigma_{11}+\sigma_{12}$ e $\lambda_2^*=\sigma_{11}-\sigma_{12}$.

Após a normalização, os autovetores ficam $(e_1^*)'=[1/\sqrt{2},\ 1/\sqrt{2}]$ e $(e_2^*)'=[1/\sqrt{2},\ -1/\sqrt{2}]$, i.é, são construídos de norma 1, correspondendo ao eixo principal da elipse com o maior autovalor, isto pode ser visto graficamente na Figura 5.1.1.

[1] Ver JOHNSON, R.A. AND WICHERN, D.W., 1992. p.131.

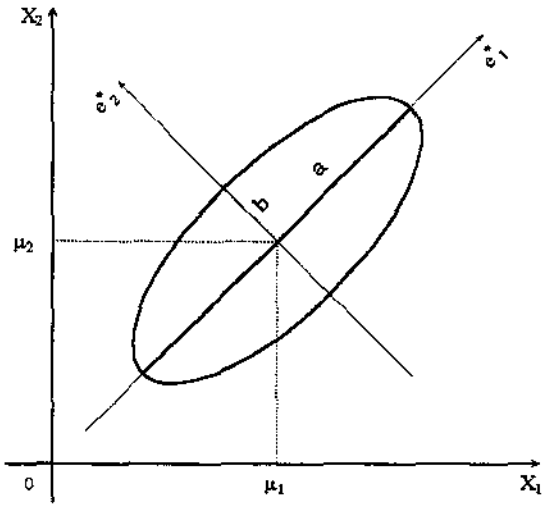


Figura 5.1.1: Contorno de densidade constante para uma distribuição normal bivariada com $\sigma_{11}=\sigma_{22}$ e $\sigma_{12}>0$. Os semi-eixos da elipse têm os valores $a=c\sqrt{\lambda_1^*}$ e $b=c\sqrt{\lambda_2^*}$

Neste caso, onde $\sigma_{11}=\sigma_{22}$, supondo λ_1^* o maior autovalor, poderíamos descrever a elipse de probabilidade constante para uma distribuição normal bivariada com: centro no ponto $(X_1,X_2)=(\mu_1,\mu_2)$, eixo principal na direção do autovetor e_1^* e, relativo às novas coordenadas (e_1^*,e_2^*), vértices nos pontos $(e_1^*,e_2^*)=\left(+c\sqrt{\lambda_1^*},0\right)$ e $(e_1^*,e_2^*)=\left(-c\sqrt{\lambda_1^*},0\right)$. O eixo secundário, por outro lado, encontra-se na direção do autovetor e_2^* com extremos nos pontos $(e_1^*,e_2^*)=\left(0,+c\sqrt{\lambda_1^*}\right)$ e $(e_1^*,e_2^*)=\left(0,-c\sqrt{\lambda_1^*}\right)$.

Para variáveis aleatórias normais correlacionadas positivamente, o eixo principal da elipse de densidade constante estará localizado paralelo à linha de 45° que passa através de μ; no entanto, se a correlação fosse negativa, $\lambda_2^*=\sigma_{11}-\sigma_{12}$ será o maior autovalor e o eixo principal da elipse de densidade constante fará angulo de 90° com a direção paralela à linha de 45° que passa através de μ.

Especificamente o seguinte resultado é certo para uma distribuição normal p-dimensional: o elipsóide sólido dos valores X satisfaz $(X-\mu)'\Sigma^{-1}(X-\mu)\leq \chi_p^2(\alpha)$, com probabilidade 1-α.

Notemos que as novas coordenadas (e_1^*, e_2^*) não são outras que as Componentes Principais Y_1 e Y_2 para as variáveis X_1 e X_2; Y_1 na direção do maior autovalor, i.é, a equação da elipse nas novas coordenadas fica

$$\frac{Y_1^2}{\left(c\sqrt{\lambda_1^*}\right)^2}+\frac{Y_2^2}{\left(c\sqrt{\lambda_2^*}\right)^2}=1$$

Sem perder generalidade, no caso de variâncias iguais, foram geradas amostras normais bivariadas com parâmetros $\mu_1=\mu_2=0$ e $\sigma_{11}=\sigma_{22}=1$; neste caso σ_{12} coincide com a correlação entre as duas variáveis.

♦ Variâncias diferentes

No caso de variâncias diferentes, os autovalores λ_1^* e λ_2^* são as soluções da equação característica

$$\left|\Sigma-\lambda^* I\right|=0 \qquad (5.1.2)$$

Os autovetores, por outro lado, foram calculados satisfazendo:

$$\Sigma\left(e^*\right)=\lambda_i^*\left(e^*\right), \text{ para } i=1,2 \qquad (5.1.3)$$

e posteriormente padronizados, ficando prontos para serem usados na construção das configurações de interesse no caso de $\sigma_{11}\neq\sigma_{22}$. Note-se que, exceto o fato de que os autovetores não fazem ângulo de 45° com os eixos, o resto acontece da mesma forma que como explicado para variâncias iguais, i.é, os contornos de probabilidade constante para uma distribuição normal bivariada são elipses dadas pela equação (5.1.1); os autovetores são construídos de norma um, correspondendo o eixo principal da elipse com o maior autovalor, λ_1^* para

variáveis aleatórias normais correlacionadas positivamente e λ_2^* se a correlação fosse negativa. Além do mais as novas coordenadas (e_1^*, e_2^*) continuam sendo as Componentes Principais Y_1 e Y_2 para as variáveis X_1 e X_2; Y_1 na direção do maior autovalor.

Sem perder generalidade, no caso de variâncias diferentes foram geradas amostras bivariadas com parâmetros $\mu_1 = \mu_2 = 0$ e $\sigma_{11} = 2\sigma_{22}$.

5.1.1 Esquema de simulação para o caso bivariado

A partir das considerações teóricas e em geral para quaisquer variâncias foram obtidas três amostras de tamanho 1000 de populações com igual matriz de variâncias e covariâncias e centralizadas na origem. O interesse não é focalizar flutuações aleatórias e sim verificar o comportamento populacional, ou seja, de configurações estáveis. Os questionamentos que nos motivam são: se as populações se configuram com um certo padrão, que impacto este padrão produz nos coeficientes? Não é de presente interesse saber como tamanhos de amostra influenciam este impacto.

Poderíamos ter optado por trabalhar algebricamente sobre as matrizes que produzem estas configurações. A opção por grandes amostras deve-se a que estudos[2] têm confirmado que, quando a razão entre o tamanho da amostra e o número de variáveis é igual a 20 ou mais, é comprovada a estabilidade dos coeficientes. Segundo isto não temos problemas de instabilidade, pelo que os valores dos coeficientes trabalhados são confiáveis.

Foi escolhido trabalhar com três grupos para facilitar a visualização gráfica do processo de simulação, também por esta ser uma quantidade não trivial e suficiente para nossos propósitos. Foram trabalhados dados bivariados e trivariados, procurando padrões de comportamento que permitissem a generalização dos resultados a p variáveis.

[2] Ver STEVENS, J., 1992. p.277.

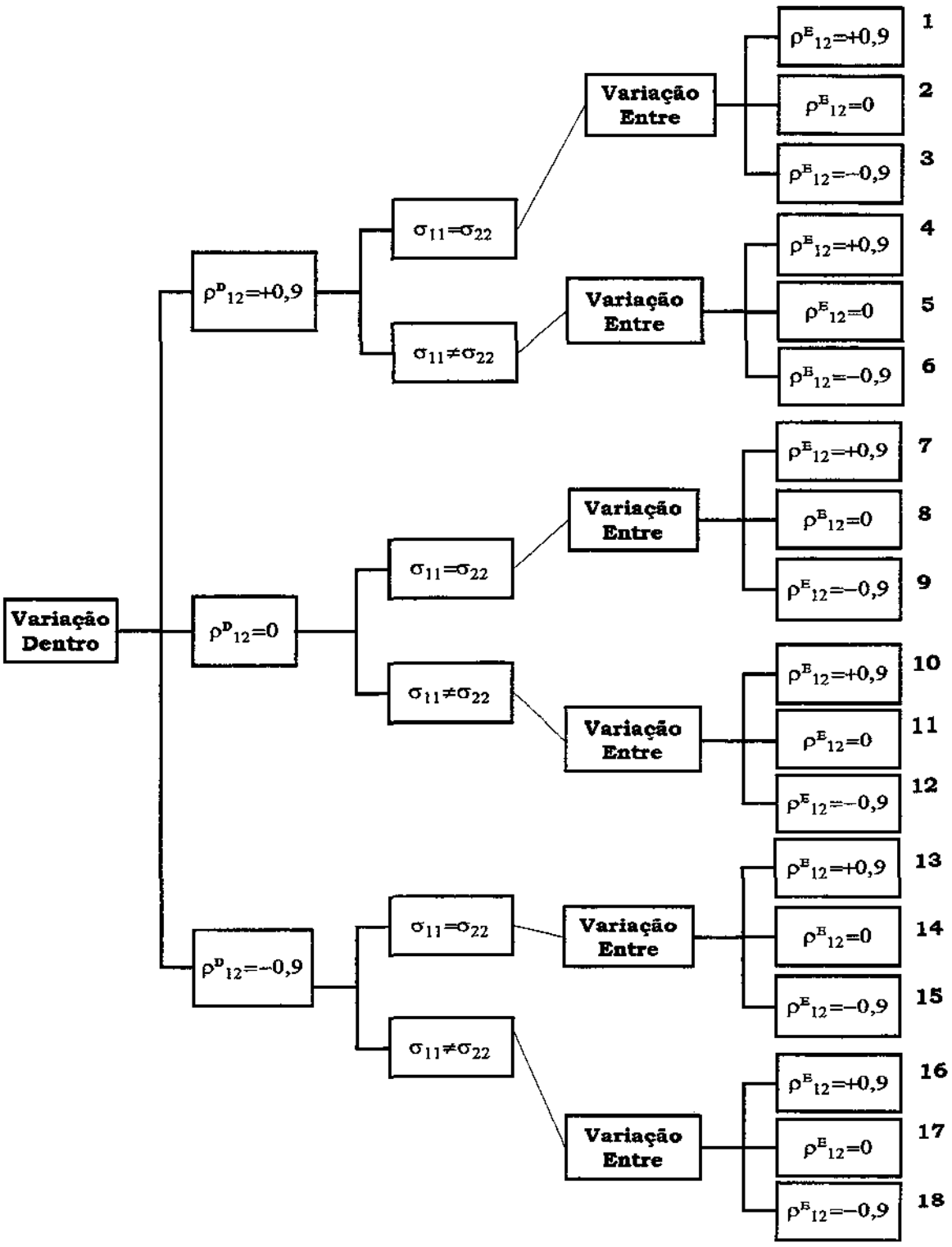


Figura 5.1.2: Esquema das simulações para o caso bivariado, à direita os números servirão para identificar cada caso.

O primeiro foi fixar um valor para a variância, $\sigma_{11} = \sigma_{22}$ ou $\sigma_{11} \neq \sigma_{22}$. A partir daqui se construíram grupos que refletissem uma estrutura de correlação dentro pré-determinada, i.é, para poder gerar inicialmente as três nuvens na origem, um valor para a correlação dentro. Desta forma a igualdade da matriz de variâncias e covariâncias através dos grupos está respeitada. Tomando em conta os nossos objetivos trabalhou-se com três configurações para a correlação dentro; alta positiva, $\rho_{12} = +0{,}9$, nula, $\rho_{12} = 0$, e alta negativa, $\rho_{12} = -0{,}9$.

Após gerar as três amostras na origem, e para cada caso de correlação dentro, duas médias foram deslocadas para obter três casos de configurações das médias dos três grupos correspondentes à correlação entre alta positiva, nula e alta negativa. Ou seja, no caso de duas variáveis, após fixar a variância, trabalhamos com 9 combinações diferentes de correlação dentro e entre, ver Figura 5.1.2.

A seguir apresentam-se os esboços das configurações dentro e entre para cada caso de variâncias diferentes.

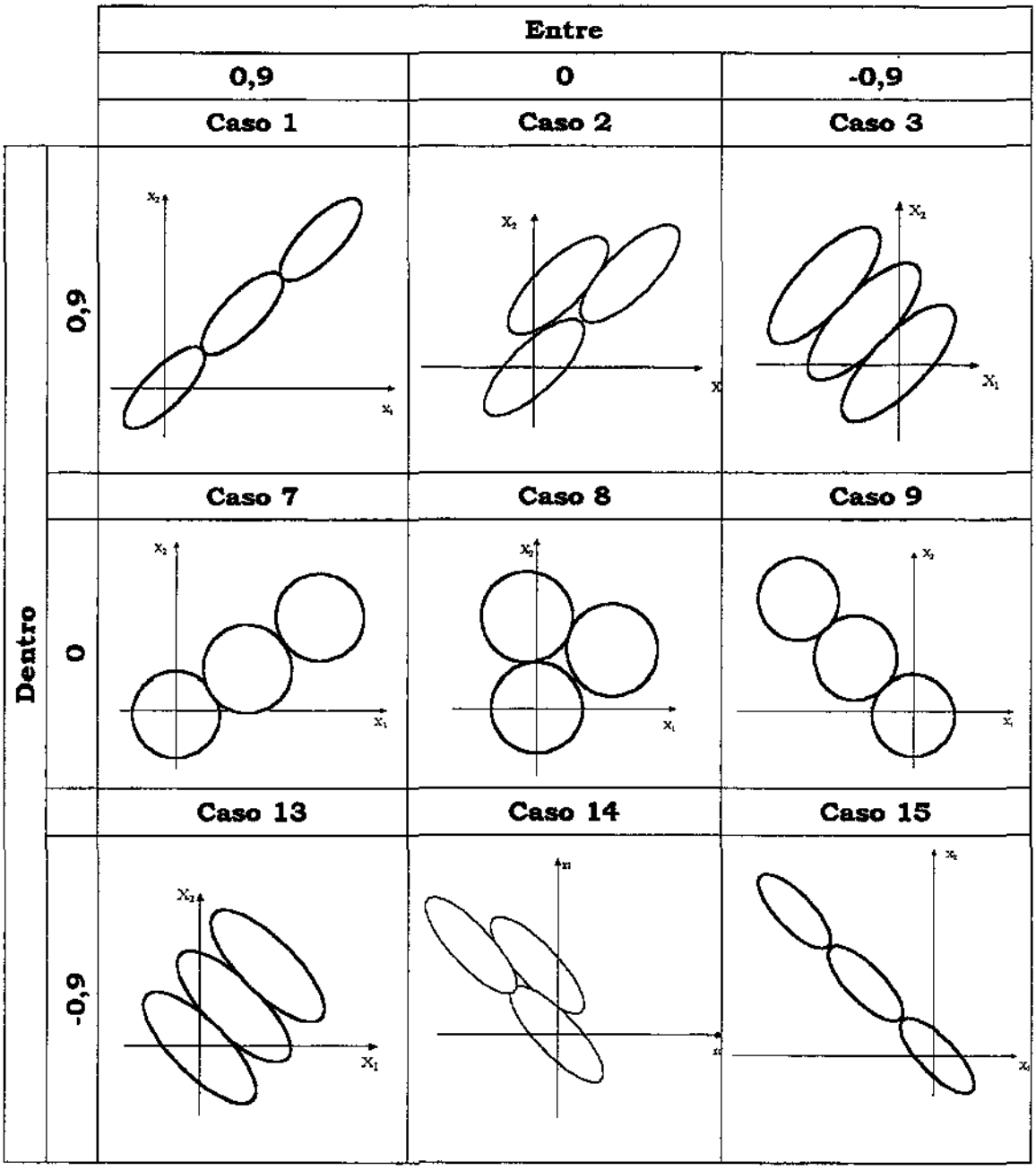


Figura 5.1.3: Exemplo das configurações correspondentes ao esquema de simulações para os casos bivariados com $\sigma_{11} = \sigma_{22}$. Por linha (por coluna) as correlações dentro (entre) dos grupos são iguais. As diagonais correspondem aos casos com correlações entre e dentro iguais.

O processo de simulação aconteceu da seguinte forma: mantendo fixa a posição da média de um dos grupos na origem, as outras duas foram localizadas no plano, calculando as novas coordenadas de tal forma que dentro de cada nuvem se encontrassem os indivíduos de cada grupo com 95% de probabilidade, ou seja, a possível mistura de indivíduos de um grupo com outro acontecerá no máximo com 5% de probabilidade para os indivíduos de cada grupo. Os cálculos específicos dependem do caso particular da variância, e serão vistos em detalhe quando particularizemos para cada um dos casos de variâncias trabalhados.

Após a localização das médias elas foram "perturbadas" para tirar o efeito da colinearidade, conseqüência da construção das configurações, e evitar problemas com a singularidade da matriz Σ_E, i.é, para evitar o problema do $|\Sigma_E| = 0$. O processo de geração dos dados, localização das médias e perturbação das coordenadas será detalhado a seguir para o caso bivariado.

♦ Variâncias iguais

1. Geração dos dados bivariados

Foram geradas 3 amostras bivariadas, cada uma de tamanho 1000. Sem perder generalidade trabalhou-se com média $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ e variâncias $\sigma_{11} = \sigma_{22} = 1$, portanto com a matriz de variâncias e covariâncias $\Sigma = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & 1 \end{bmatrix}$, coincidindo neste caso com a matriz de correlações R. O valor de $\sigma_{12} = \sigma_{21}$ dependeu da correlação dentro que foi escolhida para cada caso.

2. Localização conveniente das médias no plano

Os cálculos para localizar as amostras nas posições desejadas são detalhados a seguir para cada configuração:

2.1. *Configuração dentro alta positiva e entre alta positiva:*

2.1.1. calcular $L = c\sqrt{\lambda_1^*}$, lembrando que neste caso $c^2 = \chi_2^2$,

2.1.2. sabendo que o ângulo do primeiro autovetor $\mathbf{e}_1^*$ com X_1 é de 45°, calcular a projeção da elipse no eixo X_1, $k = \frac{L}{\sqrt{2}}$, este valor foi chamado de k , a segunda média foi para a coordenada $(2k,2k)$

2.1.3. a terceira média vai para $(4k,4k)$., ver Figura 5.1.4

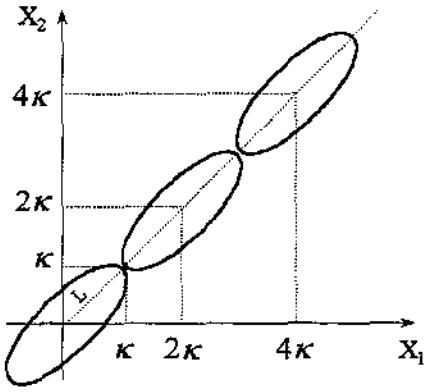


Figura 5.1.4: Exemplo do cálculo das posições dos grupos para conseguir a configuração desejada

2.2. *Configuração dentro alta positiva e entre zero:*

2.2.1. calcular $L_1 = c\sqrt{\lambda_1^*}$ e a projeção da elipse no eixo X_1, $k = \frac{L_1}{\sqrt{2}}$, desta forma a segunda média fica na coordenada $(2k,2k)$

2.2.2. calcular $L_2 = c\sqrt{\lambda_2^*}$

2.2.3. o triângulo retângulo destacado na Figura 5.1.5 têm o cateto maior igual a L_1, e o menor igual a $2L_2$. Usando Pitágoras calculamos a hipotenusa como $H = \sqrt{L_1^2 + 4L_2^2}$

2.2.4. calcular o ângulo α como $\alpha = \arccos\left(\frac{L_1}{H}\right)$ e θ como $\theta = \alpha + 45°$

2.2.5. calcular $k_x = H\cos\theta$ e $k_y = H\,\mathrm{sen}\,\theta$. A coordenada para a média da terceira amostra será (k_x, k_y).

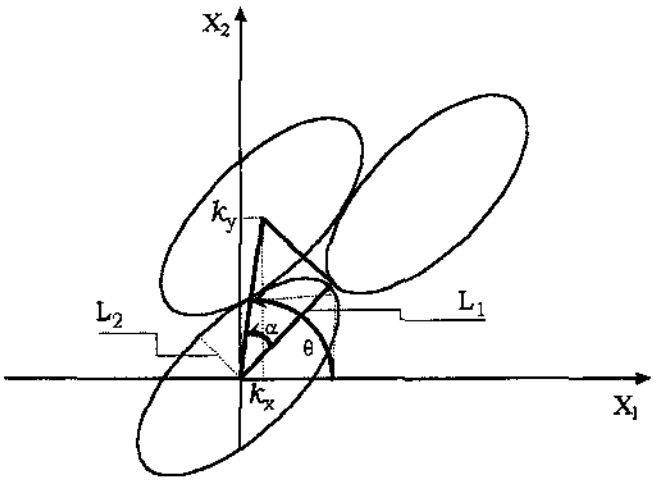


Figura 5.1.5 Cálculos para alocar as médias dos três grupos no caso da configuração dentro alta positiva e entre zero

2.3. *Configuração dentro alta positiva e entre alta negativa:*

2.3.1. Calcular $L = c\sqrt{\lambda_2^*}$, $k = \frac{L}{\sqrt{2}}$.

2.3.2. As coordenadas da segunda e terceira média ficam $(-2k, 2k)$ e $(-4k, 4k)$; ver Figura 5.1.4c para uma idéia melhor.

2.4. *Configuração dentro zero e entre alta positiva:* idem a 2.1

2.5. *Configuração dentro zero e entre zero:*

2.5.1. O cálculo de L, da projeção k e das coordenadas da média para o segundo grupos são os iguais a 2.1

2.5.2. Para localizar a terceira média neste caso, cuidou-se não somente que a separação dos grupos seguisse o mesmo padrão das outras configurações senão também que as três médias formassem um triângulo equilátero. A coordenada da terceira média ficou

$(ponto3x, ponto3y) = (k(1-\sqrt{3}), 2k - ponto3x)$. A Figura 5.1.6 mostra como ficaram os dados gerados após ter alocado as médias segundo esta configuração.

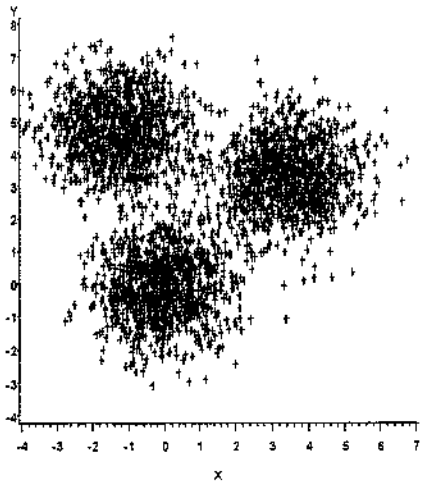


Figura 5.1.6 Gráfico das três amostras geradas e das posições na qual foram colocados os grupos.

2.6. *Configuração dentro zero e entre alta negativa:*

2.6.1. Idem ao caso 2.1, calcular $L = c\sqrt{\lambda_1^*}$ e $k = \frac{L}{\sqrt{2}}$, com a única diferença que desta vez as coordenadas das médias serão $(-2k, 2k)$ e $(-4k, 4k)$.

2.7. *Configuração dentro alta negativa e entre alta positiva:*

2.7.1. Calcular $L = c\sqrt{\lambda_2^*}$ e $k = \frac{L}{\sqrt{2}}$.

2.7.2. As coordenadas das médias serão $(2k, 2k)$ e $(4k, 4k)$. Para uma idéia mais clara da forma desta configuração ver Figura 5.1.4c.

2.8. *Configuração dentro alta negativa e entre zero:*

2.8.1. calcular $L_1 = c\sqrt{\lambda_1^*}$ e a projeção da elipse no eixo X_1, $k = \frac{L_1}{\sqrt{2}}$, desta forma a segunda média fica na coordenada $(-2k, 2k)$

2.8.2. calcular $L_2 = c\sqrt{\lambda_2^*}$

2.8.3. Parecido ao caso 2.2, usando Pitágoras calculamos a hipotenusa como $H = \sqrt{L_1^2 + 4L_2^2}$

2.8.4. calcular o ângulo α como $\alpha = \arccos\left(\frac{L_1}{H}\right)$ e θ como $\theta = \alpha + 45°$

2.8.5. calcular $k_x = H\cos\theta$ e $k_y = H\,\text{sen}\,\theta$. A coordenada para a média da terceira amostra será $(-k_x, k_y)$.

2.9. *Configuração dentro alta negativa e entre alta negativa:*

2.9.1. Calcular $L = c\sqrt{\lambda_1^*}$ e $k = \frac{L}{\sqrt{2}}$.

2.9.2. As coordenadas das médias serão $(-2k, 2k)$ e $(-4k, 4k)$. Para uma idéia mais clara da forma desta configuração ver Figura 5.1.5a.

3. Perturbação das coordenadas das médias dos grupos

Mantendo um grupo centralizado na origem, os outros dois foram perturbados para remover a colinearidade com uma uniforme $U(-2\sigma_{\bar{x}}, +2\sigma_{\bar{x}})$ que permitiu um desvio das médias dentro do intervalo de mais ou menos dois desvios padrões da média, i.é, $\pm 2\sigma_{\bar{X}} = \pm 2\frac{1}{\sqrt{1000}}$. A quantidade gerada da uniforme foi somada aos valores das coordenadas da média para o segundo e terceiro grupo; na Figura 5.1.7 pode se observar como o programa fez esta perturbação.

```
data tese.perturba;
     set tese.geranorm;
               var1=1; var2=1;
               desvio1=sqrt(var1);
               desvio2=sqrt(var2); rmil=sqrt(1000);

/* Perturbação aleatória das coordenadas no eixo x e y */
```

```
/* gerar U(-2/sqrt(1000),+2/sqrt(1000)) */

  if amostra=1 then do;
    x=xorig;
    y=yorig;
  end;

  if amostra>1 then do;
    uni01X=ranuni(12);
    uni01Y=ranuni(23);
    uniabX=2*(2/rmil)*uni01X-(2/rmil);
    uniabY=2*(2/rmil)*uni01Y-(2/rmil);

/* perturbando x e y para a segunda e terceira médias */

    x=xorig+uniabX;
    y=yorig+uniabY;

  end;
run;
```

Figura 5.1.7: Fragmento do programa em SAS que mostra o procedimento de perturbação das médias para evitar o problema de colinearidade.

Esta perturbação aleatória foi feita da mesma forma para todos os casos, como explicado anteriormente. Após estas etapas os dados ficaram prontos para entrar na Análise Discriminante de Fisher.

♦ Variâncias diferentes

Até aqui trabalhamos gerando variáveis com variâncias iguais, $\sigma_{11} = \sigma_{22}$. Na segunda etapa, apesar de saber o impacto nos coeficientes quando as variâncias das variáveis originais são diferentes, descrito na seção 3.2.1, o interesse foi ilustrá-lo.

Para conseguir amostras com variâncias diferentes que ao mesmo tempo fossem comparáveis com os casos estudados para variâncias iguais, fixamos a correlação dentro e uma relação para as variâncias, desta forma geramos novas amostras bivariadas cuidando para manter fixa a variância generalizada, com respeito aos casos análogos para variâncias iguais.

A variância generalizada permite escrever a informação das variâncias e covariâncias das p variáveis em um número simples. Ela pode ser obtida

calculando o determinante da matriz de variâncias e covariâncias total para a amostra, $|S_T|$. Mantendo fixa a variância generalizada, conseguimos comparar as nuvens obtidas com variâncias diferentes de aquelas geradas com variâncias iguais.

Os detalhes do cálculo feito para calcular quais as variâncias diferentes obtidas em cada caso, um pouco para ilustrar isto, foi o seguinte:

Seja, por exemplo o caso $\sigma_{11} \neq \sigma_{22}$, que como dissemos anteriormente foi trabalhado usando a relação $\sigma_{11} = 2\sigma_{22}$. Calculando o determinante da matriz de variâncias e covariâncias neste caso fica $\begin{vmatrix} 2\sigma_{22} & \rho_{12}\sqrt{2}\sigma_{22} \\ \rho_{12}\sqrt{2}\sigma_{22} & \sigma_{22} \end{vmatrix} = 2\left(1-\rho_{12}^2\right)\sigma_{22}^2$.

A condição imposta, de que as amostras simuladas para o caso de variâncias diferentes tenham a mesma variância generalizada do que as amostras para o caso de variâncias iguais, conseguiu-se igualando este determinante ao determinante para o caso $\sigma_{11} = \sigma_{22}$, o qual fica $\begin{vmatrix} \sigma_{22} & \rho_{12}\sigma_{22} \\ \rho_{12}\sigma_{22} & \sigma_{22} \end{vmatrix} = \left(1-\rho_{12}^2\right)\sigma_{22}^2$.

Resolvendo uma equação simples obtemos $\sigma_{22} = \frac{1}{\sqrt{2}}$, ficando a matriz de variâncias e covariâncias para o caso de variâncias diferentes da forma $\begin{bmatrix} \sqrt{2} & \rho_{12} \\ \rho_{12} & \sqrt{2}/2 \end{bmatrix}$. Faltaria somente substituir os valores das correlações entre as duas variáveis originais e obter as matrizes de variâncias e covariâncias $\begin{bmatrix} \sqrt{2} & 0{,}9 \\ 0{,}9 & \sqrt{2}/2 \end{bmatrix}$, $\begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2}/2 \end{bmatrix}$ e $\begin{bmatrix} \sqrt{2} & -0{,}9 \\ -0{,}9 & \sqrt{2}/2 \end{bmatrix}$, para cada caso particular. Portanto estes casos com $\sigma_{11} \neq \sigma_{22}$ além de preservar a variância generalizada, respeito aos casos com $\sigma_{11} = \sigma_{22}$, também manteve a mesma estrutura de correlação dentro dos grupos.

5.1.2 Recortes para as análises.

Para uma maior clareza a Figura 5.1.2 mostra o plano de simulação com todos os casos numerados. Em todos os casos geramos as amostras, as alocamos geometricamente, perturbamos as médias, fizemos Análise Discriminante de Fisher completa e os Círculos das Correlações para as correlações entre as variáveis originais e as discriminantes.

Para estudar a influência nos coeficientes da Função Linear Discriminante de Fisher, foram separados os objetivos a serem alcançados ao fazer cada uma das simulações, que estão explicados a seguir em forma de tabela. Os números com os quais identificamos os casos são os mesmos da Figura 5.1.2.

a) Para comparar a influência da correlação "DENTRO"

	Correlação "ENTRE"		
Variação Dentro	$\rho_{12}^{E} = +0{,}9$	$\rho_{12}^{E} = 0$	$\rho_{12}^{E} = -0{,}9$
$\sigma_{11} = \sigma_{22}$	1-7-13	2-8-14	3-9-15
$\sigma_{11} \neq \sigma_{22}$	4-10-16	5-11-17	6-12-18

São assim comparados os casos listados nas caselas da tabela, pois referem-se a distintas correlações dentro, mantidas constantes as configurações das variâncias "DENTRO" e correlação "ENTRE".

b) Para comparar a influência da correlação "ENTRE"

Da mesma forma na tabela seguinte são comparados os casos listados nas caselas da tabela, referentes neste caso a distintas correlações "DENTRO", mantendo constantes as configurações das variâncias "DENTRO" e correlação "DENTRO".

	Correlação "DENTRO"		
Variação Dentro	$\rho_{12}^{D} = +0,9$	$\rho_{12}^{D} = 0$	$\rho_{12}^{D} = -0,9$
$\sigma_{11} = \sigma_{22}$	1-2-3	7-8-9	13-14-15
$\sigma_{11} \neq \sigma_{22}$	4-5-6	10-11-12	16-17-18

c) Para comparar a influência de variáveis originais com variâncias diferentes.

Por último, no caso de duas variáveis, o interesse foi comparar os casos listados nas caselas da tabela, os quais, mantendo constantes as configurações de correlação "DENTRO" e "ENTRE", permitem comparar os resultados para variâncias iguais e diferentes.

	Correlação "ENTRE"		
Correlação "DENTRO"	$\rho_{12}^{E} = +0,9$	$\rho_{12}^{E} = 0$	$\rho_{12}^{E} = -0,9$
$\rho_{12}^{D} = +0,9$	1-4	7-10	13-16
$\rho_{12}^{D} = 0$	2-5	8-11	14-17
$\rho_{12}^{D} = -0,9$	3-6	9-12	15-18

Vemos que no caso da variabilidade dentro será de interesse comparar os casos de variância, das variáveis originais, iguais, $\sigma_{11} = \sigma_{22}$, e o caso de variâncias diferentes, $\sigma_{11} \neq \sigma_{22}$. Como a variabilidade entre é induzida por uma taxa de má classificação, fixada a *priori* na hora da alocação das médias geometricamente; ela não é de interesse, neste caso só serão de interesse os efeitos de correlações diferentes entre os grupos.

Os resultados para cada um destes recortes encontram-se separados no capítulo 6, nas seções 6.1.2, 6.1.2 e 6.1.3, respectivamente.

5.2 Simulação de amostras trivariadas

No caso de três variáveis, também trabalhamos com três grupos, igual que no caso anterior de duas variáveis. O primeiro passo foi escolher as combinações de correlação dentro e entre que seriam trabalhadas. Inicialmente foram considerados, como de possível interesse, nove casos de orientações fortes .

No caso da correlação dentro o interesse foi trabalhar com 9 combinações, i.é, oito combinações de orientações fortes, que são: $(+,+,+)_D$[3], $(-,+,+)_D$, $(+,-,+)_D$, $(+,+,-)_D$, $(+,-,-)_D$, $(-,+,-)_D$, $(-,-,+)_D$, $(-,-,-)_D$, e o caso $(0,0,0)_D$ que significa ausência total de correlação dentro.

Das oito combinações que dão orientações fortes só quatro foram escolhidas, elas são, $(+,+,+)_D$, $(+,-,-)_D$, $(-,+,-)_D$,$(-,-,+)_D$. As outras quatro não reproduzem matrizes positivas definidas, pelo que não servem para os nossos propósitos. Além disso notamos que as variáveis são intercambiáveis; i.é, $(+,-,-)_D$, $(-,+,-)_D$ ou $(-,-,+)_D$ representam o mesmo caso, portanto só um destes casos foi considerado. Em particular, e só para conseguir uma boa visualização gráfica das nuvens, escolhemos entre estes três o caso $(+,-,-)_D$.Ficaram então dois casos de orientações fortes, $(+,+,+)_D$, $(+,-,-)_D$, mais o caso $(0,0,0)_D$. Estas três combinações foram as escolhidas para as correlações dentro.

Para variação entre considerou-se apenas as anisotropias mais acentuadas e a isotropia total, ou seja, matrizes de correlação uniforme com valores 0,9 e 0. O caso $-0,9$ não foi considerado por esta matriz, i.é, $S_E = \begin{bmatrix} 1 & -0,9 & -0,9 \\ -0,9 & 1 & -0,9 \\ -0,9 & -0,9 & 1 \end{bmatrix}$, não

[3] Usaremos esta notação para representar abreviadamente os casos, aqui $(+,+,+)_D$, por exemplo, significa $\rho_{12}^D = +0,9$, $\rho_{13}^D = +0,9$, $\rho_{23}^D = +0,9$, o que em forma matricial poderíamos escrever como $S_D = \begin{bmatrix} 1 & 0,9 & 0,9 \\ 0,9 & 1 & 0,9 \\ 0,9 & 0,9 & 1 \end{bmatrix}$. Nos outros casos a notação é equivalente à explicada anteriormente.

ser positiva definida. Desta forma ficamos com três casos para as correlações entre os grupos, i.é, $(+,+,+)_E$[4], $(+,-,-)_E$, mais o caso $(0,0,0)_E$.

Variações distintas não são consideradas pelo resultado de similaridade entre matrizes da seção 3.2.1, só serão de interesse os casos de variáveis com variâncias iguais. Novamente, e sem perder generalidade, trabalharemos considerando $\sigma_{12} = \sigma_{13} = \sigma_{23} = 1$.

Na Figura 5.2.1 podemos ver o esquema completo das simulações para o caso trivariado, com as nove combinações de casos que ficaram no final.

[4] Neste caso estamos nos referindo abreviadamente às correlações entre os grupos, onde $(+,+,+)_E$ significa $\rho_{12}^E = +0{,}9$, $\rho_{13}^E = +0{,}9$, $\rho_{23}^E = +0{,}9$, ou $\mathbf{S}_z = \begin{bmatrix} 1 & 0{,}9 & 0{,}9 \\ 0{,}9 & 1 & 0{,}9 \\ 0{,}9 & 0{,}9 & 1 \end{bmatrix}$.

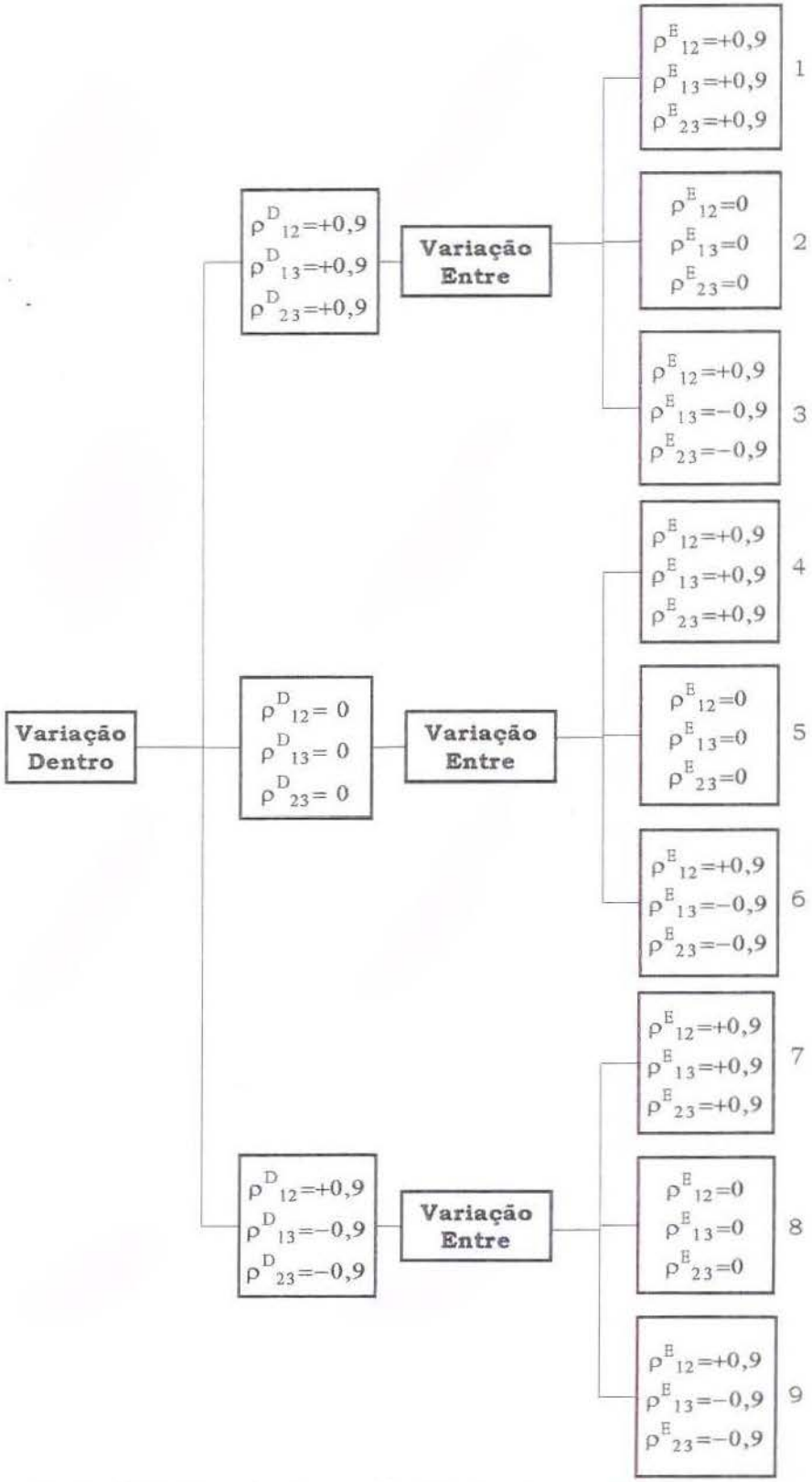


Figura 5.2.1: Esquema das simulações para o caso de três variáveis

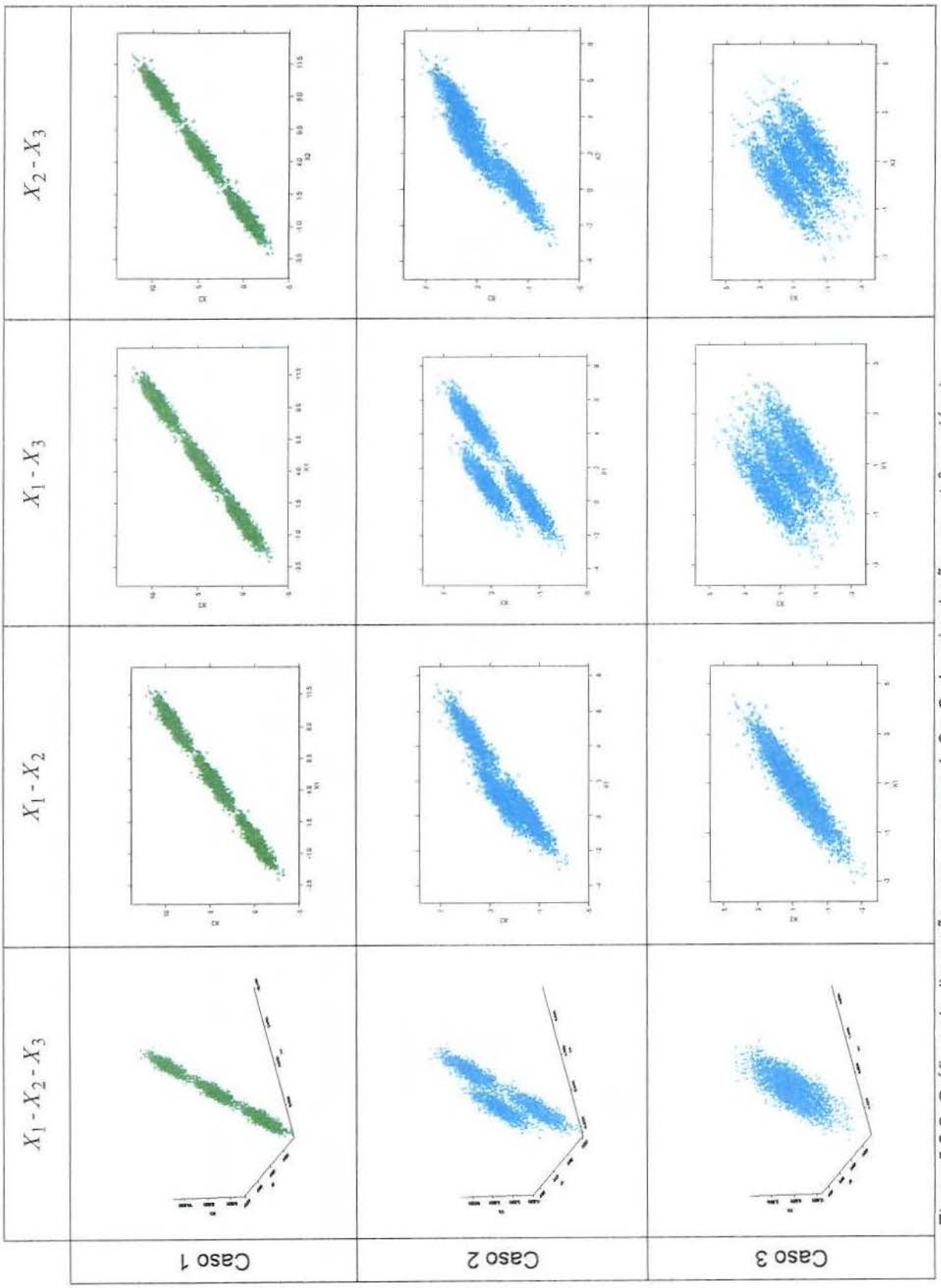


Figura 5.2.2: Gráficos de dispersão para os casos 1, 2 e 3, das simulações com três variáveis

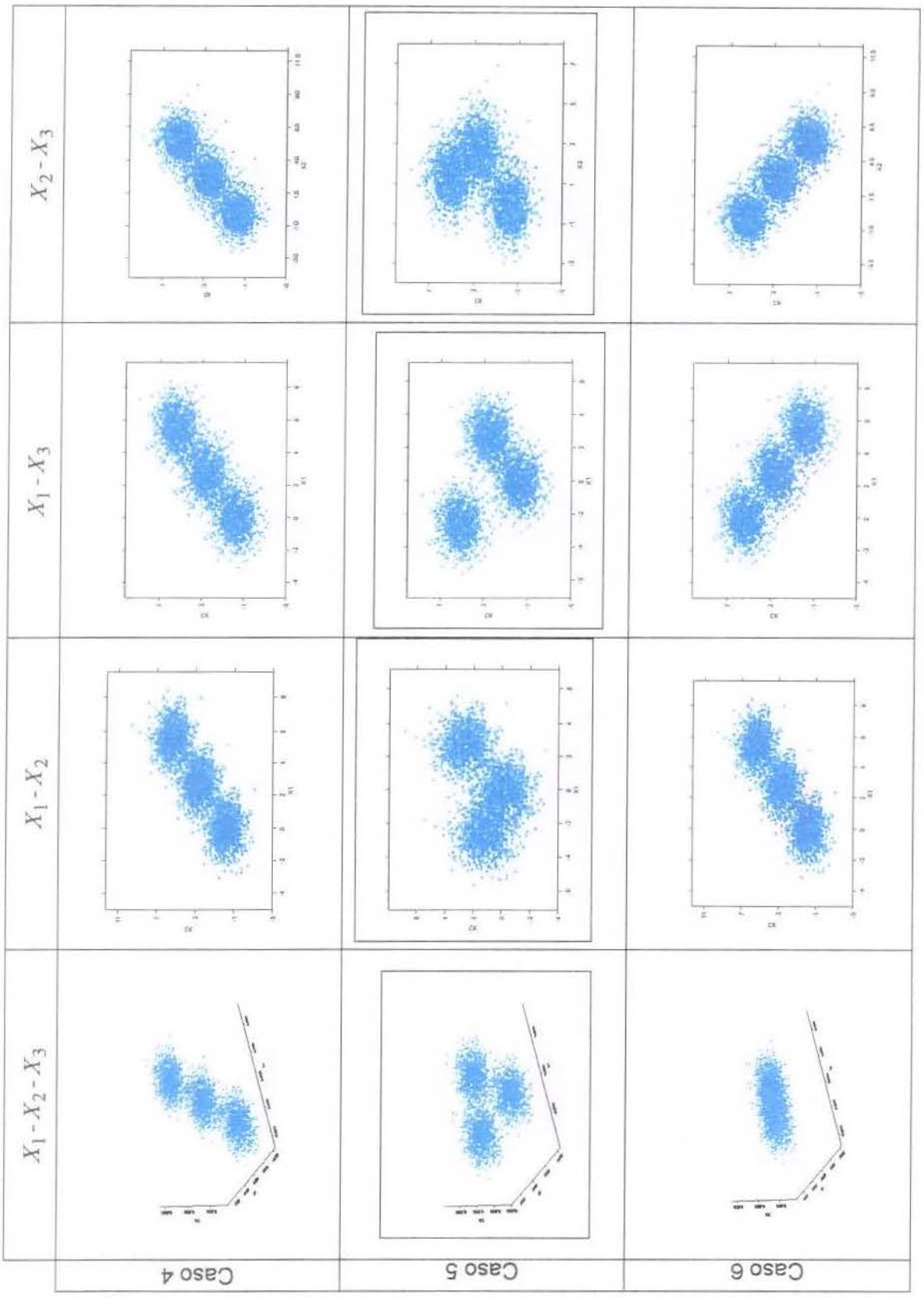


Figura 5.2.3: Gráficos de dispersão para os casos 4, 5 e 6, das simulações com três variáveis

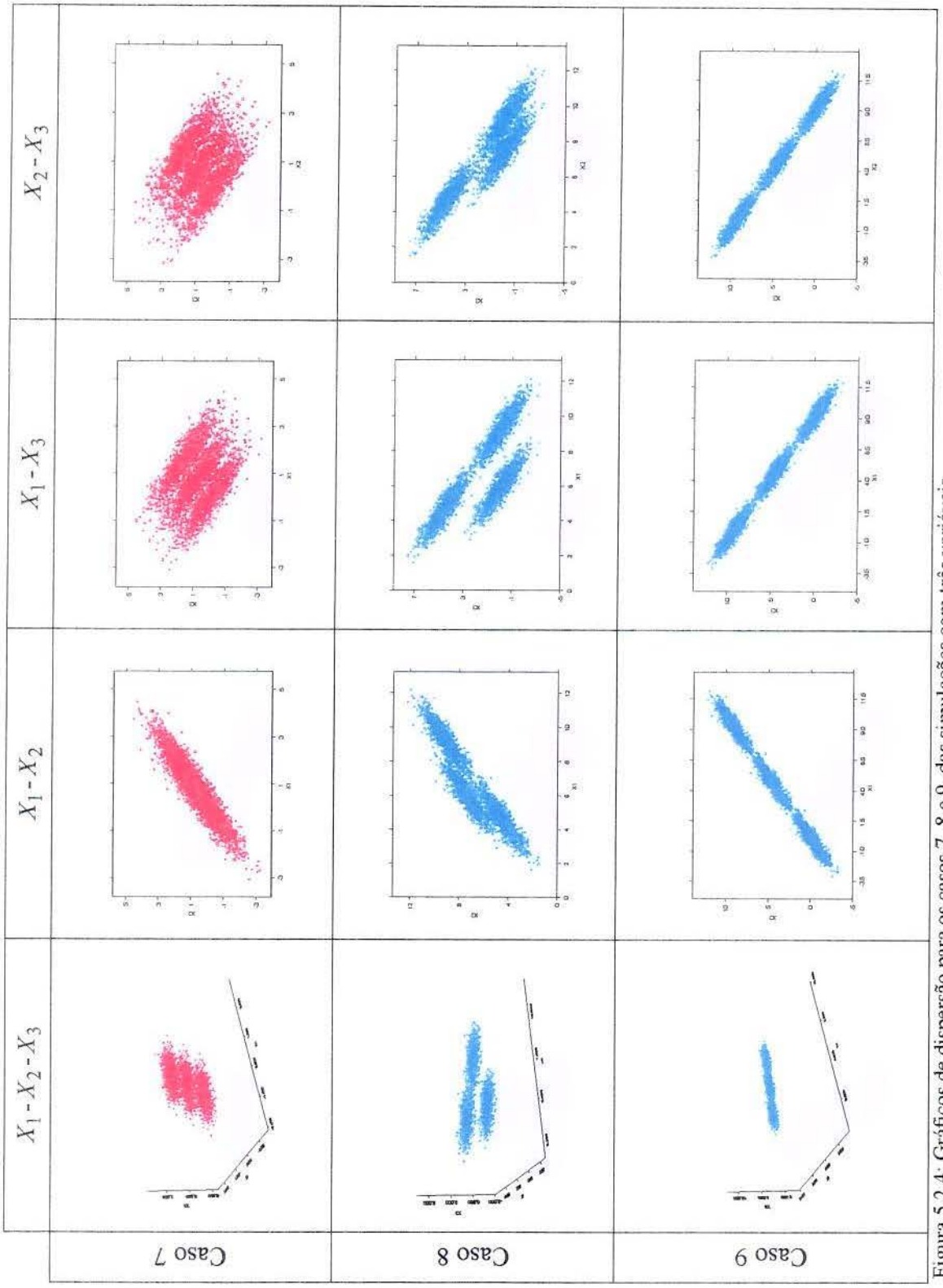


Figura 5.2.4: Gráficos de dispersão para os casos 7, 8 e 9, das simulações com três variáveis

Capítulo 6: Resultados

A idéia da interpretação dos Círculos das Correlações é muito interessante, pois quando tomamos como referência a posição relativa das setas que representam as variáveis, com respeito aos eixos **can1** e **can2**, indicam quanto os dados, vistos por momentos segundo um interesse especifico (vistos como grupos: estrutura de correlação entre, vistos dentro dos grupo: estrutura de correlação dentro e vistos como um conjunto de dados só, desprezando a estrutura de grupos: correlação para a amostra total) acompanham a direção dos eixos discriminantes.

Usando a mesma idéia anterior, individualizando agora as variáveis, pode-se dizer como cada uma em particular influi neste comportamento. Desta forma poderemos obter comportamentos iguais ou diferentes, dependendo da situação.

No caso de variâncias iguais e quando observadas as posições relativas das variáveis originais nos Círculos das Correlações, como foi explicado teoricamente nas seção 4.2.1, podemos recuperar a informação de qual a correlação dentro, entre e total entre elas, podendo ser comprovada dando uma olhada nas matrizes de correlações.

No caso da distância à origem, e seguindo o já dito na seção 4.2.2, apreenderemos qual variável discriminante será a que melhor representa cada variável original. Todos estes critérios serão mostrados a seguir com a interpretação dos dados simulados.

Após simulados os dados para cada estrutura de covariância e correlação de interesse, segundo os esquemas de simulação das Figuras 5.1.2 e 5.2.1, foi aplicada a Análise Discriminante de Fisher e obtidos os coeficientes da Função Linear Discriminante de Fisher e as correlações entre as variáveis originais e

discriminantes, tomando em conta a variação dentro, entre e total. Nos ocuparemos neste capítulo dos resultados da Análise Discriminante de Fisher, e da sua interpretação, para os casos trabalhados.

Desta forma o objetivo central neste capítulo será analisar se e como os coeficientes discriminantes e as correlações entre as variáveis originais e discriminantes permitem recuperar a estrutura de correlação, para as diferentes configurações de dados construídas, com a idéia de, ao trabalhar com dados reais, fazer o contrário, i.é, fazer Análise Discriminante de Fisher, construir os Círculos das Correlações e interpretar a importância das variáveis segundo a experiência acumulada após o estudo de simulação realizado.

Na seção 6.1 são mostrados os resultados dos coeficientes discriminantes de Fisher e os resultados das correlações entre as variáveis originais e as discriminantes para duas variáveis, tomando em conta os recortes mencionados na seção 5.1.2. Para maior facilidade na seção 6.1.1 vemos os resultados correspondentes ao recorte a) da seção 5.1.2, quando comparada a influência da correlação "DENTRO". Na seção 6.1.2, quando comparada a influência da correlação "ENTRE", recorte b) e na seção 6.1.3 quando o interesse é ver se verdadeiramente o fato das variáveis originais terem variâncias iguais ou diferentes influencia na hora de discriminar os grupos, que se corresponde com o recorte c) visto na seção 5.1.2. Resultados para os casos análogos, desta vez com três variáveis, são apresentados na seção 6.2.

Em geral, e pela forma como foram construídos os casos de interesse, o primeiro autovalor acumulará mais do 99% de variância entre os grupos, salvo algumas exceções que serão devidamente assinaladas. É por isto que em geral, nos casos trabalhados aqui com dados simulados, a importância fundamental na hora da interpretação recai sobre a **can1** .

Os Círculos das Correlações entre as variáveis originais e as discriminantes foram colocados dentro de tabelas para conseguir, em cada momento, a visualização completa dos casos de interesse.

6.1 Comportamento dos coeficientes da Função Linear Discriminante de Fisher e das correlações entre as variáveis originais e discriminantes variando as correlações "DENTRO" e "ENTRE" os grupos, para duas variáveis originais.

A seguir apresentam-se os resultados das simulações para o caso de duas variáveis. Mudanças na correlação "DENTRO" e "ENTRE" foram feitas para descobrir o efeito destas nos coeficientes da Função Linear Discriminante de Fisher.

6.1.1 Comparando a influência de mudanças na correlação "DENTRO" dos grupos, mantendo fixa a correlação "ENTRE".

Nestes casos mantemos fixa a correlação "ENTRE" os grupos e a relação entre as variâncias das variáveis originais, em alguns casos iguais, $\sigma_{11} = \sigma_{22} = 1$, e em outros diferentes, $\sigma_{11} = 2\sigma_{22}$, procurando ver se a variação da correlação "DENTRO" influencia no cálculo dos coeficientes da Função Linear Discriminante de Fisher.

- **Casos 1, 7 e 13**

A Tabela 6.1.1 mostra os coeficientes da Função Linear Discriminante de Fisher para entre alta positiva, $\rho^E = +0{,}9$, e para variáveis originais com variâncias iguais, $\sigma_{11} = \sigma_{22} = 1$. Neste caso a **can1** explica no caso 1, 99,99% e nos casos 7 e 13, 100% da variância entre os grupos, e portanto ela consegue separar muito bem os grupos.

No caso 1, com variâncias iguais e "ENTRE" e "DENTRO" altas positivas, sabemos que (ver Figura 5.1.3 a) os grupos estão separados e esta separação é muito bem captada pela primeira discriminante (ver Figura 6.1.1).

Segundo o coeficiente bruto, parece que X é levemente mais importante que Y na separação dos grupos. Os coeficientes *Dentro* e *Total* indicam também que a importância de X com respeito a Y é maior. No entanto estas diferenças de magnitude *Entre*, *Dentro* e *Bruto* nos casos 1 e 7 e maior semelhança no 13 não

parece conter informação relevante para a interpretação das variáveis. Mesmo que os sinais positivos em **can1** parecem estar associados a correlações entre os grupos positivas.

Tabela 6.1.1: Coeficientes quando a correlação entre é fixada para ser alta positiva (Casos 1-7-13)

Variação	Variáveis	CASO 1 (=,+,+)[1]		CASO 7(=,0,+)		CASO 13(=,-,+)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	2.1792	-8.9398	2.1618	-2.1486	2.9412	0.6985
	Y	2.0153	8.9420	2.1049	2.1519	2.9542	-0.6953
Dentro	X	0.5326	-2.1849	0.7092	-0.7049	2.1724	0.5159
	Y	0.4944	2.1939	0.6992	0.7148	2.1734	-0.5115
Bruto	X	0.5418	-2.2226	0.7215	-0.7171	2.2098	0.5248
	Y	0.4989	2.2139	0.6933	0.7087	2.2060	-0.5192

Ao analisar as correlações, por outro lado, vemos que os valores das correlações entre X e Y, e a discriminante **can1** para o Caso 1 a), b) e c) são altas (ver Tabela 6.1.3). A boa representação das variáveis X e Y nos Círculos das Correlações, representados na Tabela 6.1.2, permite ver como as duas variáveis contribuem grandemente na separação dos grupos, ao observar as correlações *Entre*, acompanhadas na mesma intensidade pelas correlações *Dentro* dos grupos, o qual era já sabido da configuração construída para este caso.

A correlação *Entre* quase zero de X e Y com **can2**, seguida das também baixas, -0.2198 e 0.2368, respectivamente, para a *Dentro*, dão idéia de "achatamento" na direção de **can2**, pelo que podemos imaginar as nuvens com um formato alongado na direção de **can1**, como são na realidade.

[1] Os sinais entre parênteses servem para dar uma indicação rápida de se estamos trabalhando com variâncias iguais ou diferentes e de quais as correlações dentro e entre que se correspondem com cada caso. Exemplo: (=,+,0), (≠,+,-) correspondem-se com: variâncias iguais, dentro +0,9 e entre zero; variâncias diferentes, dentro +0,9 e entre –0,9, respectivamente.

A correlação *Total*, esquecendo neste caso a separação em grupos, também mostra este comportamento para a amostra completa. Observa-se, pelo pequeno ângulo que formam entre elas, que as variáveis X e Y estão altamente correlacionadas para a amostra total.

Neste caso 1 podemos observar também que a direção de alongamento da nuvem coincide com a dos grupos e os Círculos apresentam perfis parecidos.

No caso 7, onde temos: variâncias iguais, correlação "DENTRO" zero e "ENTRE" alta positiva; vemos na Tabela 6.1.1 que, segundo os coeficientes *Bruto*, *Dentro* e *Total* a variável X teria mais influência que Y na separação dos grupos. No entanto esta diferença é também pequena (menor de 0,03) pelo qual trataremos de fazer uso das correlações entre originais e discriminantes para tirar uma conclusão sobre isto.

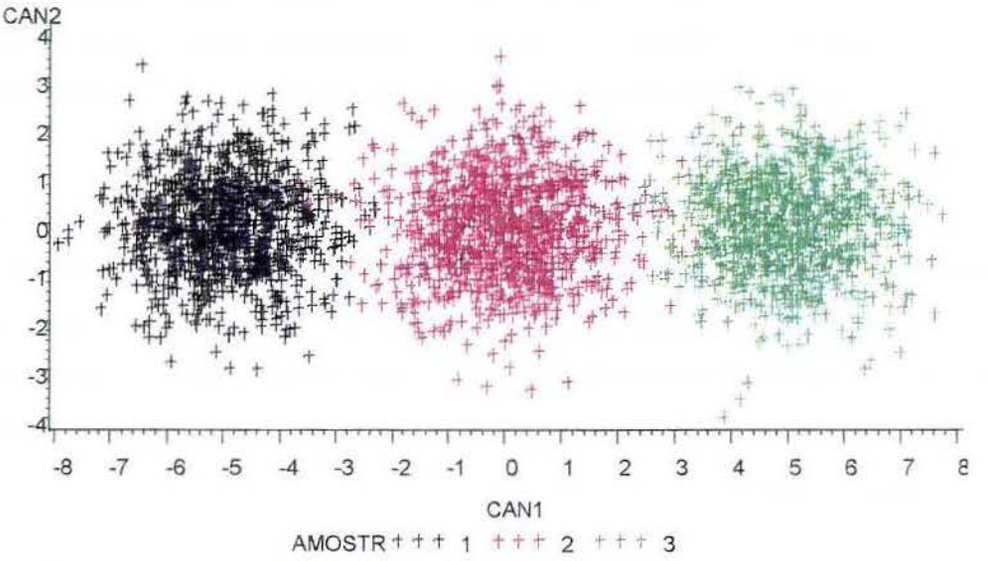


Figura 6.1.1: Gráfico das variáveis discriminantes para o caso 1 bivariado

Se observamos os gráficos das correlações para o Caso 7, na Tabela 6.1.2, vemos as boas representações de X e Y nos Círculos das Correlações. As contribuições de ambas as variáveis à separação são altas, maiores que 0,95,

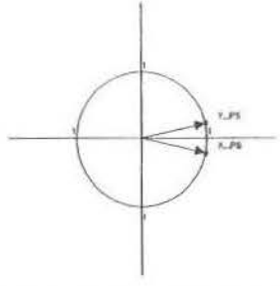
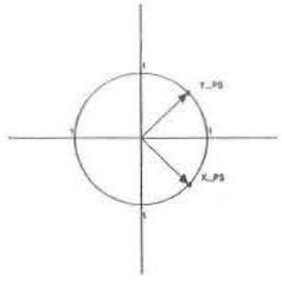
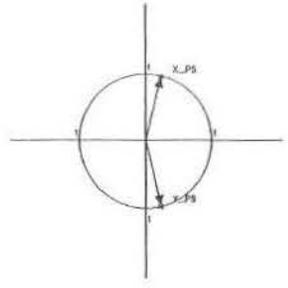
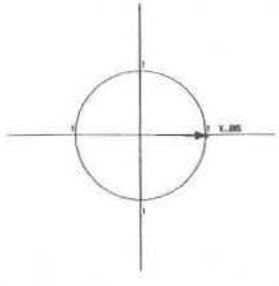
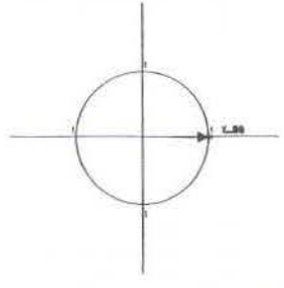
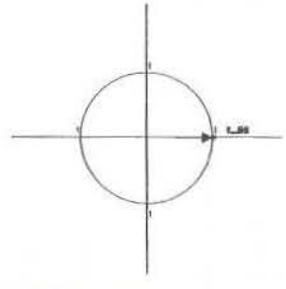
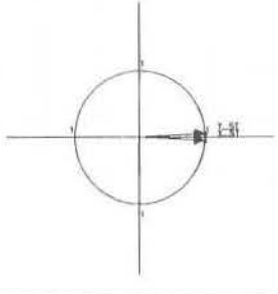
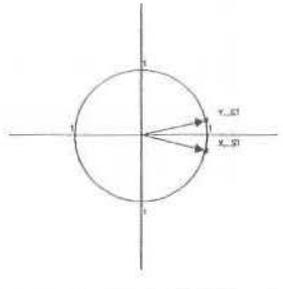
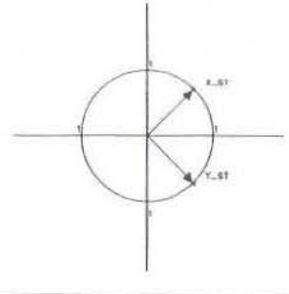
pelas correlações *Entre* os grupos. *Dentro* dos grupos acompanha, mas não com a mesma intensidade, pois $corr_D(X, \mathbf{can}_1) = 0{,}71$ e $corr_D(Y, \mathbf{can}_1) = 0{,}70$; ao mesmo tempo as correlações *Dentro* para **can2** se comportam praticamente iguais, $corr_D(X, \mathbf{can}_2) = -0{,}70$ e $corr_D(Y, \mathbf{can}_2) = 0{,}71$, o que sugere que a estrutura de correlação entre originais e discriminantes *Dentro* dos grupos acompanha da mesma forma as duas variáveis discriminantes. A direção de **can1** neste caso apresenta correlações com as variáveis originais apesar de no haver nos dados originasi.

Se somamos ao comportamento anterior o ângulo reto que aparece entre as variáveis X e Y no círculo com as correlações *Dentro* , e que indica que as variáveis X e Y são não correlacionadas "DENTRO" dos grupos, fica clara a configuração arredondada das nuvens.

Já no caso 13, os coeficientes não mostram diferenciação alguma para o caso da **can1** . Segundo isto as variáveis têm o mesmo comportamento; o confuso é que, somente olhando os coeficientes, este comportamento semelhante pode significar que as duas influenciam igual na separação dos grupos ou que, pelo contrário, nenhuma das duas aporta nada. Para tentarmos esclarecer isto vamos nos Círculos das Correlações na Tabela 6.1.2, Caso 13 a), b) e c).

No caso da estrutura de correlação *Entre* vemos que as duas variáveis acompanham perfeitamente a direção da **can1** , significando que esta variável discriminante consegue separar com sucesso os três grupos e que as duas variáveis contribuem fortemente à boa separação, como pode ser visto nos valores das correlações *Entre*, na Tabela 6.1.3, que neste caso são: $corr_E(X, \mathbf{can}_1) = 0{,}99$ e. $corr_E(Y, \mathbf{can}_1) = 0{,}99$.

Tabela 6.1.2: CÍRCULOS DAS CORRELAÇÕES para os casos 1-7-13

CASO 1 (=,+,+)	CASO 7(=,0,+)	CASO 13(=,-,+)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

Por outro lado, a estrutura de correlação *Dentro* não acompanha a orientação das médias na direção **can1**, como se pode observar na Tabela 6.1.2 Caso 13 a). Segundo esta figura podemos dizer que a orientação interna das nuvens vai na direção de **can2**, sendo que as duas variáveis estão favorecendo fortemente este comportamento, como pode ser visto pelas baixas correlações com **can1**, i.é, $corr_D(X, \mathbf{can}_1) = 0{,}22$ e $corr_D(Y, \mathbf{can}_1) = 0{,}23$; e as altas com **can2**, i.é, $corr_D(X, \mathbf{can}_2) = 0{,}97$ e $corr_D(Y, \mathbf{can}_2) = -0{,}97$. O sinal desta última correlação fala de um contraste que indica correlação negativa da variável X com Y.

O comportamento explicado anteriormente ajuda a representar a forma real das nuvens, alargadas na direção de **can2**, como pode ser visto na Figura 5.1.4 c), representação hipotética do caso 13.

Tabela 6.1.3: Correlações entre originais e discriminantes (Casos 1-7-13)

Estrutura de Correlação	Variáveis	CASO 1(=,+,+)		CASO 7(=,0,+)		CASO 13(=,-,+)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	0.9985	-0.0537	0.9733	-0.2294	0.6952	0.7187
	Y	0.9983	0.0581	0.9718	0.2356	0.6985	-0.7155
Entre	X	0.9999	-0.0017	0.9999	-0.0058	0.9999	0.0082
	Y	0.9999	0.0019	0.9999	0.0059	0.9999	-0.0081
Dentro	X	0.9755	-0.2198	0.7148	-0.6992	0.2291	0.9733
	Y	0.9715	0.2368	0.7049	0.7092	0.2310	-0.9729

Nestes três primeiros casos, onde mantivemos fixa a variação "DENTRO", $\sigma_{11} = \sigma_{22}$ e a correlação "ENTRE" como $\rho_{12}^{E} = +0{,}9$, podemos chegar às seguintes conclusões sobre a mudança na correlação "DENTRO":

⇒ Esta mudança não afeta a boa separação dos grupos conseguida por **can1** nos três casos, conforme podemos ver se percorremos estes casos na segunda linha da Tabela 6.1.2, correspondente à estrutura de correlação *Entre* os grupos. Aqui observa-se sempre as duas variáveis com altíssima correlação com **can1** e refletindo a mesma importância para a separação dos grupos.

⇒ Com a abertura crescente do ângulo entre as variáveis, conforme a correlação "DENTRO" muda de alta positiva a zero e posteriormente a alta negativa, nota-se um decrescimento no acompanhamento da estrutura de correlação *Dentro* à orientação das médias dos grupos.

⇒ Caso a parte é a estrutura de correlação *Total* para o conjunto completo dos dados onde, sem diferenciação dos grupos, mostra-se que a medida que a correlação "DENTRO" se opõe à "ENTRE" a correlação *Total* vai perdendo força, conforme ilustra a abertura dos vetores.

- **Casos 4, 10 e 16**

Na Tabela 6.1.4, a seguir, veremos três casos parecidos aos anteriores porém com a peculiaridade que neste caso teremos variâncias diferentes para as variáveis originais.

O caso que chamamos 4 corresponde a correlação "DENTRO" e "ENTRE" alta positiva e $\sigma_{11} = \sqrt{2}$, $\sigma_{22} = \sqrt{2}/2$ (variâncias escolhidas como explicado na seção 5.1.1). Neste caso o primeiro autovalor acumula o 99,99% da explicação da variação "ENTRE".

Dando uma olhada no coeficientes *Dentro* e *Bruto*, X parece ter mais importância que Y na separação dos grupos, no entanto da análise dos Círculos das Correlações vemos que a estrutura de correlação *Entre* os grupos, na Tabela 6.1.6, Caso 4 b) mostra que as duas variáveis têm correlações altas com **can1**, $corr_E(X, \mathbf{can}_1) = 0{,}99$ e $corr_e(Y, \mathbf{can}_1) = 0{,}99$, o que indica que as duas acompanham fortemente a direção de máxima variabilidade entre os grupos. Visto desde o ângulo que proporciona a estrutura de correlação *Dentro*, existe uma diferenciação da variável X no aporte à

separação dos grupos com respeito a Y, desde que $corr_D(X, \mathbf{can}_1) = 0{,}99$ e $corr_D(Y, \mathbf{can}_1) = 0{,}95$, porém as duas caminham na direção de **can1**. No caso da estrutura de correlação *Dentro*, Caso 4 a) observa-se que ela acompanha a orientação das médias dos grupos.

Da análise das estruturas de correlação podemos reconstruir a forma alongada das nuvens dentro dos grupos e entre eles, em ambos casos acompanhando a direção de **can1**.

Tabela 6.1.4: Coeficientes (Casos 4-10-16)

Variação	Variáveis	CASO 4 (≠,+,+)		CASO 10(≠,0,+)		CASO 16(≠,-,+)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	2.9143	-8.8863	4.1916	-0.0374	2.4745	0.8039
	Y	1.2816	8.9082	0.0009	1.0009	3.5163	-0.5278
Dentro	X	0.7022	-2.1412	0.9999	-0.0089	2.1288	0.6916
	Y	0.3209	2.2305	0.0009	1.0000	2.2136	-0.3322
Bruto	X	0.6007	-1.8319	0.8555	-0.0076	1.8213	0.5917
	Y	0.3851	2.6768	0.0010	1.1791	2.6728	-0.4012

No caso 10, que é onde temos correlação "DENTRO" zero, "ENTRE" alta positiva e $\sigma_{11} = \sqrt{2}$, $\sigma_{22} = \sqrt{2}/2$, o primeiro autovalor acumula 99,99% da explicação da variação entre os grupos, portanto **can1** consegue discriminar bem os grupos. Observando os coeficientes Discriminantes na Tabela 6.1.4, notamos uma diferença marcada, entre a magnitude do coeficiente da variável X com respeito ao Y, para os três coeficientes, *Total, Dentro* e *Bruto*, que levaria a pensar que X está a favor da separação dos grupos e Y contribui muito menos.

O Círculo das Correlações para a estrutura de correlação *Entre* os grupos (ver Tabela 6.1.6 Caso 10 b)) mostra que enquanto X têm uma correlação perfeita com

can1, igual a 1 (ver Tabela 6.1.5), Y têm correlação de 0,73. Portanto Y também contribui para a separação dos grupos, porém o que temos certeza é que a responsabilidade maior nesta separação é de X.

Quando analisado o comportamento das correlações para a estrutura de correlação *Dentro* dos grupos vemos que $corr_D(X, \mathbf{can}_1) = 1{,}00$ e $corr_D(Y, \mathbf{can}_1) = 0{,}009$, portanto dentro dos grupos X acompanha a orientação das médias, a favor de **can1** e Y a direção de **can2**, tendo $corr_D(Y, \mathbf{can}_2) = 0{,}99$. Este comportamento conjunto explica-se pela forma arredondada das nuvens.

Tabela 6.1.5: Correlações entre originais e discriminantes (Casos 4-10-16)

Estrutura de Correlação	Variáveis	CASO 4(≠,+,+)		CASO 10(≠,0,+)		CASO 16(≠,-,+)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	0.9994	-0.0343	1.0000	-0.0002	0.5259	0.8505
	Y	0.9969	0.0780	0.0373	0.9993	0.8011	-0.5985
Entre	X	0.9999	-0.0011	1.0000	-0.0000	0.9999	0.0083
	Y	0.9999	0.0026	0.7311	0.6822	0.9999	-0.0038
Dentro	X	0.9898	-0.1424	1.0000	-0.0009	0.1484	0.9889
	Y	0.9502	0.3116	0.0089	0.9999	0.3089	-0.9510

Tabela 6.1.6: CÍRCULOS DAS CORRELAÇÕES para os casos 4-10-16

CASO 4 (≠,+,+)	CASO 10(≠,0,+)	CASO 16(≠,-,+)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		

Por outro lado, no caso 16, com correlação "DENTRO" alta negativa, "ENTRE" alta positiva e $\sigma_{11} \neq \sigma_{22}$, os coeficientes Discriminantes, na Tabela 6.1.4, indicam que a ordem de importância seria Y e X para a **can1**.

No entanto, a estrutura de correlação *Entre* os grupos mostra que as duas variáveis têm altíssimas correlações com **can1**, $corr_E(X, \mathbf{can}_1) = 0{,}99$ e $corr_E(Y, \mathbf{can}_1) = 0{,}99$. Outra coisa é a estrutura de correlação *Dentro* dos grupos, pois neste caso as duas variáveis têm correlações baixas com **can1** e altas com **can2** (ver Tabela 6.1.5), o que indica que a configuração dentro dos grupos encontra-se em oposição à orientação das médias dos grupos. Este comportamento pode ser visto mais claramente olhando os Círculos das Correlações na Tabela 6.1.6.

Podemos concluir então, para estes três casos com variâncias diferentes e correlação "ENTRE" alta positiva, que:

⇒ as mudanças na correlação "DENTRO" provocam, no caso 10 com $\rho_{12}^{E} = 0$, uma diferenciação da importância das variáveis na separação dos grupos, neste caso a variável X é quem têm uma influência maior, ficando Y em um segundo lugar. Nos outros dois casos 4 e 16, observa-se que as duas variáveis têm um grande influência para efeitos da separação dos grupos.

⇒ Olhando a estrutura de correlação "DENTRO" dos grupos, conforme muda de alta positiva, para zero e depois para alta negativa, pode se observar que as variáveis participam menos, i.é, no caso 4 as duas variáveis aportam à separação dos grupos, já no caso 10 a Y não participa, sendo que por último, no caso 16, nenhuma das duas ajuda a separação dos grupos, exibindo altas correlações com **can2**.

⇒ O efeito das variâncias distintas aparece na magnitude dos coeficientes e nos Círculos de Correlação entre originais e canônicas.

⇒ O impacto das correlações é maior quando a nuvem dentro não tem direção preferencial, como por exemplo o caso 10.

- **Casos 2, 8 e 14**

A peculiaridade destes três casos é a ausência de correlação entre grupos e a igualdade das variâncias, i.e., $\sigma_{11} = \sigma_{22}$. O objetivo é analisar a influência de estruturas de correlação "DENTRO" diferentes, ou seja, para os três casos estudados: alta positiva, zero e alta negativa.

No caso 2, o primeiro autovalor acumula 57,08% da variância entre grupos, portanto **can1** não produz uma boa separação dos três grupos, somente consegue separar um pouco o grupo 3 dos dois primeiros, que ficam completamente misturados, como podemos ver se projetamos os indivíduos na primeira discriminante. Isto pode ser visto no plano discriminante para este caso, apresentado a seguir na Figura 6.1.2.

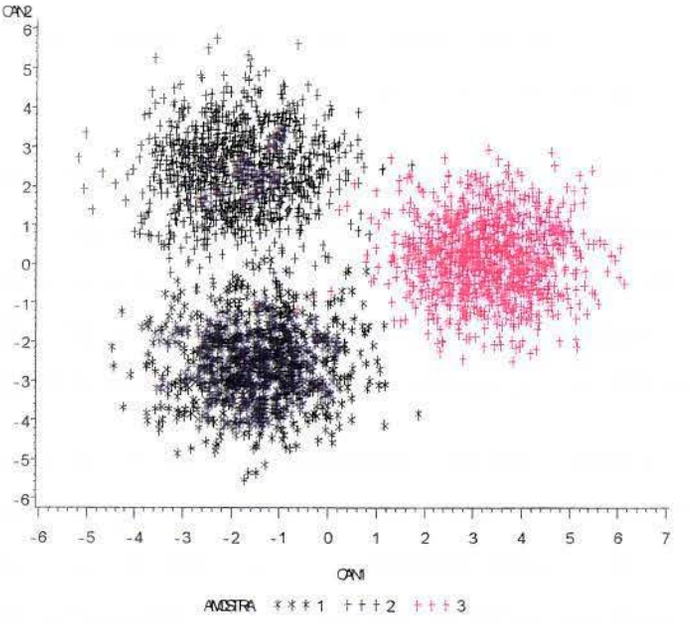


Figura 6.1.2: Gráfico das variáveis discriminantes para o caso 2

Observa-se na Tabela 6.1.7, que a magnitude dos coeficientes, no caso do *Bruto* e do *Dentro*, para a variável X é maior com respeito a Y em **can1** sendo o contrário em **can2**, i.é, Y é maior com respeito a X. Nota-se também a existência de um contraste no sinal dos coeficientes de X e Y em **can1**.

Tabela 6.1.7: Coeficientes (Casos 2-8-14)

Variação	Variáveis	CASO 2 (=,+,0)		CASO 8(=,0,0)		CASO 14(=,-,0)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-5.0445	0.9347	1.8888	1.2488	4.9789	1.0693
	Y	4.9070	1.3980	-1.2523	1.8616	4.9192	-1.2707
Dentro	X	-2.2113	0.4097	0.83418	0.5515	2.1830	0.4688
	Y	2.1628	0.6162	-0.5581	0.8297	2.1619	-0.5584
Bruto	X	-2.2494	0.4168	0.8485	0.5610	2.2207	0.4769
	Y	2.1826	0.6218	-0.5534	0.8226	2.1942	-0.5668

Dando uma olhada nos Círculos das Correlações podemos ver graficamente a existência deste contraste, Tabela 6.1.8 caso 2 b), pois acontece uma mudança de sinal da correlação com respeito a **can1**, cujos valores podem ser vistos na Tabela 6.1.9. Também a baixa correlação *Entre* com **can1** e a alta com **can2**, i.é, $corr_E(X, \mathbf{can}_2) = 0{,}96$ e $corr_E(Y, \mathbf{can}_2) = 0{,}98$, deixa ver que a Análise Discriminante de Fisher, neste caso, não consegue separar os grupos. A estrutura de correlação *Dentro* dos grupos, Tabela 6.1.8 caso 2 a), fica longe de querer acompanhar a orientação de **can1**.

É interessante assinalar que a repetição do mesmo padrão nas três estruturas de correlação para este caso indica que as variáveis X e Y estão positivamente correlacionadas entre elas, como indica o ângulo entre as duas.

Tabela 6.1.8: CÍRCULOS DAS CORRELAÇÕES para os casos 2-8-14

CASO 2 (=,+,0)	CASO 8(=,0,0)	CASO 14(=,-,0)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		

No caso 8, por outro lado, o primeiro autovalor acumula 51,49% da variância entre os grupos, pelo que **can1**, também neste outro caso, não conseguirá uma boa separação dos três grupos.

Na Tabela 6.1.7, mostra-se uma mudança no sinal dos coeficientes para X e Y, ou seja, existe um contraste entre as duas variáveis. A magnitude dos coeficientes favorece a X. Para esclarecermos sobre o comportamento da separação dos grupos vejamos na Tabela 6.1.8 caso 8 a), b) e c) os Círculos das Correlações.

Os três gráficos mostram o mesmo comportamento, i.e., contraste entre as variáveis X e Y em **can1**. Neste caso o plano **can1 - can2** mostra uma boa separação dos grupos. Isto pode ser visto graficamente na Figura 6.1.3.

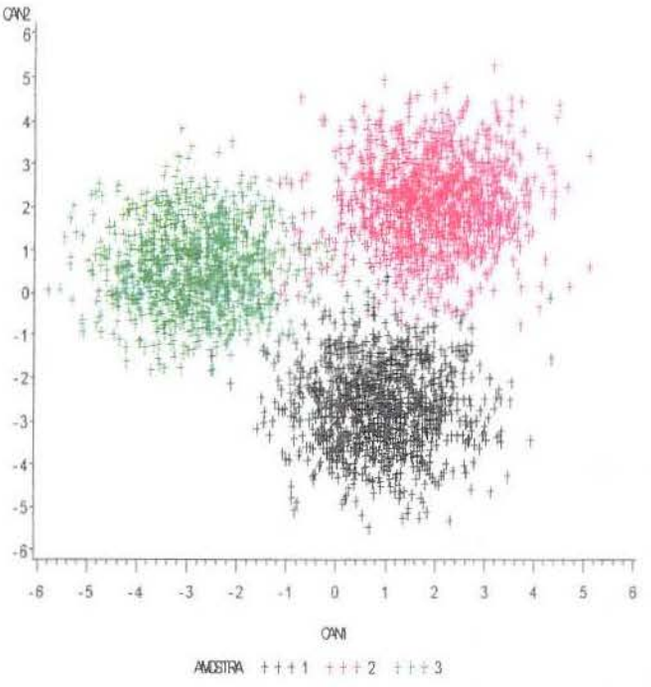


Figura 6.1.3: Gráfico do plano discriminante para o caso 8

Além disto o ângulo aproximadamente reto entre as variáveis X e Y indica ausência de correlação, fato este devido a que este caso foi construído precisamente com ausência de correlação entre grupos e dentro dos grupos.

No caso 14, por outro lado, coeficientes levemente maiores para X no caso de **can1**, ver Tabela 6.1.8, e de Y em **can2**, com um contraste entre as variáveis no último caso.

As correlações com **can1** são baixíssimas, i.e., por exemplo no caso da estrutura de correlação *Entre*, $corr_E(X, \mathbf{can}_1) = 0{,}28$ e $corr_E(Y, \mathbf{can}_1) = 0{,}23$, e altas com **can2**, i.e., $corr_E(X, \mathbf{can}_2) = 0{,}96$ e $corr_E(Y, \mathbf{can}_2) = -0{,}97$.

Tabela 6.1.9: Correlações entre originais e discriminantes (Casos 2-8-14)

Estrutura de Correlação	Variáveis	CASO 2(=,+,0)		CASO 8(=,0,0)		CASO 14(=,-,0)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-0.3052	0.9522	0.8358	0.5489	0.2780	0.9605
	Y	0.2040	0.9789	-0.5607	0.8280	0.2339	-0.9722
Entre	X	-0.3121	0.9500	0.8372	0.5467	0.2842	0.9587
	Y	0.2089	0.9779	-0.5629	0.8264	0.2393	-0.9709
Dentro	X	-0.2740	0.9617	0.8297	0.5581	0.2501	0.9682
	Y	0.1822	0.9832	-0.5515	0.8341	0.2099	-0.9777

Após a análise destes três casos podemos chegar às seguintes conclusões:

⇒ Neste caso , onde temos sempre correlação "ENTRE" os grupos nula, precisa-se das duas variáveis discriminantes para separar os grupos.

⇒ Os Círculos das Correlações, sem importar a estrutura de correlação, seja *Dentro, Entre* ou *Total,* são praticamente os mesmos. Ou seja, o valor das correlações entre originais e discriminantes pouco ou nada varia quando se muda o ponto de referência na hora de calculá-la.

- **Casos 5, 11 e 17**

Estes três casos são análogos aos anteriores, 2, 8 e 14, porém com variâncias diferentes para as variáveis X e Y. Nestes casos o primeiro autovalor acumula 57,02%, 56,89% e 72,37% da variância entre grupos, respectivamente. Junto com os três casos anteriores são os casos aonde o primeiro autovalor acumula menos porcentagem da variância entre grupos.

Para o caso 5, vemos na Tabela 6.1.10, que as magnitudes dos coeficientes Discriminantes para **can1** exibem um contraste entre os valores para as variáveis X e Y, sendo o coeficiente para Y maior em valor absoluto, para os três coeficientes.

Tabela 6.1.10: Coeficientes (Casos 5-11-17)

Variação	Variáveis	CASO 5(≠,+,0)		CASO 11(≠,0,0)		CASO 17(≠,-,0)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-4.9358	1.3538	-0.1017	2.2652	5.0274	1.3143
	Y	5.0454	0.9833	2.5402	0.0936	5.1510	-1.0867
Dentro	X	-2.1714	0.5955	-0.0448	0.9990	2.1633	0.5655
	Y	2.2100	0.4307	0.9993	0.0368	2.1879	-0.4615
Bruto	X	-1.8576	0.5095	-0.0383	0.8546	1.8507	0.4838
	Y	2.6523	0.5169	1.1783	0.0434	2.6413	-0.5572

Se olharmos as correlações, Tabela 6.1.11, vemos que são grandes para **can2**.

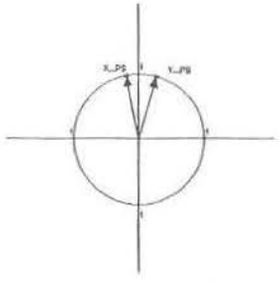
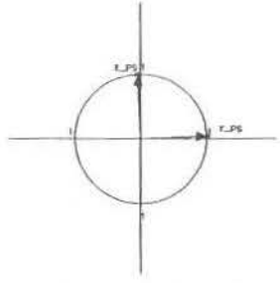
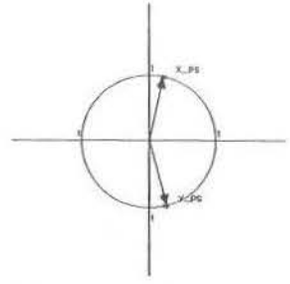
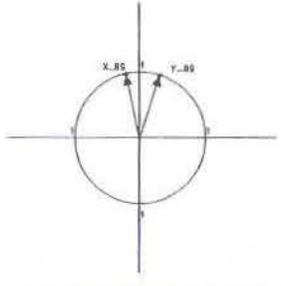
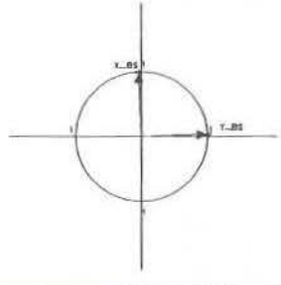
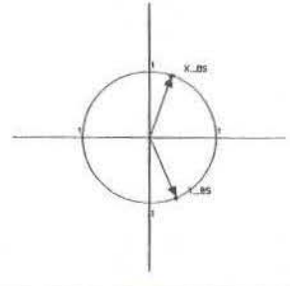
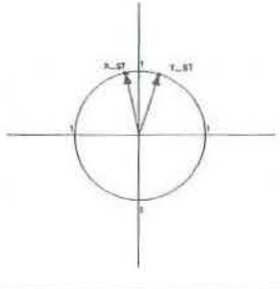
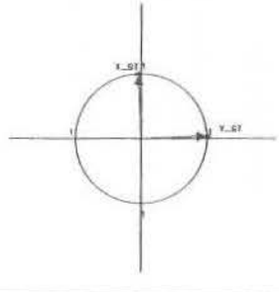
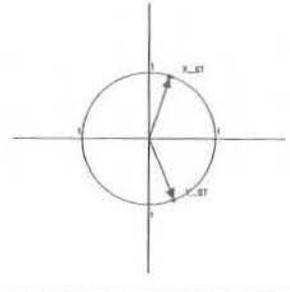
No caso 11 os coeficientes de Y para **can1** são de maior magnitude que os de X. Se olharmos nas correlações vemos que, de fato, a separação dos grupos é fortemente apoiada pela variável Y, parecendo X ter uma importância inferior neste caso. *Dentro* dos grupos acontece o mesmo comportamento.

Por outro lado, no caso 17, a magnitude dos coeficientes é muito parecida, embora os valores para Y sejam maiores.

Tabela 6.1.11: Correlações entre originais e discriminantes (Casos 5-11-17)

Estrutura de Correlação	Variáveis	CASO 5(≠,+,0)		CASO 11(≠,0,0)		CASO 17(≠,-,0)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-0.2139	0.9768	-0.0413	0.9991	0.3049	0.9523
	Y	0.2945	0.9556	0.9992	0.0400	0.3688	-0.9295
Entre	X	-0.2190	0.9757	-0.0423	0.9991	0.3231	0.9463
	Y	0.3012	0.9535	0.9992	0.0390	0.3897	-0.9209
Dentro	X	-0.1913	0.9815	-0.0368	0.9993	0.2064	0.9784
	Y	0.2645	0.9643	0.9989	0.0448	0.2529	-0.9674

Tabela 6.1.12: CÍRCULOS DAS CORRELAÇÕES para os casos 5-11-17

CASO 5(≠,+,0)	CASO 11(≠,0,0)	CASO 17(≠,-,0)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

Olhando detidamente estes casos podemos chegar às seguintes conclusões:

⇒ O efeito da mudança na correlação "DENTRO" faz com que: nenhuma das duas variáveis tenham importância quando é alta positiva, que somente uma das variáveis favoreça a separação; no caso de "DENTRO" também zero, e no caso de "DENTRO" alta negativa tampouco X e Y influenciam na separação, só que desta vez, além de ter correlações altas com **can1**, X e Y estão em contraste.

⇒ Dentro de cada caso os Círculos não mudam, eles observam o mesmo comportamento, ver Tabela 6.1.12.

⇒ Para todos estes casos, onde a correlação "ENTRE" é zero, precisa-se das duas dimensões para separar os grupos.

- **Casos 3, 9 e 15**

Estes casos têm em comum $\sigma_{11} = \sigma_{22}$ e "ENTRE" alta negativa. As peculiaridades são quanto à correlação "DENTRO" pois, o caso 3 tem positiva, o 9 zero e o 15 alta negativa. O primeiro autovalor para os casos 3 e 9 acumula 100% da variação entre grupos, o caso 15 acumula 99,99%, pelo que **can1** consegue uma boa separação dos grupos.

Tabela 6.1.13: Coeficientes (Casos 3-9-15)

Variação	Variáveis	CASO 3 (=,+,-)		CASO 9(=,0,-)		CASO 15(=,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-2.9520	0.6987	2.1709	2.1294	2.0588	8.8842
	Y	2.9780	0.6921	-2.1250	2.1321	-2.1422	8.8831
Dentro	X	-2.1885	0.5179	0.7139	0.7002	0.5040	2.1752
	Y	2.1905	0.5091	-0.7059	0.7082	-0.5234	2.1706
Bruto	X	-2.2262	0.5269	0.7262	0.7123	0.5127	2.2127
	Y	2.2105	0.5138	-0.6999	0.7022	-0.5313	2.2031

Na Tabela 6.1.13, caso 3, vemos que os coeficientes apenas se diferenciam em valor absoluto, i.é, a diferença que observa-se diz respeito aos sinais dos coeficientes invertidos para as duas variáveis. Os Círculos das Correlações mostram mais claramente este comportamento.

Na Tabela 6.1.15, caso 3 b) observam-se as variáveis com correlações iguais mas com sinais opostos, i.é, $corr_E(X, \mathbf{can_1}) = 0{,}99$ e $corr_E(Y, \mathbf{can_1}) = -0{,}99$, como pode ser visto na Tabela 6.1.14. Isto diz respeito à direção de **can1** neste caso (paralela à reta $Y + X = 1$ para este caso). Por outro lado a estrutura de correlação *Dentro* não acompanha a direção das médias dos grupos, pelas baixas correlações de X e Y com **can1**, -0,22 e 0,23, e as altas com **can2**, 0,97 e 0,97, respectivamente.

O exame do gráfico, caso 3 c), para a estrutura de correlação *Total* (sem considerar a estrutura de grupos) revela uma idéia do que acontece com estruturas de correlações, *Dentro* e *Entre*, opostas ao mostrar que não existirá influência diferenciada de nenhuma das duas variáveis, significando que a amostra total têm configuração arredondada.

Tabela 6.1.14: Correlações entre originais e discriminantes (Casos 3-9-15)

Estrutura de Correlação	Variáveis	CASO 3(=,+,-)		CASO 9(=,0,-)		CASO 15(=,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-0.6920	0.7218	0.9726	0.2321	0.9983	0.0574
	Y	0.6985	0.7155	-0.9714	0.2371	-0.9984	0.0551
Entre	X	-0.9999	0.0087	0.9999	0.006087	0.9999	0.0021
	Y	0.9999	0.0086	-0.9999	0.0062	-0.9999	0.0020
Dentro	X	-0.2263	0.9740	0.7082	0.7059	0.9721	0.2344
	Y	0.2303	0.9731	-0.7002	0.713890	-0.9741	0.2257

No caso 9, a leve diferença dos coeficientes não parece estar significando uma maior influência da variável X na separação dos grupos. Indo para os Círculos das

Correlações vemos que **can1** consegue uma excelente separação dos grupos, influenciada de igual maneira por ambas variáveis, sendo que elas apresentam um contraste na correlação com **can1**, i.é, $corr_E(X,\mathbf{can}_1)=0{,}99$ e $corr_E(Y,\mathbf{can}_1)=-0{,}99$, como pode ser visto na Tabela 6.1.14.

Já na figura caso 9 a), Tabela 6.1.15, os valores das correlações *Dentro* dos grupos são praticamente iguais para ambas variáveis. Isto está mostrando uma configuração circular dentro dos grupos, conseqüência da correlação zero, neste caso.

Para o caso 15 os coeficientes, na Tabela 6.1.13, mostram uma leve inclinação (menos que 0,02 em valor absoluto) a maior influencia por parte de Y. Ajudando-os dos Círculos das Correlações para chegar a uma conclusão, vemos, na Tabela 6.1.15 caso 15 b), que se consegue uma excelente separação entre os grupos, com altas correlações *Entre* de X e Y com **can1**, valores que podem ser vistos na Tabela 6.1.14.

A diferença dos casos 3 e 9, a estrutura de correlação *Dentro* dos grupos, cuja representação gráfica pode ser vista na Tabela 6.1.15 caso 15 a), sim acompanha a orientação das médias dos grupos, mostrando também um contraste entre as correlações das variáveis X e Y.

Comparando estes três casos podemos dizer que:

⇒ Os coeficientes, na Tabela 6.1.13, não oferecem informação clara do que acontece com os dados originais.

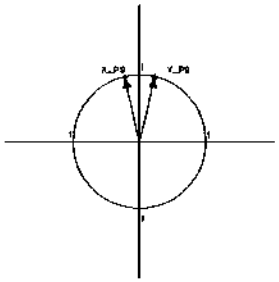
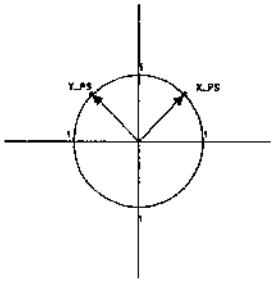
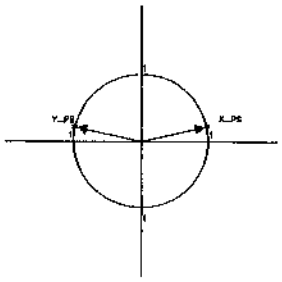
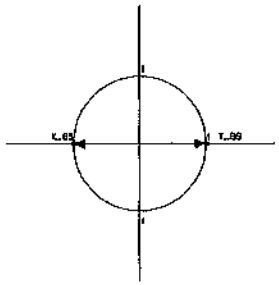
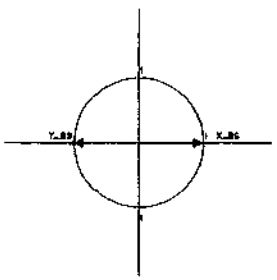
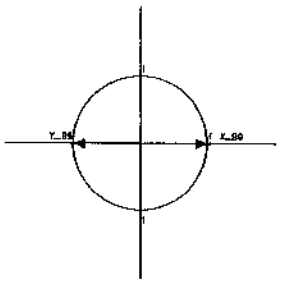
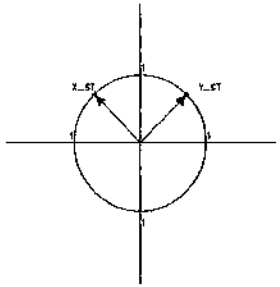
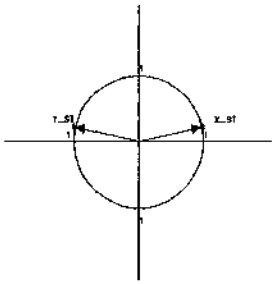
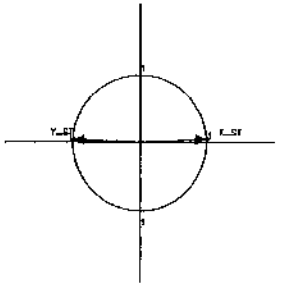
⇒ **can1** garante uma boa separação dos grupos relativo aos Círculos referentes à estrutura de correlação *Entre*. As correlações *Entre* não foram afeitadas.

⇒ A estrutura de correlação *Dentro* vão de 3-9-15 refletindo o acréscimo devido ao alongamento da nuvem dentro dos grupos ir da discordância, caso 3, para a concordância, caso 15.

A estrutura de correlação *Total* dá uma idéia do que acontece com nuvem de todos os pontos . Se observamos bem podemos interpretar como que a nuvem dos 3000 pontos vai mudando o seu aspecto de circular para bem alongada, conforme a

correlação das variáveis originais, *Dentro*, vai mudando de alta positiva para zero e por último para alta negativa.

Tabela 6.1.15: CÍRCULOS DAS CORRELAÇÕES para os casos 3-9-15

CASO 3 (=,+,-)	CASO 9(=,0,-)	CASO 15(=,-,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

- **Casos 6, 12 e 18**

Na Tabela 6.1.16, estão os coeficientes Discriminantes de Fisher para os casos 6, 12 e 18, análogos aos anteriores, 3, 9 e 16, com a diferença de que neste caso $\sigma_{11} \neq \sigma_{22}$.

Tabela 6.1.16: Coeficientes (Casos 6-12-18)

Variação	Variáveis	CASO 6(≠,+,-)		CASO 12(≠,0,-)		CASO 18(≠,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-2.4846	0.8038	4.1850	0.0374	2.7814	8.8476
	Y	3.5380	0.5234	-0.0169	1.0007	-1.4174	8.8660
Dentro	X	-2.1440	0.6936	0.9999	0.0089	0.6713	2.1353
	Y	2.2291	0.3298	-0.0169	0.9998	-0.3533	2.2103
Bruto	X	-1.8342	0.5934	0.8555	0.0076	0.5743	1.8268
	Y	2.6753	0.3958	-0.0199	1.1789	-0.4266	2.6688

No caso 6, os coeficientes, Tabela 6.1.16, favorecem à variável Y e mostram contraste com X. As correlações de X e Y com as discriminantes, na Tabela 6.1.18 caso 6 b), observa-se uma perfeita separação dos grupos onde ressalta um forte contraste entre as variáveis X e Y. Na estrutura de correlação *Dentro* vemos como ela não acompanha a direção das médias dos grupos, pois apresenta altas correlações com **can2**, i.é, $corr_D(X, \mathbf{can}_2) = 0{,}99$ e $corr_D(Y, \mathbf{can}_2) = 0{,}95$, aqui percebemos uma maior influencia da variável X neste comportamento.

Respeito ao caso 12, que é aquele onde os grupos têm correlação *Dentro* zero, se olharmos nos coeficientes, na Tabela 6.1.16, vemos em todos os casos a diferença marcada entre as magnitudes de X e Y, onde é X quem têm os maiores valores. Por outro lado, o comportamento observado na estrutura de correlação *Dentro* dos grupos é diferente pois enquanto X ajuda dentro dos grupos a acompanhar a orientação das

médias, Y faz exatamente o efeito contrário, como dizem os valores de correlação onde $corr_D(X, \mathbf{can}_1) = 0{,}99$ e $corr_D(Y, \mathbf{can}_2) = 0{,}99$ (ver Tabela 6.1.17, caso 12).

Aprofundando na estrutura de correlações *Entre* vemos, além do contraste entre X e Y, a superioridade de X na hora da separação dos grupos, pois temos $corr_E(X, \mathbf{can}_1) = 1{,}00$ e $corr_E(Y, \mathbf{can}_1) = -0{,}73$, isto pode verse mais claramente na Tabela 6.1.18 caso 12 b). Y por seu lado, não acompanha a direção de nenhuma das discriminantes, na estructura *Entre*.

O caso 18, Tabela 6.1.16, mostra que, segundo os coeficientes, é a variável X quem têm coeficientes de maior magnitude. Os Círculos das Correlações, no entanto, revelam que as variáveis originais conseguem correlações muito altas com **can1**, portanto as duas contribuem muito na separação dos grupos (ver Tabela 6.1.18, caso 18 b)). X contribui, em contraste com Y, no acompanhamento da orientação das médias, como podemos ver na estrutura de correlação *Dentro* dos grupos, Tabela 6.1.18 caso 18 a), com correlações com **can1** de 0,98 e –0,95, para X e Y, respectivamente.

Tabela 6.1.17: Correlações entre originais e discriminantes (Casos 6-12-18)

Estrutura de Correlação	Variáveis	CASO 6(≠,+,-)		CASO 12(≠,0,-)		CASO 18(≠,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	-0.5213	0.8533	0.9999	0.0040	0.9992	0.0381
	Y	0.8005	0.5992	-0.0374	0.9993	-0.9971	0.0747
Entre	X	-0.9999	0.0086	1.0000	0.0001	0.9999	0.0013
	Y	0.9999	0.0039	-0.7304	0.6829	-0.9999	0.0026
Dentro	X	-0.1463	0.9892	0.9998	0.0169	0.9874	0.1578
	Y	0.3078	0.9514	-0.0089	0.9999	-0.9539	0.2999

Para estes três casos podemos fazer as seguintes observações:

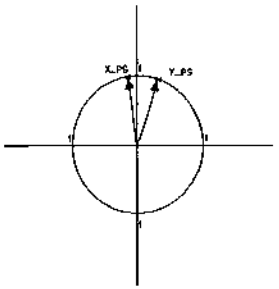
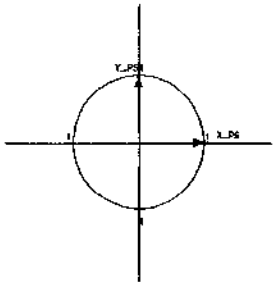
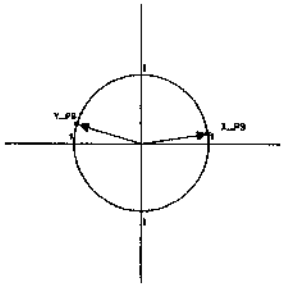
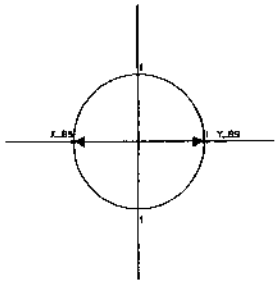
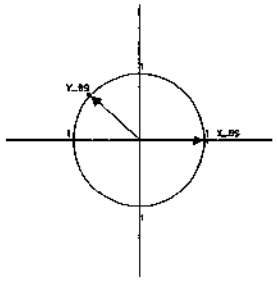
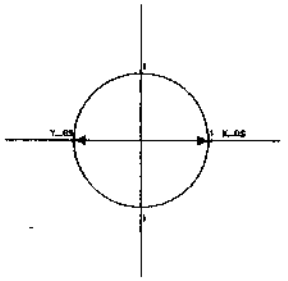
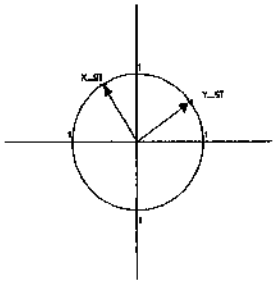
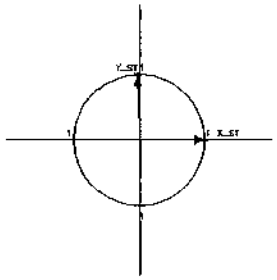
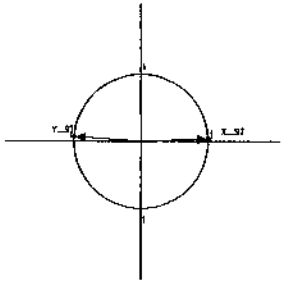
⇒ A mudança na correlação entre X e Y dentro dos grupos não influencia, para estes casos 6, 12 e 18, na separação dos grupos conseguida pela variável **can1**. Todos conseguem uma excelente separação, comparativamente no caso 12 a variável Y teve uma correlação *Entre* menor com **can1** do que nos casos 6 e 18.

⇒ Quanto à estrutura de correlação *Dentro* entre as discriminantes e as originais podemos dizer que a estrutura de correlação "DENTRO" dos grupos não acompanha a direção da **can1**, no caso 6, acompanha parcialmente, no caso 12, pois X acompanha perfeitamente enquanto Y acompanha a direção da segunda canônica; e por último no caso 18 a *Dentro* acompanha muito bem.

⇒ A estrutura de correlação *Total*, ajuda a entender o comportamento dos dados, vistos sem estrutura de grupos. Os valores de correlação *Total* observados indicam que no caso 6 a nuvem terá uma forma arredondada, no 18 forma elíptica, devido às altas correlações das duas variáveis com **can1**, e no caso 12 uma forma intermédia entre estas duas.

⇒ É importante observar que as diferenças de variância tem mais impacto nas correlações quando a nuvem dentro dos grupos tem estrutura fraca de correlação.

Tabela 6.1.18: CÍRCULOS DAS CORRELAÇÕES para os casos 6-12-18

CASO 6(≠,+,-)	CASO 12(≠,0,-)	CASO 18(≠,-,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

6.1.2 Comparando a influência de mudanças na correlação "ENTRE" os grupos, mantendo fixa a correlação "DENTRO".

Nesta seção o interesse é ir variando a correlação "ENTRE" os grupos e ver como isto influencia na magnitude dos coeficientes Discriminantes e nas correlações entre originais e discriminantes, mantendo fixa a correlação "DENTRO" dos grupos.

Como na seção 6.1.1 foram analisados os casos um a um, nesta seção vamos diretamente à comparação entre os casos de interesse.

- **Casos 1, 2 e 3**

Na Tabela 6.1.19, temos os coeficientes Discriminantes de Fisher para os casos 1, 2 e 3. Estes casos têm em comum correlação "DENTRO" dos grupos alta positiva e variâncias iguais para X e Y. Os primeiros autovalores acumulam, para cada caso, 99,99, 57,08 e 100% da variância entre os grupos, só será de interesse a análise de **Can1**.

Tabela 6.1.19: Coeficientes (Casos 1-2-3)

Variação	Variáveis	CASO 1 (=,+,+)[2]		CASO 2 (=,+,0)		CASO 3 (=,+,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	2.1792	-8.9398	-5.0445	0.9347	-2.9520	0.6987
	Y	2.0153	8.9420	4.9070	1.3980	2.9780	0.6921
Dentro	X	0.5326	-2.1849	-2.2113	0.4097	-2.1885	0.5179
	Y	0.4944	2.1939	2.1628	0.6162	2.1905	0.5091
Bruto	X	0.5418	-2.2226	-2.2494	0.4168	-2.2262	0.5269
	Y	0.4989	2.2139	2.1826	0.6218	2.2105	0.5138

[2] Os sinais entre parêntesis servem para dar uma indicação rápida de se estamos trabalhando com variâncias iguais ou diferentes e de quais as correlações dentro e entre que se correspondem com cada caso. Exemplo: (=,+,0), (≠,+,-) correspondem-se com: variâncias iguais, dentro +0,9 e entre zero; variâncias diferentes, dentro +0,9 e entre –0,9, respectivamente.

Analisando comparativamente estes casos podemos chegar às seguintes conclusões sobre o comportamento dos coeficientes e das correlações entre originais e discriminantes no caso da mudança na correlação "ENTRE":

⇒ Segundo os valores dos coeficientes; para o caso 1, "ENTRE" alta positiva, Se observamos os coeficientes *Bruto* e *Dentro* vemos que a variável X parece ser mais importante na separação dos grupos, embora esta diferença não seja muito marcada, no caso 2, "ENTRE" zero, acontece o mesmo que no caso 1 a exceção do contraste existente entre X e Y, como pode se notar pela mudança de sinal. No caso 3, "ENTRE" alta negativa, seguindo a magnitude dos coeficientes ficamos sem informação, a não ser que X e Y influenciam, ou não influenciam, com a mesma força, também elas estão em contraste. Procurando mais informação analisaremos as correlações entre as discriminantes e as originais, Tabela 6.1.20 e para facilitar a interpretação destes valores vamos nos Círculos das Correlações na Tabela 6.1.21.

Tabela 6.1.20: Correlações entre originais e discriminantes (Casos 1-2-3)

Estrutura de Correlação	Variáveis	CASO 1(=,+,+)		CASO 2(=,+,0)		CASO 3(=,+,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	0.9985	-0.0537	-0.3052	0.9522	-0.6920	0.7218
	Y	0.9983	0.0581	0.2040	0.9789	0.6985	0.7155
Entre	X	0.9999	-0.0017	-0.3121	0.9500	-0.9999	0.0087
	Y	0.9999	0.0019	0.2089	0.9779	0.9999	0.0086
Dentro	X	0.9755	-0.2198	-0.2740	0.9617	-0.2263	0.9740
	Y	0.9715	0.2368	0.1822	0.9832	0.2303	0.9731

⇒ A estrutura de correlação *Entre* os grupos mostra, a través das altas correlações com **Can1**, como quando a correlação "ENTRE" é alta positiva ou alta negativa, casos 1 e 3, **Can1** consegue uma excelente separação dos

grupos, existindo um contraste entre X e Y quando as correlações "DENTRO" e "ENTRE" são opostas, o que acontece no caso 3. No caso 2, às baixas correlações de X e Y com **Can1**, i.é, $corr_E(X, \mathbf{can}_1) = -0{,}31$ e $corr_E(Y, \mathbf{can}_1) = 0{,}21$, indicam que são necessárias as duas variáveis canônicas para mostrar a boa separação dos grupos. A estrutura de correlação *Dentro*, por outro lado, só acompanha a orientação das médias dos grupos no caso 1, nos outros dois casos a orientação é contrária.

Pelo padrão observado também podemos ressaltar os seguintes aspectos:

⇒ A correlação *Entre* tem um efeito dominante nas correlações das variáveis originais e a primeira canônica.

⇒ Baixa correlação *Entre* não implica pouca discriminação, apenas revela a falta de direção preferencial ao longo dos grupos.

Tabela 6.1.21: CÍRCULOS DAS CORRELAÇÕES para os casos 1-2-3

CASO 1 (=,+,+)	CASO 2 (=,+,0)	CASO 3 (=,+,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		

- **Casos 4, 5 e 6**

Os casos 4, 5 e 6 têm em comum correlação "DENTRO" dos grupos alta positiva, igual aos casos 1, 2 e 3, porém as variâncias entre X e Y são diferentes. Esta é a diferença com o caso anterior. Comparando estes três casos podemos chegar às seguintes conclusões:

⇒ Usando os coeficientes, na Tabela 6.1.22, não se consegue entender com claridade o que está acontecendo com a mudança de correlação "ENTRE". Vemos que no caso 4 a magnitude do coeficiente para X é maior a diferença dos casos 5 e 6, onde o coeficiente maior é o da variável Y. No entanto mostram-se contrastes entre X e Y na primeira variável discriminante nos casos 5 e 6.

Tabela 6.1.22: Coeficientes (Casos 4-5-6)

Variação	Variáveis	CASO 4 (≠,+,+)		CASO 5(≠,+,0)		CASO 6(≠,+,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	2.9143	-8.8863	-4.9358	1.3538	-2.4846	0.8038
	Y	1.2816	8.9082	5.0454	0.9833	3.5380	0.5234
Dentro	X	0.7022	-2.1412	-2.1714	0.5955	-2.1440	0.6936
	Y	0.3209	2.2305	2.2100	0.4307	2.2291	0.3298
Bruto	X	0.6007	-1.8319	-1.8576	0.5095	-1.8342	0.5934
	Y	0.3851	2.6768	2.6523	0.5169	2.6753	0.3958

⇒ Os Círculos das Correlações esclarecem que nos casos 4 e 6, com "ENTRE" alta positiva e alta negativa, respectivamente, **can1** consegue separar bem os grupos. Nestes casos as variáveis X e Y influenciam fortemente para conseguir esta separação, pois acompanham a direção de **can1** ao longo dos grupos, com a peculiaridade da existência de um contraste entre elas no caso

6. No caso 5, as baixas correlações com **can1** revelam a falta de direção preferencial ao longo dos grupos, devido à própria configuração deste caso.

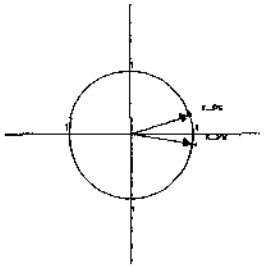
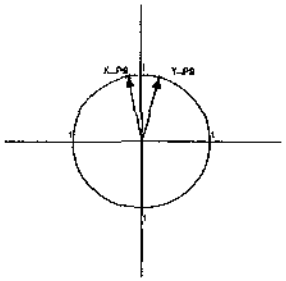
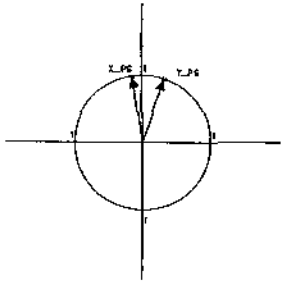
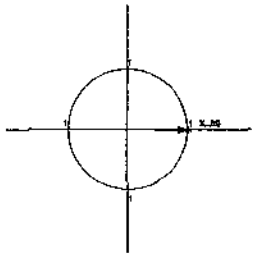
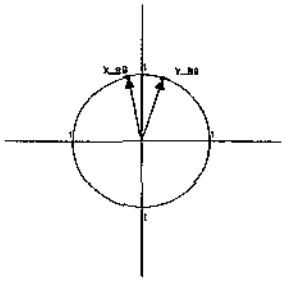
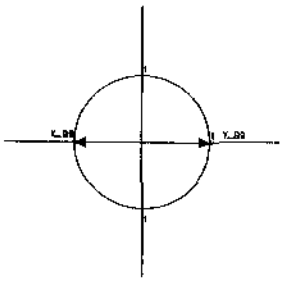
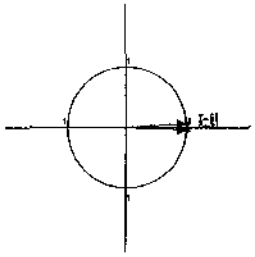
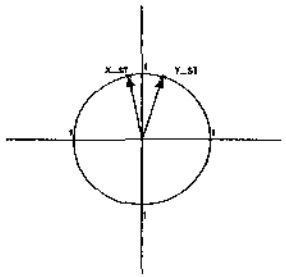
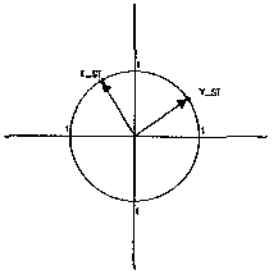
⇒ Quanto à estrutura de correlação *Dentro*, no caso 4, podemos dizer que ela acompanha à orientação da **can1**, no sendo assim no caso 6 onde vá em direção oposta.

⇒ Podemos chegar à conclusão então que, quando temos variâncias diferentes, as duas variáveis favorecem a separação dos grupos quando temos "DENTRO" e "ENTRE" altas positivas as duas, caso 4, e positiva e negativa, como no caso 6.

Tabela 6.1.23: Correlações entre originais e discriminantes (Casos 4-5-6)

Estrutura de Correlação	Variáveis	CASO 4(≠,+,+)		CASO 5(≠,+,0)		CASO 6(≠,+,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	0.9994	-0.0343	-0.2139	0.9768	-0.5213	0.8533
	Y	0.9969	0.0780	0.2945	0.9556	0.8005	0.5992
Entre	X	0.9999	-0.0011	-0.2190	0.9757	-0.9999	0.0086
	Y	0.9999	0.0026	0.3012	0.9535	0.9999	0.0039
Dentro	X	0.9898	-0.1424	-0.1913	0.9815	-0.1463	0.9892
	Y	0.9502	0.3116	0.2645	0.9643	0.3078	0.9514

Tabela 6.1.24: CÍRCULOS DAS CORRELAÇÕES para os casos 4-5-6

CASO 4 (≠,+,+)	CASO 5(≠,+,0)	CASO 6(≠,+,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

- **Casos 7, 8 e 9**

Os casos 7, 8 e 9 têm variâncias iguais e correlações "DENTRO" dos grupos nula. Os primeiros autovalores nestes casos acumulam, respectivamente, 100, 51,49 e 100% da variância entre os grupos.

Ao fazer uma análise comparativa da influencia de mudanças na correlação "ENTRE" notam-se os seguintes aspectos importantes:

⇒ Os casos 7 e 9 não mostram praticamente diferenças entre os coeficientes (ver Tabela 6.1.25), só podemos falar de uma magnitude levemente maior dos coeficientes para a variável X. Nos caso 8 e 9 existe contraste entre X e Y. O caso 8 mostra coeficientes maiores para X, respeito a Y.

Tabela 6.1.25: Coeficientes (Casos 7-8-9)

Variação	Variáveis	CASO 7(=,0,+)		CASO 8(=,0,0)		CASO 9(=,0,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	2.1618	-2.1486	1.8888	1.2488	2.1709	2.1294
	Y	2.1049	2.1519	-1.2523	1.8616	-2.1250	2.1321
Dentro	X	0.7092	-0.7049	0.83418	0.5515	0.7139	0.7002
	Y	0.6992	0.7148	-0.5581	0.8297	-0.7059	0.7082
Bruto	X	0.7215	-0.7171	0.8485	0.5610	0.7262	0.7123
	Y	0.6933	0.7087	-0.5534	0.8226	-0.6999	0.7022

⇒ A estrutura de correlação *Entre*, para os casos 7 e 9, indica que as duas variáveis influenciam na separação entre os grupos. No caso 8 as correlações *Entre* são: $corr_E(X, \mathbf{can}_1) = 0{,}84$ e $corr_E(Y, \mathbf{can}_1) = -0{,}56$, $corr_E(X, \mathbf{can}_2) = 0{,}56$ e $corr_E(Y, \mathbf{can}_2) = 0{,}82$. Este comportamento conjunto explica-se da forma arredondada das nuvens.

⇒ Nestes casos, com correlação "DENTRO" zero, a estrutura de correlação *Dentro* dos grupos, analisando conforme aos Círculos das Correlações e apoiando-nos nos valores da Tabela 6.1.26, comporta-se da seguinte forma: quando a correlação "ENTRE" é alta positiva (caso 7) a correlação *Dentro* dos grupos acompanha; quando a "ENTRE" é alta negativa a *Dentro* não acompanha. Por ultimo, quando a "ENTRE" é zero a correlação *Dentro* acompanha parcialmente.

Tabela 6.1.26: Correlações entre originais e discriminantes (Casos 7-8-9)

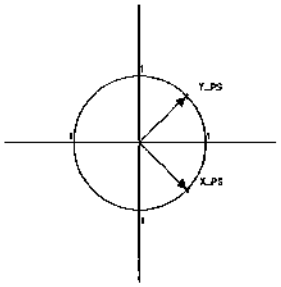
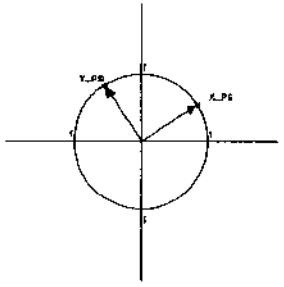
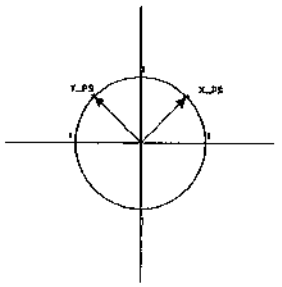
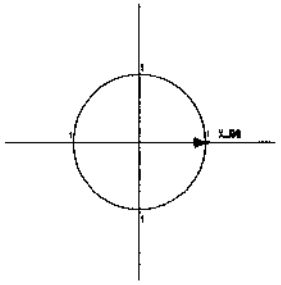
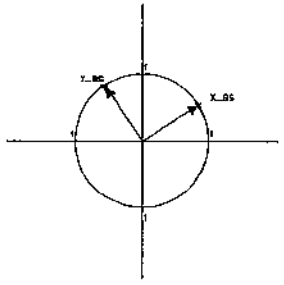
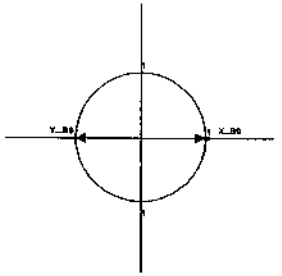
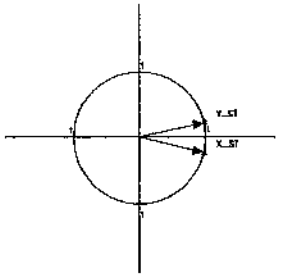
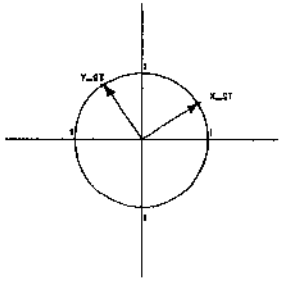
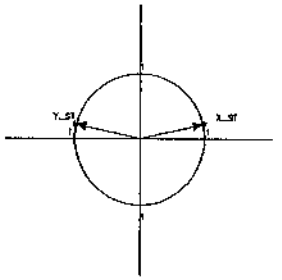
Estrutura de Correlação	Variáveis	CASO 7(=,0,+)		CASO 8(=,0,0)		CASO 9(=,0,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	0.9733	-0.2294	0.8358	0.5489	0.9726	0.2321
	Y	0.9718	0.2356	-0.5607	0.8280	-0.9714	0.2371
Entre	X	0.9999	-0.0058	0.8372	0.5467	0.9999	0.0060
	Y	0.9999	0.0059	-0.5629	0.8264	-0.9999	0.0062
Dentro	X	0.7148	-0.6992	0.8297	0.5581	0.7082	0.7059
	Y	0.7049	0.7092	-0.5515	0.8341	-0.7002	0.7138

Além disto podemos fazer as seguintes observações:

⇒ Correlações baixas entre as originais e $\mathbf{can}_1$ podem ser devidas à ausência de orientação nos grupos.

⇒ Contraste aparecem por ausência de alongamento ao longo dos grupos ou por orientação de estrutura de correlação negativa ao longo dos grupos.

Tabela 6.1.27: CÍRCULOS DAS CORRELAÇÕES para os casos 7-8-9

CASO 7(=,0,+)	CASO 8(=,0,0)	CASO 9(=,+,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

- **Casos 10, 11 e 12**

Estes três casos são análogos aos anteriores, com a diferença de que $\sigma_{11} \neq \sigma_{22}$. Fazendo uma análise cuidadosa podemos chegar as seguintes conclusões:

⇒ Nos casos 10 e 12 os coeficientes para X são maiores em magnitude. Já no caso 11, Y parece ser a variável com maior influencia na separação dos grupos, segundo os coeficientes da Tabela 6.1.28.

Tabela 6.1.28: Coeficientes (Casos 10-11-12)

Variação	Variáveis	CASO 10(≠,0,+)		CASO 11(≠,0,0)		CASO 12(≠,0,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	4.1916	-0.0374	-0.1017	2.2652	4.1850	0.0374
	Y	0.0009	1.0009	2.5402	0.0936	-0.0169	1.0007
Dentro	X	0.9999	-0.0089	-0.0448	0.9990	0.9999	0.0089
	Y	0.0009	1.0000	0.9993	0.0368	-0.0169	0.9998
Bruto	X	0.8555	-0.0076	-0.0383	0.8546	0.8555	0.0076
	Y	0.0010	1.1791	1.1783	0.0434	-0.0199	1.1789

⇒ A estrutura de correlação *Entre*, na Tabela 6.1.29, mostra que as variáveis que têm coeficientes de magnitude maior, influenciam mais na separação dos grupos, manifestando-se através de correlações altíssimas com **can1**, i.é, $corr_E(X, \mathbf{can}_1) = 1{,}00$, nos casos 10 e 12 e $corr_E(Y, \mathbf{can}_1) = 0{,}99$, no caso 11. Por outro lado a variável Y, para 10 e 12, não influenciará igual que X nesta separação, dado que apresenta aproximadamente as mesmas correlações com as duas variáveis discriminantes, pelo que não tem direção preferencial, nestes casos. No caso 11 a variável X têm correlação quase zero com **can1**,

i.e., $corr_E(X, \mathbf{can}_1) = -0{,}04$. Nos casos 11 e 12 as variáveis X e Y estão em contraste.

Tabela 6.1.29: Correlações entre originais e discriminantes (Casos 10-11-12)

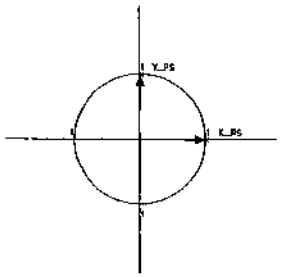
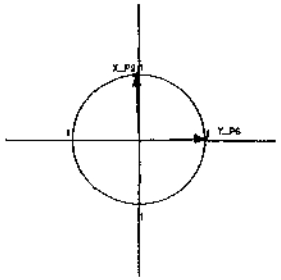
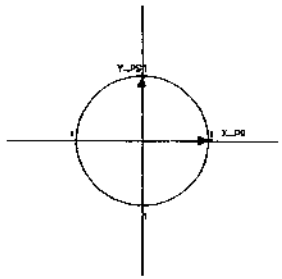
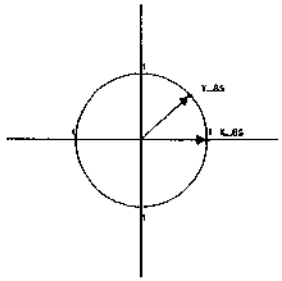
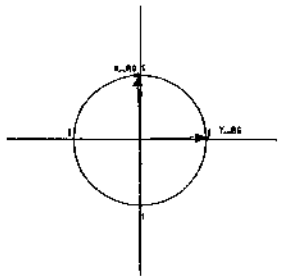
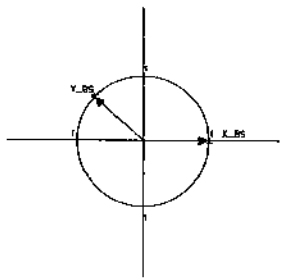
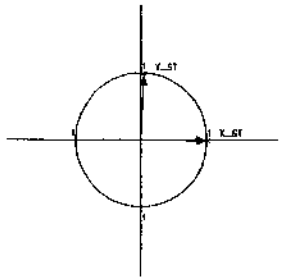
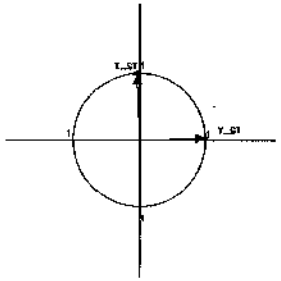
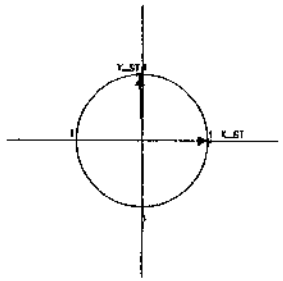
Estrutura de Correlação	Variáveis	CASO 10(≠,0,+)		CASO 11(≠,0,0)		CASO 12(≠,0,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	1.0000	-0.0002	-0.0413	0.9991	0.9999	0.0040
	Y	0.0373	0.9993	0.9992	0.0400	-0.0374	0.9993
Entre	X	1.0000	-0.0000	-0.0423	0.9991	1.0000	0.0001
	Y	0.7311	0.6822	0.9992	0.0390	-0.7304	0.6829
Dentro	X	1.0000	-0.0009	-0.0368	0.9993	0.9998	0.0169
	Y	0.0089	0.9999	0.9989	0.0448	-0.0089	0.9999

⇒ A estrutura de correlação *Dentro*, por outro lado exibe às variáveis X, com correlações 1,00 e 0.99 nos casos 10 e 12, e Y com correlação de 0,99 no caso 11, com **can1**. Em cada caso a outra variável têm correlações altas, desta vez com **can2**, e portanto oposta à direção da máxima separação dos grupos **can1**.

Além disto, observamos que:

⇒ Alterar variâncias em estruturas dentro dos grupos de ausência de correlação impacta as correlações entre originais e canônicas, favorecendo as direções de maior variabilidade dentro dos grupos com maior correlação com a primeira canônica.

Tabela 6.1.30: CÍRCULOS DAS CORRELAÇÕES para os casos 10-11-12

CASO 10(≠,0,+)	CASO 11(≠,0,0)	CASO 12(≠,0,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

- **Casos 13, 14 e 15**

Estes casos têm em comum variâncias iguais para X e Y e correlação "ENTRE" alta negativa entre elas. Desta vez os primeiros autovalores acumulam 100, 56,84 e 99,99% da variação entre grupos. Observando os coeficientes da Tabela 6.1.31, e os Círculos das Correlações, Tabela 6.1.33, com seus respectivos valores na Tabela 6.1.31, podemos fazer as seguintes observações:

⇒ Olhando os coeficientes Discriminantes, não dá para chegar em uma conclusão respeito a qual variável influencia mais, só para dizer que as duas têm coeficientes quase iguais pelo que elas parecem ter a mesma influência na separação dos grupos.

Tabela 6.1.31: Coeficientes (Casos 13-14-15)

Variação	Variáveis	CASO 13(=,-,+)		CASO 14(=,-,0)		CASO 15(=,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	2.9412	0.6985	4.9789	1.0693	2.0588	8.8842
	Y	2.9542	-0.6953	4.9192	-1.2707	-2.1422	8.8831
Dentro	X	2.1724	0.5159	2.1830	0.4688	0.5040	2.1752
	Y	2.1734	-0.5115	2.1619	-0.5584	-0.5234	2.1706
Bruto	X	2.2098	0.5248	2.2207	0.4769	0.5127	2.2127
	Y	2.2060	-0.5192	2.1942	-0.5668	-0.5313	2.2031

⇒ Estudando as correlações entre originais e discriminantes vemos que nos casos onde temos correlação alta positiva e alta negativa "ENTRE", casos 13 e 15, a estrutura de correlação *Entre*, Tabela 6.1.33 b), exibe uma excelente separação dos grupos, com a única diferença de que X e Y estão em contraste no caso 15. No caso 14, onde temos correlação "ENTRE" nula, as correlações de X e Y com **can1** são baixas, indicando que a não preferencia

das variáveis canônicas por uma direção determinada (ver o esboço da Figura 5.1.6 b).

Tabela 6.1.32: Correlações entre originais e discriminantes (Casos 13-14-15)

Estrutura de Correlação	Variáveis	CASO 13(=,-,+)		CASO 14(=,-,0)		CASO 15(=,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	0.6952	0.7187	0.2780	0.9605	0.9983	0.0574
	Y	0.6985	-0.7155	0.2339	-0.9722	-0.9984	0.0551
Entre	X	0.9999	0.0082	0.2842	0.9587	0.9999	0.0021
	Y	0.9999	-0.0081	0.2393	-0.9709	-0.9999	0.0020
Dentro	X	0.2291	0.9733	0.2501	0.9682	0.9721	0.2344
	Y	0.2310	-0.9729	0.2099	-0.9777	-0.9741	0.2257

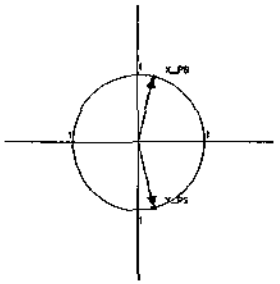
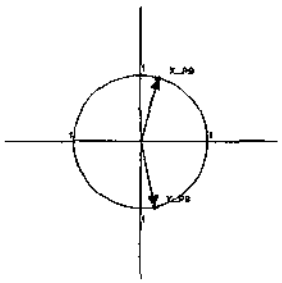
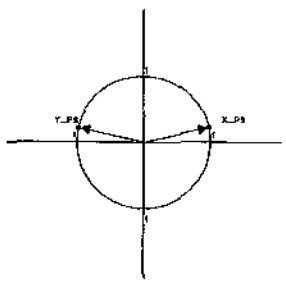
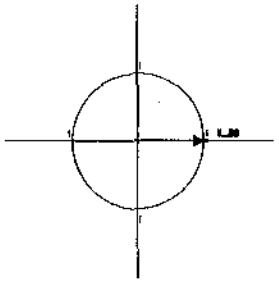
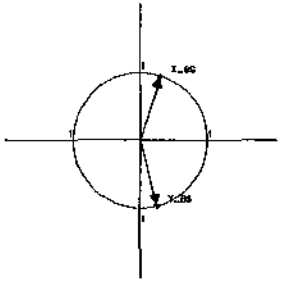
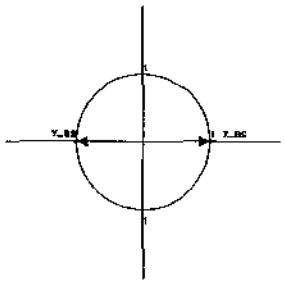
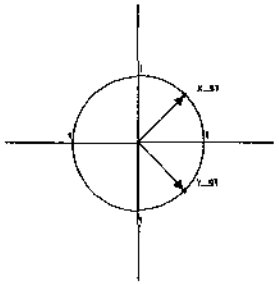
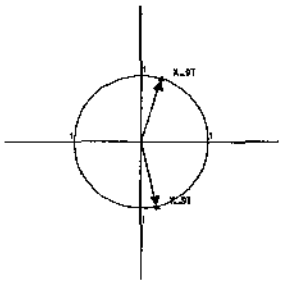
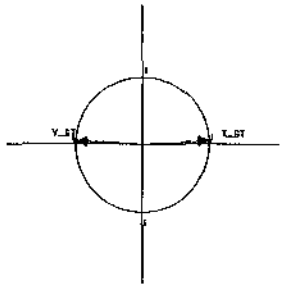
⇒ Por outro lado, a estrutura de correlação *Dentro* dos grupos mostra que só no caso 15 ela acompanha a orientação das médias, desta vez exibindo um contraste entre X e Y. Nos casos 13 e 14 a *Dentro* não consegue acompanhá-las, como mostram as baixas correlações com **can1**.

Portanto podemos observar que:

⇒ A estrutura de correlação das variáveis originais com a primeira canônica está determinada pela estrutura *Entre.*

⇒ A estrutura das correlações *Dentro* comparecem com correlações baixas devido ao fato do alongamento ao longo dos grupos ser discordante da estrutura da nuvem dentro dos grupos.

Tabela 6.1.33: CÍRCULOS DAS CORRELAÇÕES para os casos 13-14-15

CASO 13(=,-,+)	CASO 14(=,-,0)	CASO 15(=,-,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

- **Casos 16, 17 e 18**

Os casos 16, 17 e 18 são análogos ao 13, 14 e 15 anteriores, com a diferença de que neste caso $\sigma_{11} \neq \sigma_{22}$. Ao respeito destes casos podemos dizer que:

⇒ Os coeficientes de Y no caso 16 e X no caso 18, Tabela 6.1.34, são maiores em magnitude pelo que podemos dizer que o efeito destas variáveis é maior, na separação dos grupos. No caso 17 a variável Y têm coeficientes maiores. Embora o dito anteriormente é verdade, as diferenças entre os coeficientes não são grandes.

Tabela 6.1.34: Coeficientes (Casos 16-17-18)

Variação	Variáveis	CASO 16(≠,-,+)		CASO 17(≠,-,0)		CASO 18(≠,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	2.4745	0.8039	5.0274	1.3143	2.7814	8.8476
	Y	3.5163	-0.5278	5.1510	-1.0867	-1.4174	8.8660
Dentro	X	2.1288	0.6916	2.1633	0.5655	0.6713	2.1353
	Y	2.2136	-0.3322	2.1879	-0.4615	-0.3533	2.2103
Bruto	X	1.8213	0.5917	1.8507	0.4838	0.5743	1.8268
	Y	2.6728	-0.4012	2.6413	-0.5572	-0.4266	2.6688

⇒ Os Círculos das Correlações, na Tabela 6.1.36, com a estrutura de correlação *Entre*, deixam ver como nos casos 16 e 18 as variáveis tem uma efeito importante na separação dos grupos, com a diferença de que no caso 18 X e Y estão em contraste. O caso 17, com suas baixas correlações *Entre* com $\mathbf{can}_1$ e altas com $\mathbf{can}_2$, indica que a primeira canônica não mostra direção preferencial para ver a separação dos grupos.

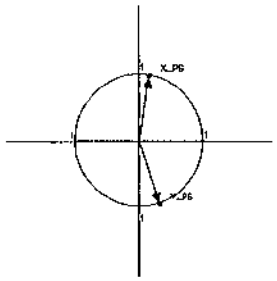
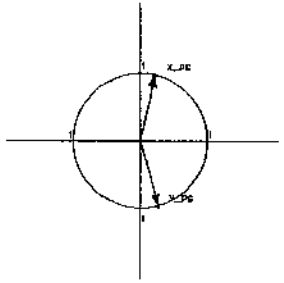
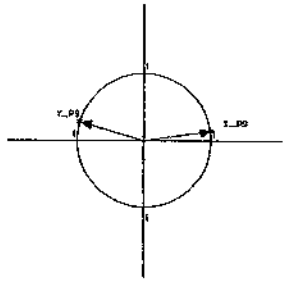
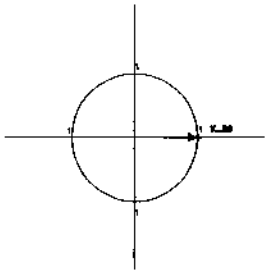
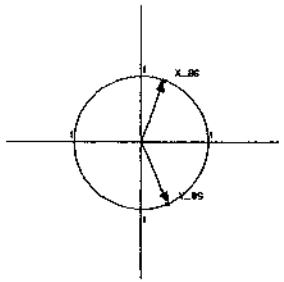
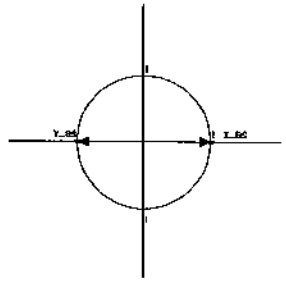
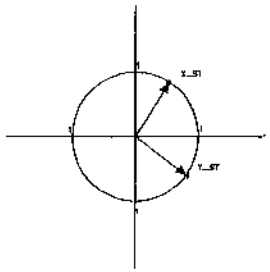
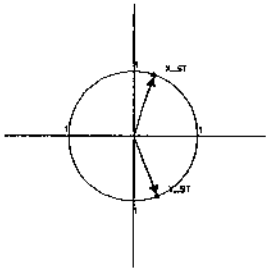
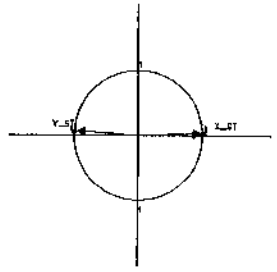
⇒ Quanto a *Dentro*, quando coincide o sentido da correlação *Dentro* e *Entre* as variáveis acompanham a direção de $\mathbf{can}_1$, como podemos ver no caso 18. Quando estas direções são discordantes, casos 16 e 17, a Dentro não acompanha. Os valores destas correlações podem ser vistos na Tabela 6.1.35.

Tabela 6.1.35: Correlações entre originais e discriminantes (Casos 16-17-18)

Estrutura de Correlação	Variáveis	CASO 16(≠,-,+)		CASO 17(≠,-,0)		CASO 18(≠,-,-)	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X	0.5259	0.8505	0.3049	0.9523	0.9992	0.0381
	Y	0.8011	-0.5985	0.3688	-0.9295	-0.9971	0.0747
Entre	X	0.9999	0.0083	0.3231	0.9463	0.9999	0.0013
	Y	0.9999	-0.0038	0.3897	-0.9209	-0.9999	0.0026
Dentro	X	0.1484	0.9889	0.2064	0.9784	0.9874	0.1578
	Y	0.3089	-0.9510	0.2529	-0.9674	-0.9539	0.2999

⇒ A estrutura de correlação *Dentro*, por outro lado, mostra como somente o caso 18 acompanha a orientação das médias, sendo que nos outros casos X e Y têm baixíssimas correlações com **can1**.

Tabela 6.1.36: CÍRCULOS DAS CORRELAÇÕES para os casos 16-17-18

CASO 16(≠,-,+)	CASO 17(≠,-,0)	CASO 18(≠,-,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

6.1.3 Para comparar a influência da variância das variáveis originais, mantendo fixas as correlações, "DENTRO" e "ENTRE", entre as variáveis originais[3]

O objetivo desta seção é comparar a influencia de variâncias iguais ou diferentes nos coeficientes discriminantes de Fisher e nas correlações entre variáveis originais e discriminantes. Cada par de casos têm a mesma estrutura de correlação "DENTRO" e "ENTRE" os grupos para as variáveis X e Y, a única coisa que os diferença é a variância, em um caso $\sigma_{11} = \sigma_{22}$ e em outro $\sigma_{11} \neq \sigma_{22}$.

- **Casos 1 e 4**

Nas Tabelas 6.1.37 e 6.1.38 podem ser vistos os coeficientes discriminantes de Fisher e as correlações entre originais e canônicas, respectivamente.

Tabela 6.1.37: Coeficientes (Casos 1 e 4)

Variação	Variáveis	CASO 1 (=,+,+)[4]		CASO 4 (≠,+,+)	
		Can1	Can2	Can1	Can2
Total	X	2.1792	-8.9398	2.9143	-8.8863
	Y	2.0153	8.9420	1.2816	8.9082
Dentro	X	0.5326	-2.1849	0.7022	-2.1412
	Y	0.4944	2.1939	0.3209	2.2305
Bruto	X	0.5418	-2.2226	0.6007	-1.8319
	Y	0.4989	2.2139	0.3851	2.6768

Os coeficientes, por um lado, mostram magnitude maior para a variável X, neste caso de variâncias diferentes é esta variável a que trouxe maior variância das originais.

3 Isto pode já ter sido observado nas seções 6.1.1 e 6.1.2, coloca-se só para detalhar e esclarecer um pouco mais este aspecto da influencia da mudança da variância.

4 Os sinais entre parêntesis servem para dar uma indicação rápida de se estamos trabalhando com variâncias iguais ou diferentes e de quais as correlações dentro e entre que se correspondem com cada caso. Exemplo: (=,+,0), (≠,+,-) correspondem-se com: variâncias iguais, dentro +0,9 e entre zero; variâncias diferentes, dentro +0,9 e entre -0,9, respectivamente.

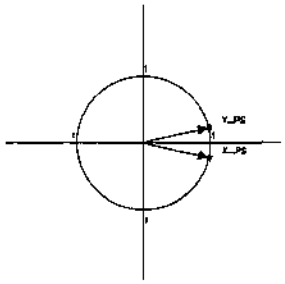
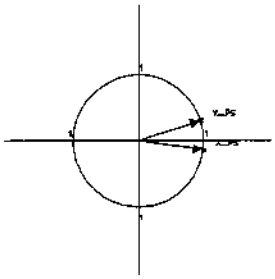
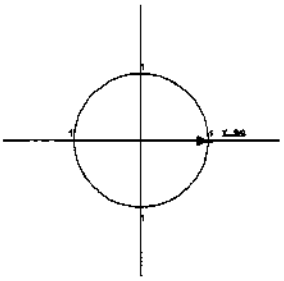
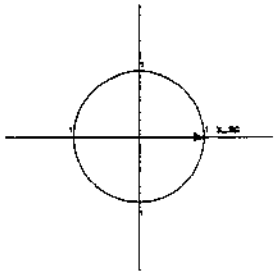
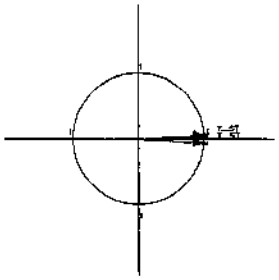
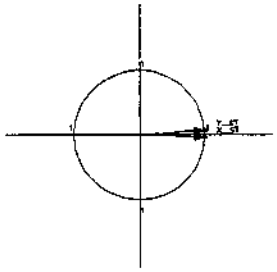
Tabela 6.1.38: Correlações entre originais e discriminantes (casos 1 e 4)

Estrutura de Correlação	Variáveis	CASO 1(=,+,+)		CASO 4(≠,+,+)	
		Can1	Can2	Can1	Can2
Total	X	0.9985	-0.0537	0.9994	-0.0343
	Y	0.9983	0.0581	0.9969	0.0780
Entre	X	0.9999	-0.0017	0.9999	-0.0011
	Y	0.9999	0.0019	0.9999	0.0026
Dentro	X	0.9755	-0.2198	0.9898	-0.1424
	Y	0.9715	0.2368	0.9502	0.3116

As correlações não apresentam comportamentos diferenciados entre um caso ou outro, isto pode ser corroborado nos Círculos das Correlações neste caso, Tabela 6.1.39, ou simplesmente olhando nos coeficientes.

Ou seja, quando temos estruturas de correlação Dentro e Entre iguais e altas positivas a mudança de variâncias não produz resultados diferentes.

Tabela 6.1.39: CÍRCULOS DAS CORRELAÇÕES para os casos 1 e 4.

CASO 1 (=,+,+)	CASO 4 (≠,+,+)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

- **Casos 2 e 5**

Os casos 2 e 5 têm em comum correlação alta positiva dentro dos grupos e nula entre os grupos, diferenciando-se nas variâncias das variáveis X e Y. Quanto a estes casos podemos dizer que os valores dos coeficientes discriminantes de Fisher são aproximadamente iguais para ambos casos.

Tabela 6.1.40: Coeficientes (Casos 2 e 5)

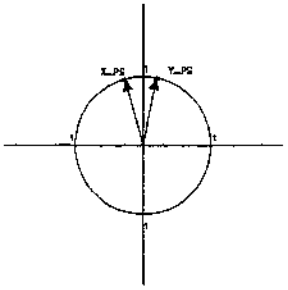
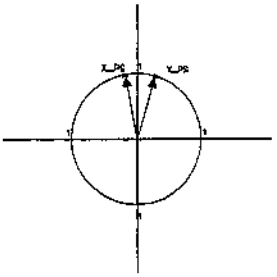
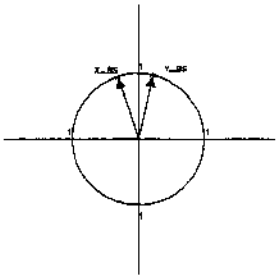
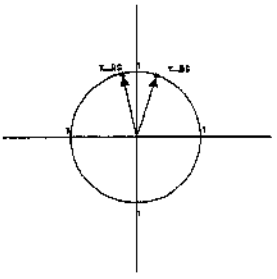
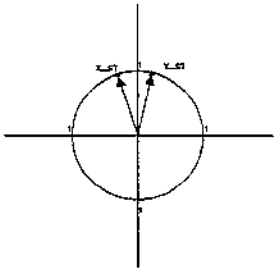
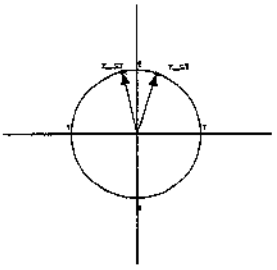
Variação	Variáveis	CASO 2 (=,+,0)		CASO 5(≠,+,0)	
		Can1	Can2	Can1	Can2
Total	X	-5.0445	0.9347	-4.9358	1.3538
	Y	4.9070	1.3980	5.0454	0.9833
Dentro	X	-2.2113	0.4097	-2.1714	0.5955
	Y	2.1628	0.6162	2.2100	0.4307
Bruto	X	-2.2494	0.4168	-1.8576	0.5095
	Y	2.1826	0.6218	2.6523	0.5169

Por outro lado as correlações comportam-se praticamente iguais para ambos casos, como podemos ver nos Círculos das Correlações na Tabela 6.1.42.

Tabela 6.1.41: Correlações entre originais e discriminantes (casos 2 e 5)

Estrutura de Correlação	Variáveis	CASO 2(=,+,0)		CASO 5(≠,+,0)	
		Can1	Can2	Can1	Can2
Total	X	-0.3052	0.9522	-0.2139	0.9768
	Y	0.2040	0.9789	0.2945	0.9556
Entre	X	-0.3121	0.9500	-0.2190	0.9757
	Y	0.2089	0.9779	0.3012	0.9535
Dentro	X	-0.2740	0.9617	-0.1913	0.9815
	Y	0.1822	0.9832	0.2645	0.9643

Tabela 6.1.42: CÍRCULOS DAS CORRELAÇÕES para os casos 2 e 5.

CASO 2 (=,+,0)	CASO 5(≠,+,0)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

- **Casos 3 e 6**

Estes casos têm em comum correlação "DENTRO" dos grupos alta positiva e "ENTRE" alta negativa. Observando os coeficientes na Tabela 6.1.43 e as correlações *Entre* originais e discriminantes na Tabela 6.1.45, com suas representações gráficas nos Círculos das Correlações da Tabela 6.1.44, podemos chegar às seguintes afirmações:

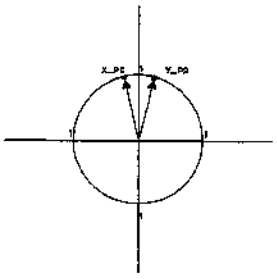
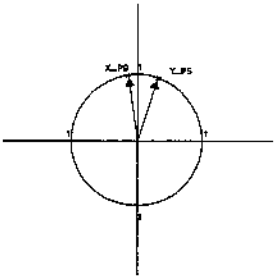
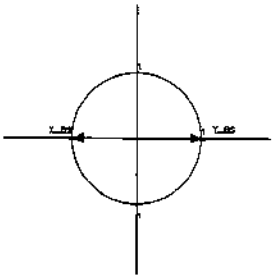
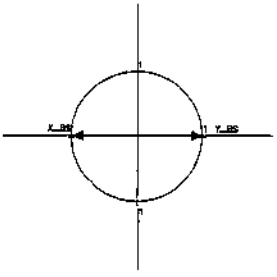
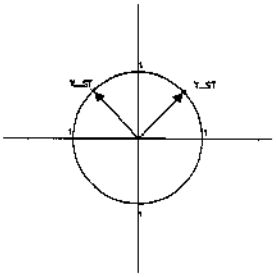
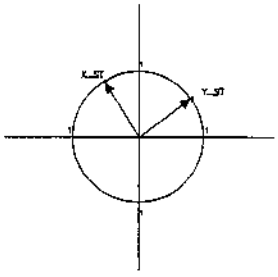
As duas variáveis têm a mesma importância no caso 3, visto pelos coeficientes de magnitude parecidas. No caso 6 os coeficientes parecem privilegiar à Y. Também os coeficientes de X e Y mostram contraste nos dois casos.

Tabela 6.1.43: Coeficientes (Casos 3 e 6)

Variação	Variáveis	CASO 3 (=,+,-)		CASO 6(≠,+,-)	
		Can1	Can2	Can1	Can2
Total	X	-2.9520	0.6987	-2.4846	0.8038
	Y	2.9780	0.6921	3.5380	0.5234
Dentro	X	-2.1885	0.5179	-2.1440	0.6936
	Y	2.1905	0.5091	2.2291	0.3298
Bruto	X	-2.2262	0.5269	-1.8342	0.5934
	Y	2.2105	0.5138	2.6753	0.3958

Olhando os Círculos das Correlações vemos que as duas variáveis, influenciam fortemente à separação dos grupos, mostrando o contraste visto através dos sinais dos coeficientes. A possível maior importância de Y, pelo comportamento visto nos coeficientes, no caso 6 não parece ter um efeito real.

Tabela 6.1.44: CÍRCULOS DAS CORRELAÇÕES para os casos 3 e 6.

CASO 3 (=,+,-)	CASO 6(≠,+,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

Podemos dizer então que a mudança na variância, neste caso, não produz influências na separação dos grupos, mantendo-se praticamente o mesmo comportamento do caso 3 para o caso 6.

Tabela 6.1.45: Correlações entre originais e discriminantes (casos 3 e 6)

Estrutura de Correlação	Variáveis	CASO 3(=,+,-)		CASO 6(≠,+,-)	
		Can1	Can2	Can1	Can2
Total	X	-0.6920	0.7218	-0.5213	0.8533
	Y	0.6985	0.7155	0.8005	0.5992
Entre	X	-0.9999	0.0087	-0.9999	0.0086
	Y	0.9999	0.0086	0.9999	0.0039
Dentro	X	-0.2263	0.9740	-0.1463	0.9892
	Y	0.2303	0.9731	0.3078	0.9514

- **Casos 7 e 10**

Os casos 7 e 10 correspondem-se a correlação "DENTRO" dos grupos zero e "ENTRE" alta positiva, com variâncias entre X e Y iguais e diferentes, respectivamente.

Tabela 6.1.46: Coeficientes (Casos 7 e 10)

Variação	Variáveis	CASO 7(=,0,+)		CASO 10(≠,0,+)	
		Can1	Can2	Can1	Can2
Total	X	2.1618	-2.1486	4.1916	-0.0374
	Y	2.1049	2.1519	0.0009	1.0009
Dentro	X	0.7092	-0.7049	0.9999	-0.0089
	Y	0.6992	0.7148	0.0009	1.0000
Bruto	X	0.7215	-0.7171	0.8555	-0.0076
	Y	0.6933	0.7087	0.0010	1.1791

Respeito a estes casos podemos fazer as seguintes observações:

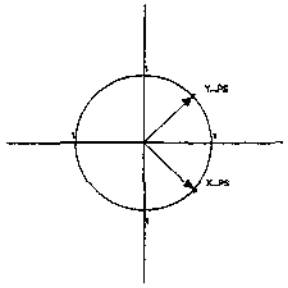
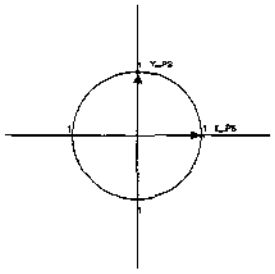
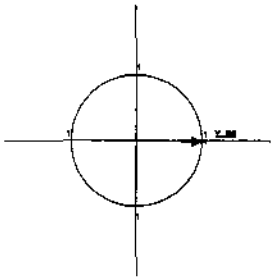
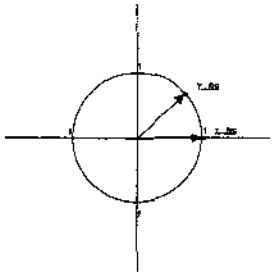
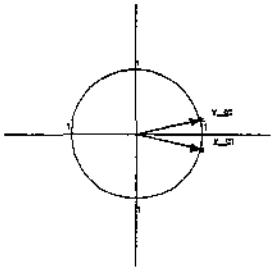
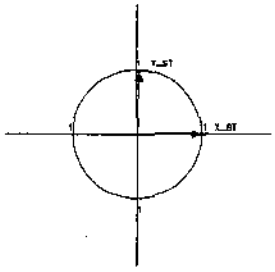
Em primeiro lugar, os coeficientes Discriminantes, Tabela 6.1.46, mostram grande diferenciação entre um caso e outro. No caso 10 o coeficiente para X é muito maior que o do Y. No caso 7 os coeficientes para as variáveis X e Y não se diferenciam entre si.

Tabela 6.1.47: Correlações entre originais e discriminantes (casos 7 e 10)

Estrutura de Correlação	Variáveis	CASO 7(=,0,+)		CASO 10(≠,0,+)	
		Can1	Can2	Can1	Can2
Total	X	0.9733	-0.2294	1.0000	-0.0002
	Y	0.9718	0.2356	0.0373	0.9993
Entre	X	0.9999	-0.0058	1.0000	-0.0000
	Y	0.9999	0.0059	0.7311	0.6822
Dentro	X	0.7148	-0.6992	1.0000	-0.0009
	Y	0.7049	0.7092	0.0089	0.9999

Se vamos nas correlações, Tabela 6.1.47, e olhamos nos Círculos, na Tabela 6.1.48, vemos que as estruturas de correlação *Entre, Dentro* e *Total,* sofrem um impacto significativo quando mudamos a variância no caso de ausência de correlação entre os grupos.

Tabela 6.1.48: CÍRCULOS DAS CORRELAÇÕES para os casos 7 e 10.

CASO 7(=,0,+)	CASO 10(≠,0,+)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

- **Casos 8 e 11**

Nestes casos temos ausência de correlação tanto "DENTRO" quanto "ENTRE" os grupos, a diferença entre eles é que no caso 8 X e Y têm variâncias iguais e no caso 11 diferentes.

Tabela 6.1.49: Coeficientes (Casos 8 e 11)

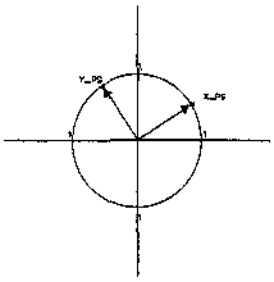
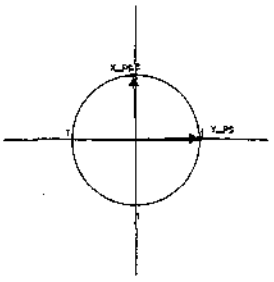
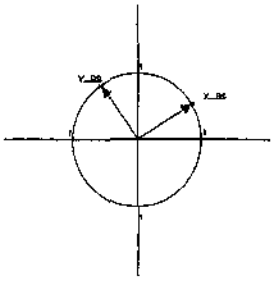
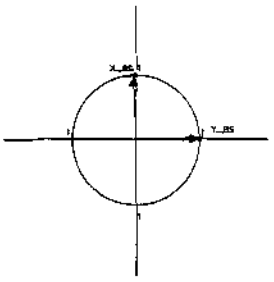
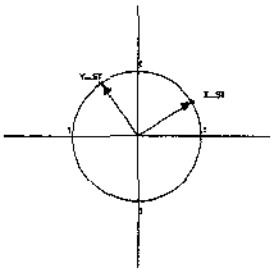
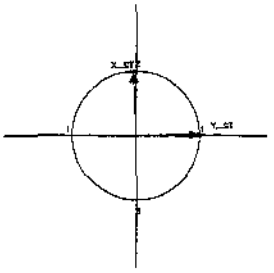
Variação	Variáveis	CASO 8(=,0,0)		CASO 11(≠,0,0)	
		Can1	Can2	Can1	Can2
Total	X	1.8888	1.2488	-0.1017	2.2652
	Y	-1.2523	1.8616	2.5402	0.0936
Dentro	X	0.83418	0.5515	-0.0448	0.9990
	Y	-0.5581	0.8297	0.9993	0.0368
Bruto	X	0.8485	0.5610	-0.0383	0.8546
	Y	-0.5534	0.8226	1.1783	0.0434

Para estes casos podemos fazer as seguintes observações:

Em ambos casos os coeficientes, Tabela 6.1.49, estão em contraste. Quanto à magnitude para os coeficientes podemos dizer que, no caso 8, os coeficientes para X são maiores, ao contrário do caso 10 onde são os de Y.

As correlações, Tabela 6.1.51, mostram também, igual que os coeficientes, o contraste existente entre as variáveis X e Y.

Tabela 6.1.50: CÍRCULOS DAS CORRELAÇÕES para os casos 8 e 11.

CASO 8(=,0,0)	CASO 11(≠,0,0)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

Nestes casos uma coisa curiosa é que como a estrutura de correlação "ENTRE" e "DENTRO" são iguais a zero, os Círculos das Correlações estão repetindo sempre o mesmo padrão. Desta forma para cada caso os três gráficos são iguais.

Tabela 6.1.51: Correlações entre originais e discriminantes (casos 8 e 11)

Estrutura de Correlação	Variáveis	CASO 8(=,0,0)		CASO 11(≠,0,0)	
		Can1	Can2	Can1	Can2
Total	X	0.8358	0.5489	-0.0413	0.9991
	Y	-0.5607	0.8280	0.9992	0.0400
Entre	X	0.8372	0.5467	-0.0423	0.9991
	Y	-0.5629	0.8264	0.9992	0.0390
Dentro	X	0.8297	0.5581	-0.0368	0.9993
	Y	-0.5515	0.8341	0.9989	0.0448

Nestes casos a mudança na variância das variáveis originais provoca diferenças marcadas nas respectivas estruturas de correlação.

- **Casos 9 e 12**

Os casos 9 e 12 têm em comum correlação entre X e Y zero "DENTRO" dos grupos e alta negativa "ENTRE" os grupos. A diferença é que no caso 9 estas variáveis têm variâncias iguais e no caso 12 diferentes.

Tabela 6.1.52: Coeficientes (Casos 9 e 12)

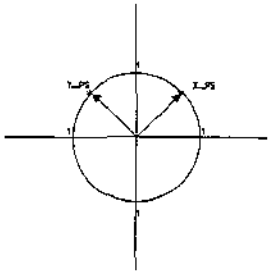
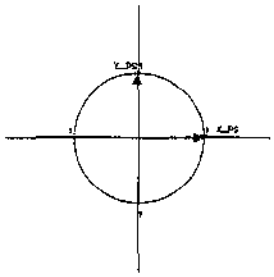
Variação	Variáveis	CASO 9(=,+,-)		CASO 12(≠,0,-)	
		Can1	Can2	Can1	Can2
Total	X	2.1709	2.1294	4.1850	0.0374
	Y	-2.1250	2.1321	-0.0169	1.0007
Dentro	X	0.7139	0.7002	0.9999	0.0089
	Y	-0.7059	0.7082	-0.0169	0.9998
Bruto	X	0.7262	0.7123	0.8555	0.0076
	Y	-0.6999	0.7022	-0.0199	1.1789

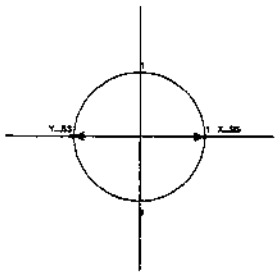
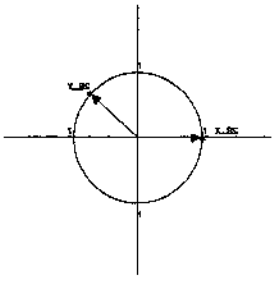
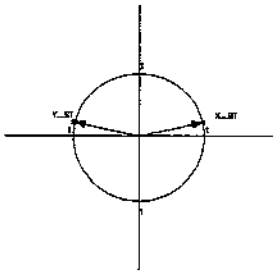
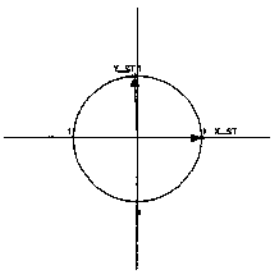
Nestes casos que tem em comum com os quatro vistos anteriormente, ausência de correlação DENTRO dos grupos, também podemos observar marcadas diferenças tanto entre os coeficientes quanto entre as estruturas de correlação *Entre*, *Dentro* e *Total*.

Tabela 6.1.53: Correlações entre originais e discriminantes (casos 9 e 12)

Estrutura de Correlação	Variáveis	CASO 9(=,0,-)		CASO 12(≠,0,-)	
		Can1	Can2	Can1	Can2
Total	X	0.9726	0.2321	0.9999	0.0040
	Y	-0.9714	0.2371	-0.0374	0.9993
Entre	X	0.9999	0.006087	1.0000	0.0001
	Y	-0.9999	0.0062	-0.7304	0.6829
Dentro	X	0.7082	0.7059	0.9998	0.0169
	Y	-0.7002	0.713890	-0.0089	0.9999

Tabela 6.1.54: CÍRCULOS DAS CORRELAÇÕES para os casos 9 e 12.

CASO 9(=,0,-)	CASO 12(≠,0,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	

CASO 9(=,0,-)	CASO 12(≠,0,-)
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

Nos casos 13 e 16, 14 e 17 e 15 e 18 poderão observar-se aproximadamente os mesmos padrões de comportamento, tanto para os coeficientes quanto para as correlações entre originais e discriminantes, com variâncias iguais ou diferentes.

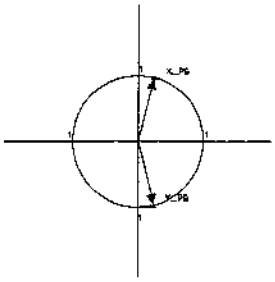
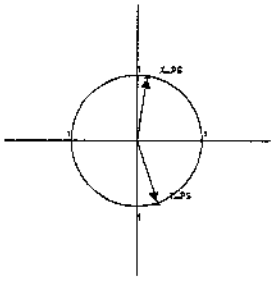
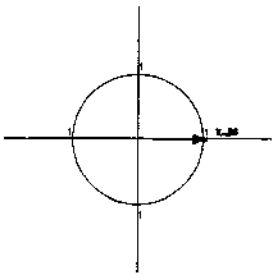
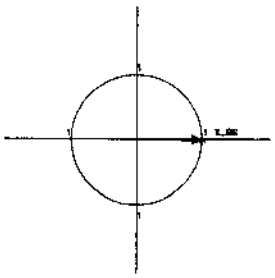
- **Casos 13 e 16**

Os casos 13 e 16 têm em comum correlação entre X e Y "DENTRO" dos grupos alta negativa e "ENTRE" os grupos alta positiva. A diferença é nas variâncias de X e Y pois no caso 13 são iguais e no 16 diferentes.

Tabela 6.1.55: Coeficientes (Casos 13 e 16)

Variação	Variáveis	CASO 13(=,-,+)		CASO 16(≠,-,+)	
		Can1	Can2	Can1	Can2
Total	X	2.9412	0.6985	2.4745	0.8039
	Y	2.9542	-0.6953	3.5163	-0.5278
Dentro	X	2.1724	0.5159	2.1288	0.6916
	Y	2.1734	-0.5115	2.2136	-0.3322
Bruto	X	2.2098	0.5248	1.8213	0.5917
	Y	2.2060	-0.5192	2.6728	-0.4012

Tabela 6.1.56: CÍRCULOS DAS CORRELAÇÕES para os casos 13 e 16.

CASO 13(=,-,+)	CASO 16(≠,-,+)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	

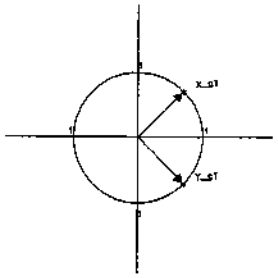
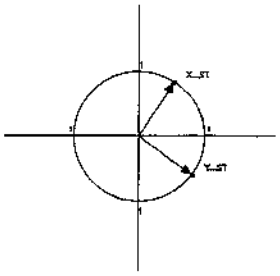
CASO 13(=,-,+)	CASO 16(≠,-,+)
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

Tabela 6.1.57: Correlações entre originais e discriminantes (casos 13 e 16)

Estrutura de Correlação	Variáveis	CASO 13(=,-,+)		CASO 16(≠,-,+)	
		Can1	Can2	Can1	Can2
Total	X	0.6952	0.7187	0.5259	0.8505
	Y	0.6985	-0.7155	0.8011	-0.5985
Entre	X	0.9999	0.0082	0.9999	0.0083
	Y	0.9999	-0.0081	0.9999	-0.0038
Dentro	X	0.2291	0.9733	0.1484	0.9889
	Y	0.2310	-0.9729	0.3089	-0.9510

- **Casos 14 e 17**

Estes casos têm em comum correlação, entre as variáveis originais, "DENTRO" alta negativa e "ENTRE" zero. O caso 14 com $\sigma_{11} = \sigma_{22}$ e o caso 17 com $\sigma_{11} \neq \sigma_{22}$.

Tabela 6.1.58: Coeficientes (Casos 14 e 17)

Variação	Variáveis	CASO 14(=,-,0)		CASO 17(≠,-,0)	
		Can1	Can2	Can1	Can2
Total	X	4.9789	1.0693	5.0274	1.3143
	Y	4.9192	-1.2707	5.1510	-1.0867

Dentro	X	2.1830	0.4688	2.1633	0.5655
	Y	2.1619	-0.5584	2.1879	-0.4615
Bruto	X	2.2207	0.4769	1.8507	0.4838
	Y	2.1942	-0.5668	2.6413	-0.5572

Tabela 6.1.59: CÍRCULOS DAS CORRELAÇÕES para os casos 14 e 17.

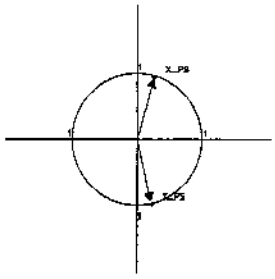
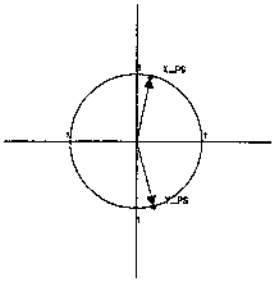
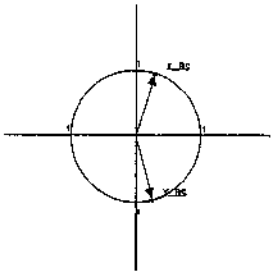
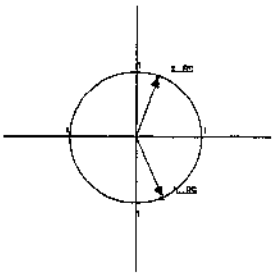
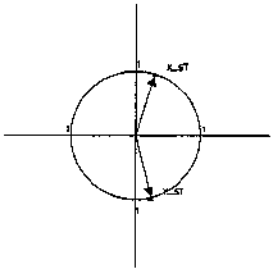
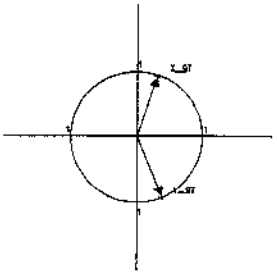
CASO 14(=,-,0)	CASO 17(≠,-,0)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

Tabela 6.1.60: Correlações entre originais e discriminantes (casos 14 e 17)

Estrutura de Correlação	Variáveis	CASO 14(=,-,0)		CASO 17(≠,-,0)	
		Can1	Can2	Can1	Can2
Total	X	0.2780	0.9605	0.3049	0.9523
	Y	0.2339	-0.9722	0.3688	-0.9295
Entre	X	0.2842	0.9587	0.3231	0.9463
	Y	0.2393	-0.9709	0.3897	-0.9209
Dentro	X	0.2501	0.9682	0.2064	0.9784
	Y	0.2099	-0.9777	0.2529	-0.9674

- **Casos 15 e 18**

Os casos 15 e 18 têm em comum a correlação "DENTRO" e "ENTRE" para X e Y, em ambos casos alta negativa.

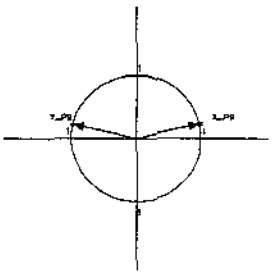
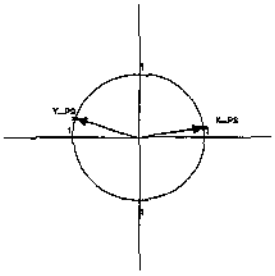
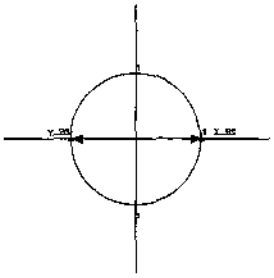
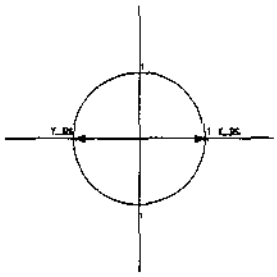
Tabela 6.1.61: Coeficientes (Casos 15 e 18)

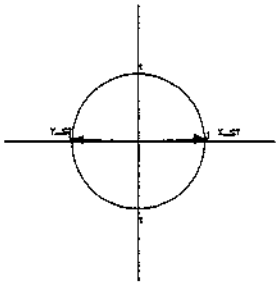
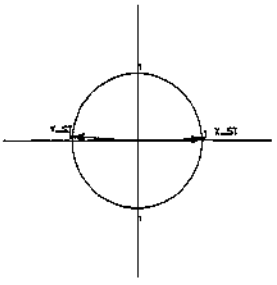
Variação	Variáveis	CASO 15(=,-,-)		CASO 18(≠,-,-)	
		Can1	Can2	Can1	Can2
Total	X	2.0588	8.8842	2.7814	8.8476
	Y	-2.1422	8.8831	-1.4174	8.8660
Dentro	X	0.5040	2.1752	0.6713	2.1353
	Y	-0.5234	2.1706	-0.3533	2.2103
Bruto	X	0.5127	2.2127	0.5743	1.8268
	Y	-0.5313	2.2031	-0.4266	2.6688

Tabela 6.1.62: Correlações entre originais e discriminantes (casos 15 e 18)

Estrutura de Correlação	Variáveis	CASO 15(=,-,-)		CASO 18(≠,-,-)	
		Can1	Can2	Can1	Can2
Total	*X*	0.9983	0.0574	0.9992	0.0381
	Y	-0.9984	0.0551	-0.9971	0.0747
Entre	*X*	0.9999	0.0021	0.9999	0.0013
	Y	-0.9999	0.0020	-0.9999	0.0026
Dentro	*X*	0.9721	0.2344	0.9874	0.1578
	Y	-0.9741	0.2257	-0.9539	0.2999

Tabela 6.1.63: CÍRCULOS DAS CORRELAÇÕES para os casos 15 e 18.

CASO 15(=,-,-)	CASO 18(≠,-,-)
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS	
	
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS	
	

CASO 15(=,-,-)	CASO 18(≠,-,-)
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL	
	

6.2 Comportamento dos coeficientes da Função Linear Discriminante de Fisher e das correlações entre as variáveis originais e discriminantes no caso de três variáveis.

Aqui começa análise dos resultados para os dados simulados com três variáveis. Lembremos que os casos de três variáveis foram trabalhados para variâncias iguais, isto significa que em todos os casos a matriz de variâncias e covariâncias coincide com a matriz de correlações.

6.2.1 Para comparar a influência da correlação "DENTRO" dos grupos, entre as variáveis originais e as discriminantes.

Nestes casos mantemos fixa a correlação "ENTRE" os grupos, procurando ver se a variação da correlação "DENTRO" influencia no cálculo dos coeficientes da Função Linear Discriminante de Fisher e nas correlações, e analisar o que isto têm a ver com o poder discriminativo das variáveis discriminantes.

- **Casos 1, 4 e 7**

⇒ Estes casos têm em comum uma matriz de correlação esférica "ENTRE" os grupos com valor de $+0{,}9$. O que varia são as matrizes de correlação "DENTRO" dos grupos: para o caso 1 ela coincide com a matriz de correlação "ENTRE", no caso 4 ela é a identidade e no caso 7 ela têm os valores $\rho_{12}^{D} = +0{,}9$, $\rho_{13}^{D} = -0{,}9$ e $\rho_{23}^{D} = -0{,}9$. Nestes três casos **can1** responde pelo 100% da variância entre grupos, o que é esperado se examinarmos o gráfico dos dados simulados nas páginas 93, 94 e 95.

⇒ A primeira variável tem coeficientes menores em X_3 no caso 1, no qual no plano X_1 - X_2 já está a separação entre grupos e os demais nada acrescentam.

⇒ A semelhança dos coeficientes *Bruto* e *Dentro* reflete a igualdade das variâncias a um.

⇒ O *Total* tem maior magnitude como simples efeito da variabilidade total.

⇒ No caso 4 a contribuição das variáveis é bastante equilibrada nos três coeficientes, não há uma projeção privilegiada como no caso 1, isto se explica pela ausência de correlação DENTRO.

⇒ A magnitude maior de X_3 no caso 7, ilustrada nos gráficos dos dados simulados, mencionados anteriormente e que são as Figuras 5.2.2 até 5.2.4, deve-se a maior visibilidade da separação dos grupos estar nessa direção.

⇒ As correlações nos casos 1-4-7 e o exame dos Círculos de Correlações, Tabela 6.2.3, mostram que as correlações *Dentro* caem a medida que o alongamento da nuvem dentro dos grupos difere do alongamento ao longo dos grupos. As baixas correlações dentro, no caso 7, correspondem à orientação dentro "oposta" a entre.

⇒ As correlações entre são triviais, as totais exibem, caso 7, ou não, caso 1, as diferenças de orientação apontadas.

Tabela 6.2.1: Coeficientes (Casos 1-4-7)

Variação	Variáveis	CASO 1		CASO 4		CASO 7	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	1.8505	-9.3573	1.4884	-1.3805	1.5550	-2.5326
	X_2	1.6570	8.4417	1.4561	2.0228	1.4490	2.6813
	X_3	0.6930	0.9194	1.3043	-0.6400	2.9391	-0.1398
Dentro	X_1	0.4527	-2.2892	0.5789	-0.5369	1.2738	-2.0746
	X_2	0.4070	2.0734	0.5770	0.8016	1.1887	2.1995
	X_3	0.1743	0.2313	0.5287	-0.2595	2.3966	-0.1140
Bruto	X_1	0.4697	-2.3751	0.6006	-0.5571	1.3216	-2.1525
	X_2	0.4205	2.1422	0.5856	0.8135	1.2281	2.2724
	X_3	0.1756	0.2329	0.5223	-0.2564	2.5212	-0.1199

Tabela 6.2.2: Correlações entre originais e discriminantes (Casos 1-4-7)

Estrutura de Correlação	Variáveis	CASO 1		CASO 4		CASO 7	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	0.9984	-0.0548	0.9510	-0.2142	0.5978	-0.0015
	X_2	0.9981	0.0564	0.9478	0.3142	0.5961	0.3620
	X_3	0.9966	0.0115	0.9437	-0.1064	0.6034	-0.1776
Entre	X_1	1.0000	-0.0000	1.0000	0.0000	1.0000	0.0000
	X_2	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
	X_3	1.0000	0.0000	1.0000	0.0000	1.0000	-0.0000
Dentro	X_1	0.9734	-0.2241	0.6072	-0.5509	0.2046	-0.0019
	X_2	0.9693	0.2297	0.5939	0.7933	0.2037	0.4415
	X_3	0.9448	0.0457	0.5781	0.5781	0.2074	-0.2179

Analisando os coeficientes da Tabela 6.2.1 e as correlações da Tabela 6.2.2, podemos fazer as seguintes afirmações:

⇒ Os coeficientes para o caso 1 indicam a ordem de importância das variáveis como X_1, X_2 e X_3, igualmente no caso 4. Já no caso 7 a ordem de importância das variáveis segundo os coeficientes é X_3, X_1 e X_2.

⇒ Se olhamos nos Círculos das Correlações vemos que a estrutura de correlação Entre os grupos, ver Tabela 6.2.3 b), exibe correlações triviais com **can1**, em todos os casos igual a um e zero com **can2**, consequência óbvia dos dados gerados.

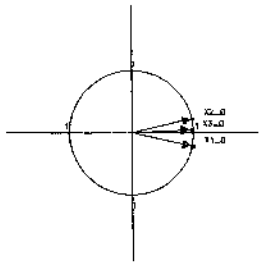
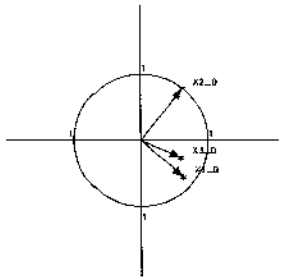
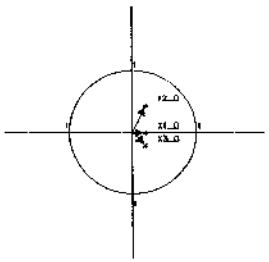
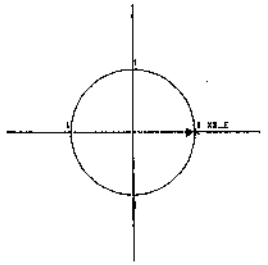
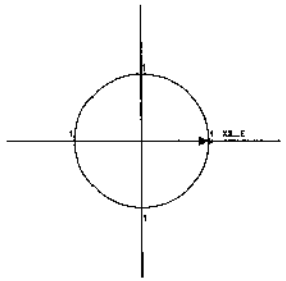
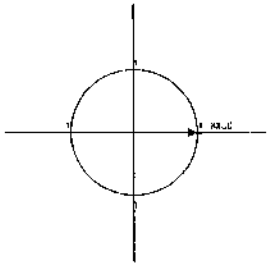
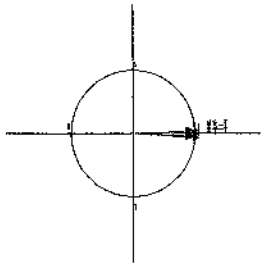
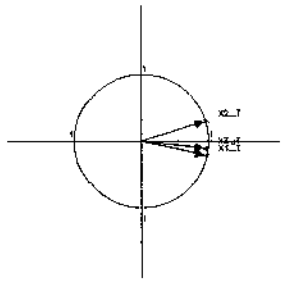
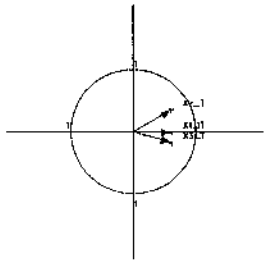
⇒ Além disso, no caso 1, a estrutura de correlação *Dentro* dos grupos acompanha muito bem a *Entre*, com altas correlações das três variáveis com **can1**, i.e., $corr_D(X_1, \mathbf{can}_1) = 0{,}97$, $corr_D(X_2, \mathbf{can}_1) = 0{,}97$ e $corr_D(X_3, \mathbf{can}_1) = 0{,}94$. Agora, este comportamento mudará a medida que muda a matriz de correlações "DENTRO" dos grupos, já nos casos 4 e 7 a estrutura de correlação *Dentro* não acompanham igual como no caso 1.

⇒ No caso 4 a ausência de correlação dentro reflete-se em correlações positivas e baixas com **can1**, devido a forte correlação positiva entre os grupos.

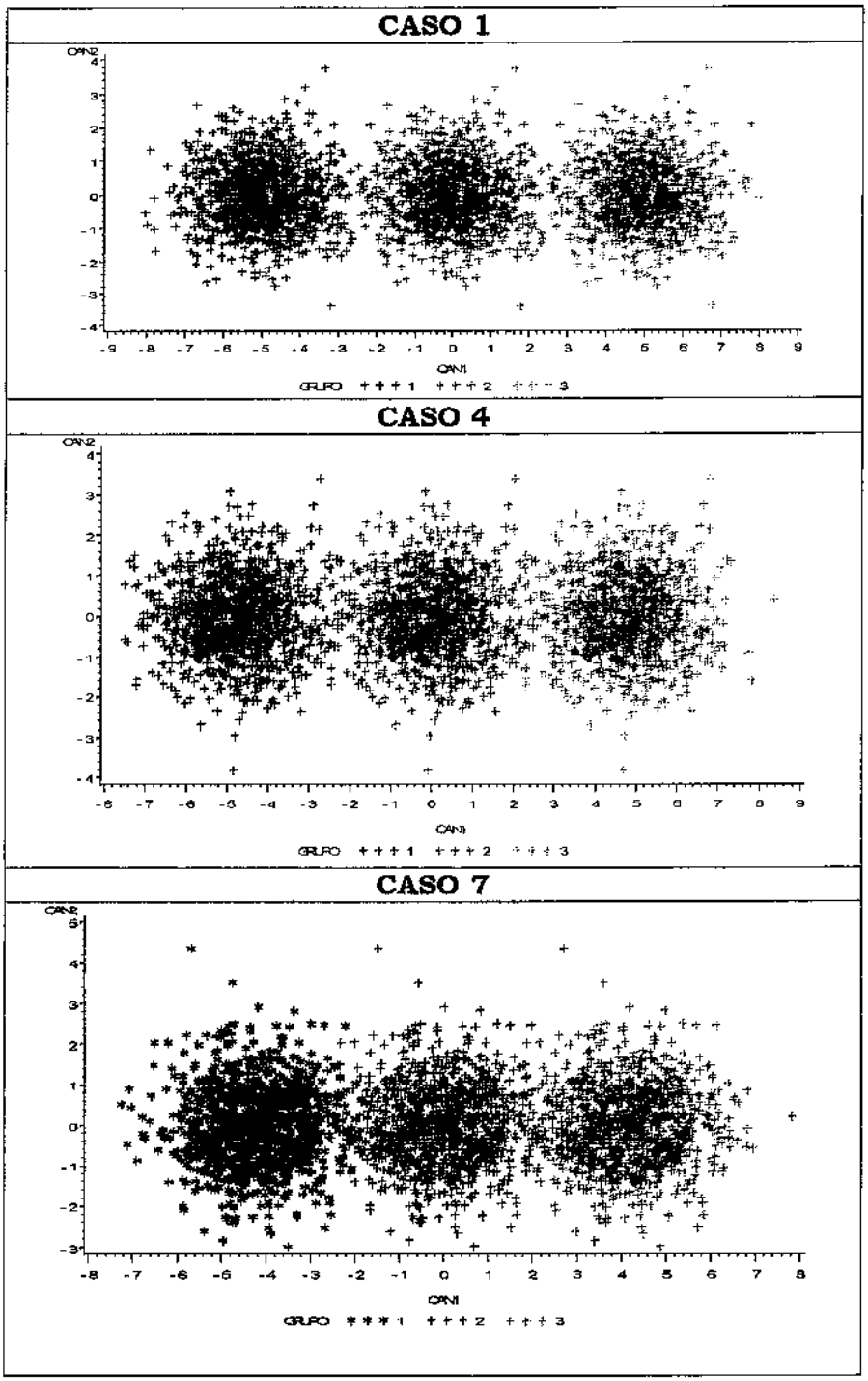
⇒ No caso 7 as correlações aparecem mais baixas devido às direções ortogonais do alongamento dentro e ao longo dos grupos.

⇒ Vistos estes casos podemos dizer que para estas combinações de correlações a *Dentro* nem sempre acompanhará a *Entre*. No caso 1, onde as duas estruturas coincidem, sim fica claro que elas se acompanham sem problemas.

Tabela 6.2.3: CÍRCULOS DAS CORRELAÇÕES para os casos 1-4-7

CASO 1	CASO 4	CASO 7
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

Para ter uma idéia apresentamos os planos discriminantes para os casos 1, 4 e 7, como pode se observar são muito semelhantes.



- **Casos 2, 5 e 8**

A característica destes casos é que têm em comum uma matriz de correlação uniforme "ENTRE" os grupos com valor de 0, ou seja, os três grupos são incorrelatos nestes casos. O que varia são as matrizes de correlação "DENTRO" dos grupos: para o caso 2 é uniforme igual a +0,9, no caso 5 ela coincide com a matriz de correlação "ENTRE", no caso 8 têm os valores $\rho_{12}^{D} = +0{,}9$, $\rho_{13}^{D} = -0{,}9$ e $\rho_{23}^{D} = -0{,}9$.

Tabela 6.2.4: Coeficientes (Casos 2-5-8)

Variação	Variáveis	CASO 2		CASO 5		CASO 8	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	-3.4132	-0.0107	1.9091	1.2383	-4.6217	1.5659
	X_2	-2.4343	0.1804	0.0821	0.8792	-0.0004	0.6363
	X_3	5.3185	2.0922	-1.6175	1.4387	-5.5377	-0.1291
Dentro	X_1	-1.5172	-0.0047	0.77034	0.4996	-2.0319	0.6884
	X_2	-1.0938	0.0810	0.0536	0.5746	-0.0002	0.2860
	X_3	2.3662	0.9308	-0.6797	0.6046	-2.1920	-0.0511
Bruto	X_1	-1.5746	-0.0049	0.7992	0.5184	-2.1081	0.7143
	X_2	-1.1301	0.0837	0.0544	0.5831	-0.0002	0.2955
	X_3	2.3830	0.9374	-0.6715	0.5973	-2.3059	-0.0537

Nestes casos o primeiro autovalor acumula o valor de 56,85, no caso 2, 66,65% no caso 5. Analisando os coeficientes da Tabela 6.2.4 e as correlações da Tabela 6.2.5, podemos fazer as seguintes afirmações:

⇒ A primeira direção canônica exibe contrastes pela isotropia entre grupos (ausência de correlação entre os grupos)

⇒ A magnitude apenas reflete a igualdade de variâncias dentro, pois os coeficientes *Dentro* e *Bruto* apresentam-se semelhantes em cada caso.

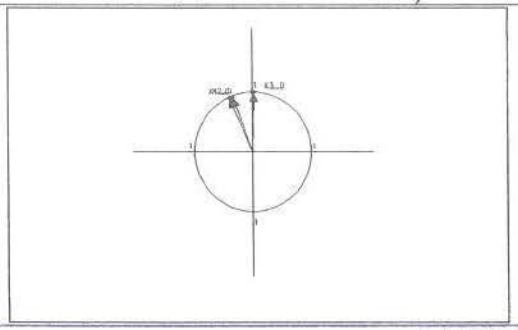
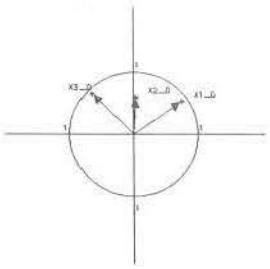
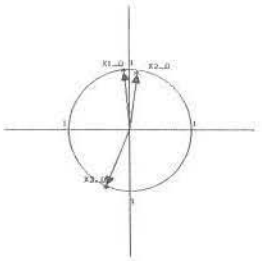
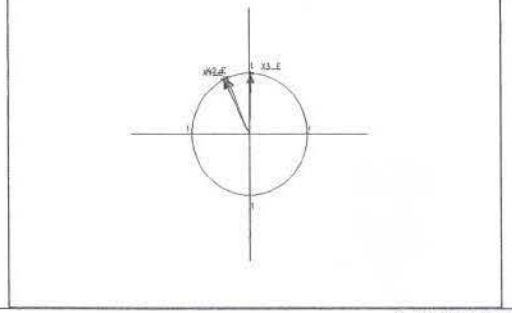
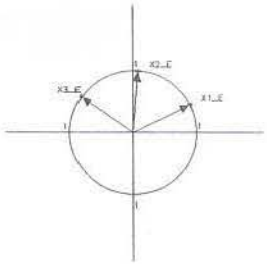
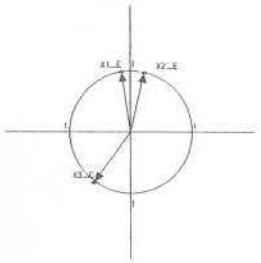
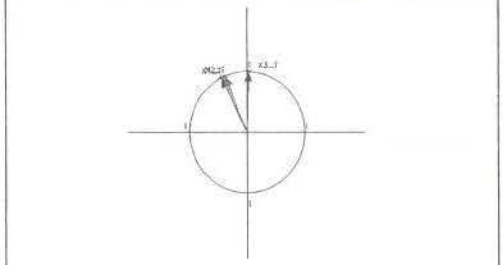
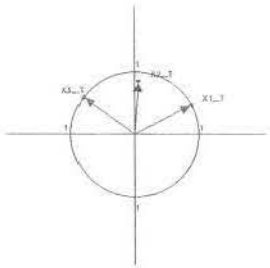
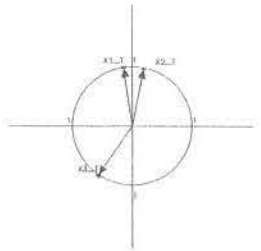
⇒ A manutenção dos sinais dos contrastes nas correlações deve-se, dentro de cada grupo, à ausência de oposição da ENTRE, pela ausência de correlação nestes casos.

⇒ Dos Círculos de Correlação, na correlação *Entre*, nota-se que necessitamos do plano canônico para visualizar discriminação e que a estrutura *Dentro* é similar à *Entre*. Em ambos casos as variáveis tem baixas correlações com $\mathbf{can}_1$, casos 2 e 8, ou não tem preferência por nenhuma das duas direções, caso 5.

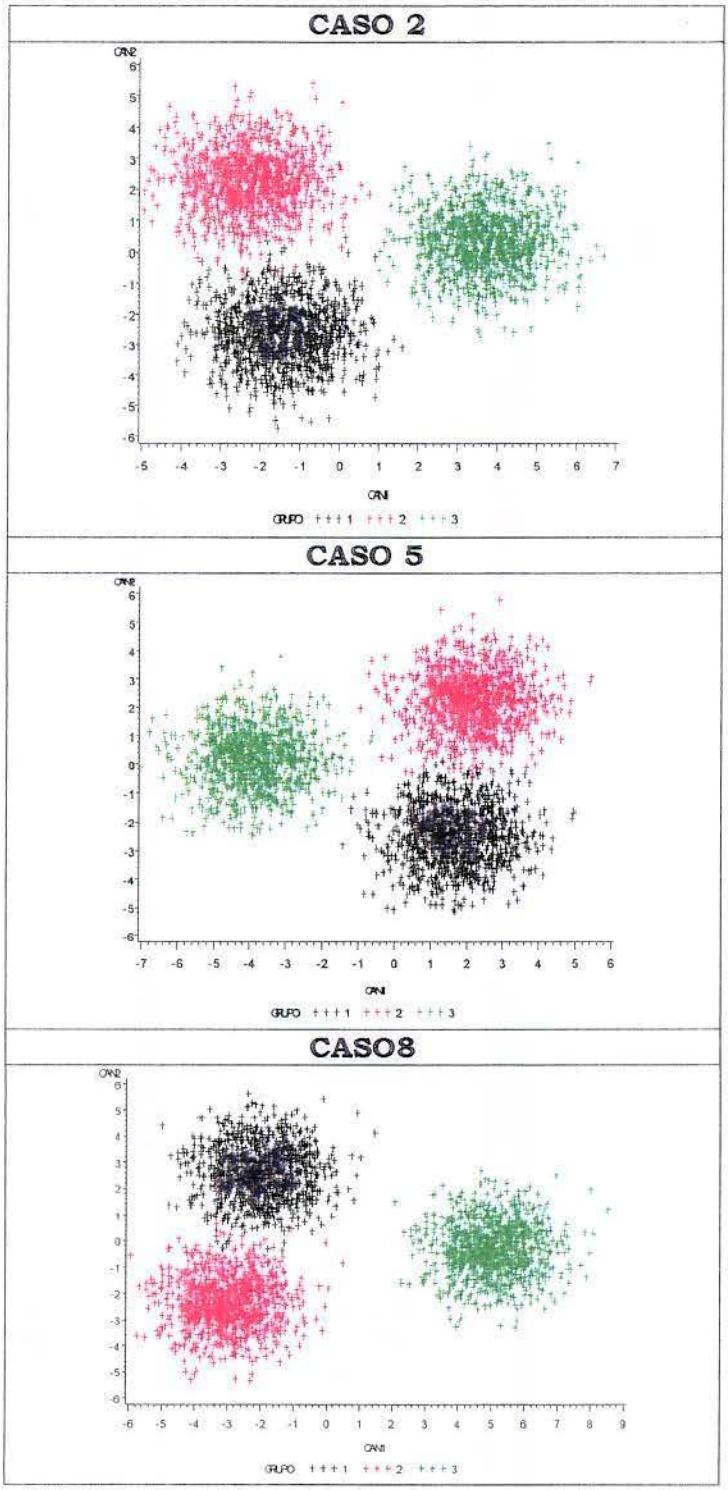
Tabela 6.2.5: Correlações entre originais e discriminantes (Casos 2-5-8)

Estrutura de Correlação	Variáveis	CASO 2		CASO 5		CASO 8	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	-0.4261	0.8998	0.8664	0.4694	-0.1302	0.9902
	X_2	-0.3714	0.9220	0.0609	0.8488	0.2065	0.9695
	X_3	0.0298	0.9995	-0.7829	0.5971	-0.5613	-0.8265
Entre	X_1	-0.4366	0.8996	0.8898	0.4562	-0.1396	0.9902
	X_2	-0.3816	0.9243	0.0756	0.9971	0.2226	0.9748
	X_3	0.0305	0.9995	-0.8108	0.5852	-0.5886	-0.8083
Dentro	X_1	-0.3808	0.9005	0.7339	0.5320	-0.0798	0.9903
	X_2	-0.3283	0.9128	0.0318	0.5939	0.1238	0.9483
	X_3	0.0266	0.9993	-0.6368	0.6497	-0.3822	-0.9180

Tabela 6.2.6: CÍRCULOS DAS CORRELAÇÕES para os casos 2-5-8

CASO 2	CASO 5	CASO 8
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

Os planos discriminantes nestes caso ficam:



- **Casos 3, 6 e 9**

Estes casos têm em comum a matriz de correlações "ENTRE" os grupos, i.e.,. $\rho_{12}^{E} = +0{,}9$, $\rho_{13}^{E} = -0{,}9$ e $\rho_{23}^{E} = -0{,}9$, a matriz de correlações "DENTRO" dos grupos varia da mesma forma que o caso 1, 4 e 7, para três variáveis, tratado anteriormente.

Ao respeito destes três casos podemos fazer os seguintes comentários:

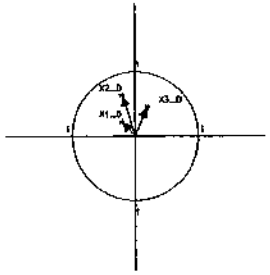
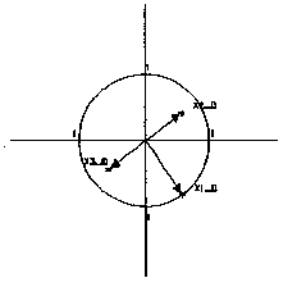
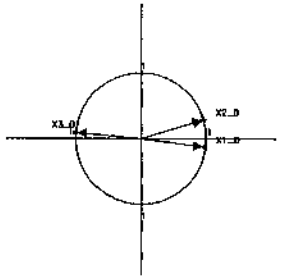
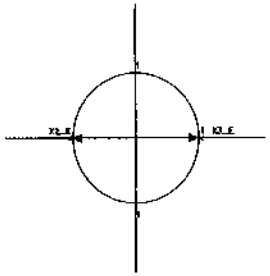
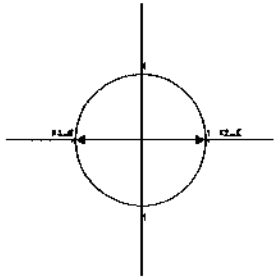
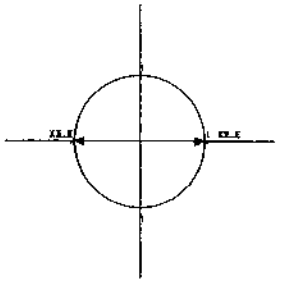
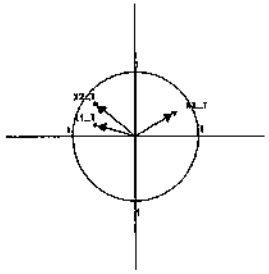
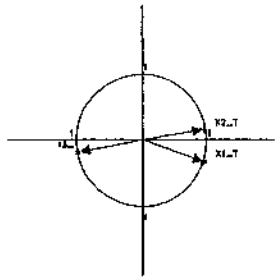
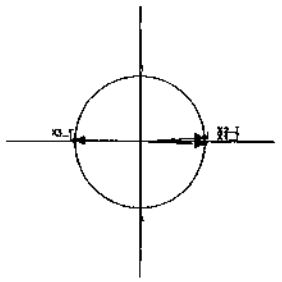
⇒ Observam-se contrastes pela presença de correlações negativas entre os grupos.

⇒ Os Círculos de Correlação ao longo da estrutura entre exibem a importância das variáveis no contraste, transmitindo a estrutura de correlação entre.

Tabela 6.2.7: Coeficientes (Casos 3-6-9)

Variação	Variáveis	CASO 3		CASO 6		CASO 9	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	-1.5620	-2.2937	1.5576	-2.0497	1.2501	-5.6117
	X_2	-1.6576	2.5828	1.4648	1.0859	1.1447	10.1611
	X_3	3.0978	0.2876	-1.4738	-0.9750	-1.8656	4.5430
Dentro	X_1	-1.2505	-1.8362	0.6058	-0.7973	0.3058	-1.3728
	X_2	-1.3290	2.0708	0.5805	0.4303	0.2811	2.4958
	X_3	2.5061	0.2326	-0.5974	-0.3952	-0.4505	1.0970
Bruto	X_1	-1.2974	-1.9051	0.6286	-0.8272	0.3173	-1.4244
	X_2	-1.3731	2.1394	0.5891	0.4367	0.2905	2.5785
	X_3	2.5239	0.2343	-0.5902	-0.3904	-0.4739	1.1541

Tabela 6.2.8: CÍRCULOS DAS CORRELAÇÕES para os casos 3-6-9

CASO 3	CASO 6	CASO 9
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

⇒ A estrutura *Dentro* apresenta baixas correlações no caso 3, pela oposição das direções, passando a valores razoáveis no caso 6 e reproduzindo a concordância das direções simuladas no caso 9.

Tabela 6.2.9: Correlações entre originais e discriminantes (Casos 3-6-9)

Estrutura de Correlação	Variáveis	CASO 3		CASO 6		CASO 9	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	-0.6205	0.1822	0.9479	-0.3182	0.9976	-0.0277
	X_2	-0.6188	0.5083	0.9447	0.1654	0.9973	0.0717
	X_3	0.6087	0.3639	-0.9406	-0.1718	-0.9984	0.0254
Entre	X_1	-1.0000	-0.0000	1.0000	-0.0000	1.0000	-0.0000
	X_2	-1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
	X_3	1.0000	-0.0000	-1.0000	-0.0000	-1.0000	0.0000
Dentro	X_1	-0.1997	0.2277	0.5738	-0.8184	0.9590	-0.1133
	X_2	-0.1988	0.6342	0.5613	0.4177	0.9550	0.2920
	X_3	0.1938	0.4499	-0.5464	-0.4241	-0.9724	0.1053

6.2.2 Para comparar a influência de diferentes matrizes de correlação "ENTRE".

Esta seção têm como objetivos esclarecer se diferentes combinações de correlação entre X_2, X_1 e X_3, mantendo fixa a matriz de correlações "DENTRO", modificam a resposta dos coeficientes da Função Linear Discriminante de Fisher e das correlações entre as variáveis originais e as discriminantes.

- **Casos 1, 2 e 3**

Neste caso mantemos fixa a matriz de correlações "DENTRO" dos grupos, a qual é uma matriz de correlação esférica com valor $+0{,}9$.

Tabela 6.2.10: Coeficientes (Casos 1-2-3)

Variação	Variáveis	CASO 1		CASO 2		CASO 3	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	1.8505	-9.3573	-3.4132	-0.0107	-1.5620	-2.2937
	X_2	1.6570	8.4417	-2.4343	0.1804	-1.6576	2.5828
	X_3	0.6930	0.9194	5.3185	2.0922	3.0978	0.2876
Dentro	X_1	0.4527	-2.2892	-1.5172	-0.0047	-1.2505	-1.8362
	X_2	0.4070	2.0734	-1.0938	0.0810	-1.3290	2.0708
	X_3	0.1743	0.2313	2.3662	0.9308	2.5061	0.2326
Bruto	X_1	0.4697	-2.3751	-1.5746	-0.0049	-1.2974	-1.9051
	X_2	0.4205	2.1422	-1.1301	0.0837	-1.3731	2.1394
	X_3	0.1756	0.2329	2.3830	0.9374	2.5239	0.2343

Para cada um destes casos já foram feitas algumas observações na seção 6.2.1. Agora veremos comparativamente os casos com diferentes matrizes ENTRE. Olhando nos coeficientes da Tabela 6.2.10, nos Círculos das Correlações, Tabela 6.2.12 e nos valores das correlações na Tabela 6.2.11, podemos fazer as seguintes observações:

⇒ A aparição de contraste nos casos 2 e 3 deriva-se de discordâncias entre as estruturas de correlação DENTRO e ENTRE para os referidos casos, a diferença do caso 1, onde estas estruturas apresentam comportamentos idénticos.

⇒ A semelhança entre os valores dos coeficientes *Bruto* e *Dentro* reflete a igualdade das variâncias a um.

⇒ A estrutura de correlação *Entre* os grupos mostra, para a correlação com a primeira discriminante nos casos 1 e 3, como as três variáveis influenciam grandemente na separação dos grupos, com o contraste decorrente da oposição das estruturas de correlação DENTRO e ENTRE. Já no caso 2 as

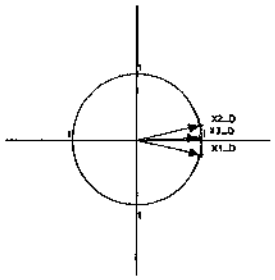
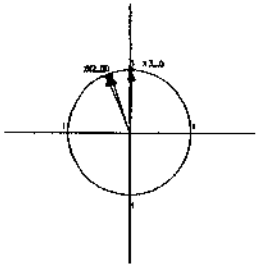
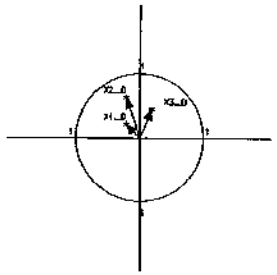
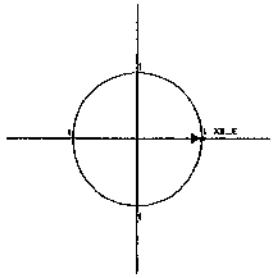
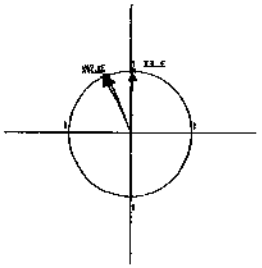
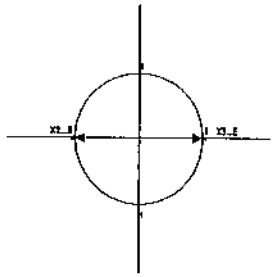
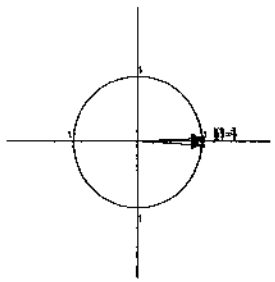
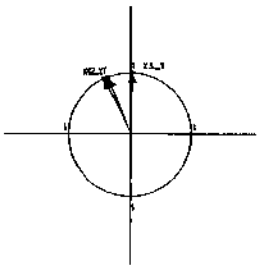
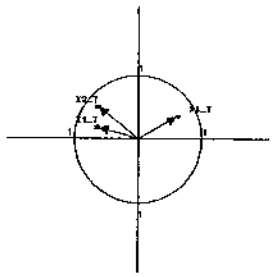
baixas correlações com $\mathbf{can_1}$ e altas com $\mathbf{can_2}$ indicam a necesidade das duas canônicas para ver a separação entre grupos e a não preferencialidade de $\mathbf{can_1}$ por uma direção especifica.

⇒ Por outro lado a estrutura de correlação *Dentro* não acompanha em todos os casos a orientação das médias. A exceção do caso 1, onde acompanha muito bem, os outros caso não mostram o mesmo comportamento. No caso 2 observa-se o mesmo comportamento da *Entre*, devido precisamente a ausência de correlação ENTRE os grupos.No caso 3 a correlação *Dentro* dos grupos é baixa nas duas canônicas, neste caso, para todas as variáveis.

Tabela 6.2.11: Correlações entre originais e discriminantes (Casos 1-2-3)

Estrutura de Correlação	Variáveis	CASO 1		CASO 2		CASO 3	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	0.9984	-0.0548	-0.4261	0.8998	-0.6205	0.1822
	X_2	0.9981	0.0564	-0.3714	0.9220	-0.6188	0.5083
	X_3	0.9966	0.0115	0.0298	0.9995	0.6087	0.3639
Entre	X_1	1.0000	-0.0000	-0.4366	0.8996	-1.0000	-0.0000
	X_2	1.0000	0.0000	-0.3816	0.9243	-1.0000	0.0000
	X_3	1.0000	0.0000	0.0305	0.9995	1.0000	-0.0000
Dentro	X_1	0.9734	-0.2241	-0.3808	0.9005	-0.1997	0.2277
	X_2	0.9693	0.2297	-0.3283	0.9128	-0.1988	0.6342
	X_3	0.9448	0.0457	0.0266	0.9993	0.1938	0.4499

Tabela 6.2.12: CÍRCULOS DAS CORRELAÇÕES para os casos 1-2-3

CASO 1	CASO 2	CASO 3
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

- **Casos 4, 5 e 6**

Estes casos têm em comum uma matriz de correlações "DENTRO" dos grupos igual à identidade, as correlações "ENTRE" os grupos variam, dependendo do caso.

Tabela 6.2.13: Coeficientes (Casos 4-5-6)

Variação	Variáveis	CASO 4		CASO 5		CASO 6	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	1.4884	-1.3805	1.9091	1.2383	1.5576	-2.0497
	X_2	1.4561	2.0228	0.0821	0.8792	1.4648	1.0859
	X_3	1.3043	-0.6403	-1.6175	1.4387	-1.4738	-0.9750
Dentro	X_1	0.5789	-0.5369	0.77034	0.4996	0.6058	-0.7973
	X_2	0.5770	0.8016	0.0536	0.5746	0.5805	0.4303
	X_3	0.5287	-0.2595	-0.6797	0.6046	-0.5974	-0.3952
Bruto	X_1	0.6006	-0.5571	0.7992	0.5184	0.6286	-0.8272
	X_2	0.5856	0.8135	0.0544	0.5831	0.5891	0.4367
	X_3	0.5223	-0.2564	-0.6715	0.5973	-0.5902	-0.3904

Respeito a estes casos podemos dizer o seguinte:

⇒ Observa-se contrastes quando as estruturas de correlação ENTRE e DENTRO dos grupos estão discordantes, que es o caso de 5 e 6.

⇒ Igualmente igualdade dos valores dos coeficientes *Dentro* e *Bruto* nos três casos, decorre da igualdade das variâncias a um.

⇒ Os Círculos das Correlações, Tabela 6.2.15, mostra na estrutura de correlação *Entre* que nos casos 4 e 6 as variáveis te, altas correlações com **can1**, em contraste no caso 6, pela presença de correlações ENTRE negativas. No caso 5

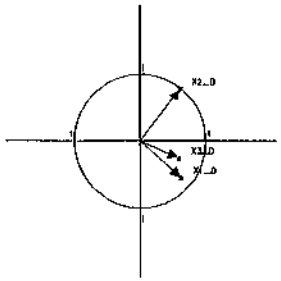
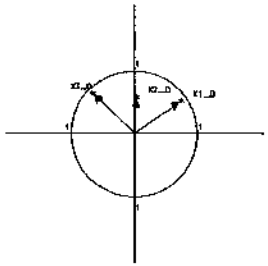
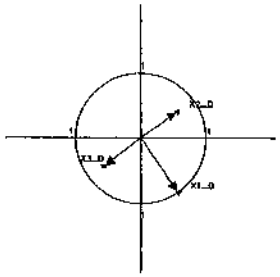
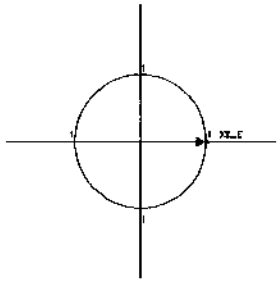
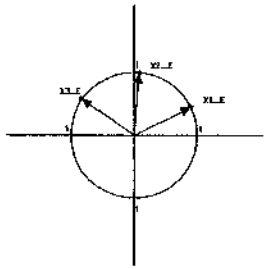
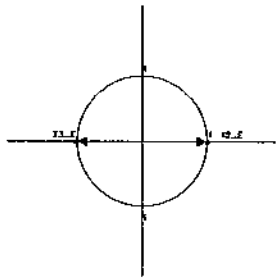
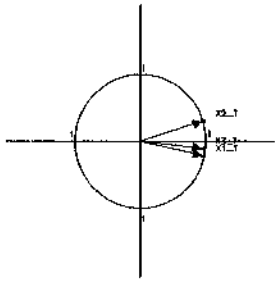
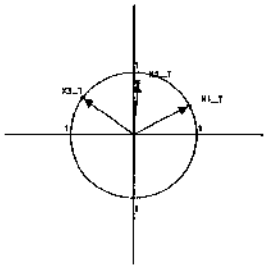
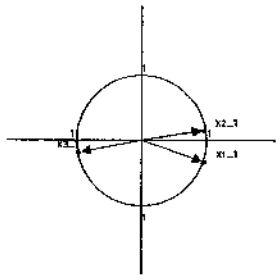
a ausência de correlação tanto DENTRO como ENTRE os grupos faz conque se precise das duas variáveis discriminantes para ver a separação dos grupos.

⇒ A estrutura de correlação *Dentro* dos grupos mostra como a ausência de correlação DENTRO dos grupos nos três casos faz com que as variáveis em geral não apresentem direção preferencial com as discriminantes. No entanto conservam-se os contrastes entre as variáveis, respeito aos observados na estrutura de correlação *Entre*.

Tabela 6.2.14: Correlações entre originais e discriminantes (Casos 4-5-6)

Estrutura de Correlação	Variáveis	CASO 4		CASO 5		CASO 6	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	0.9510	-0.2142	0.8664	0.4694	0.9479	-0.3182
	X_2	0.9478	0.3142	0.0609	0.8488	0.9447	0.1654
	X_3	0.9437	-0.1064	-0.7829	0.5971	-0.9406	-0.1718
Entre	X_1	1.0000	0.0000	0.8898	0.4562	1.0000	-0.0000
	X_2	1.0000	0.0000	0.0756	0.9971	1.0000	0.0000
	X_3	1.0000	0.0000	-0.8108	0.5852	-1.0000	-0.0000
Dentro	X_1	0.6072	-0.5509	0.7339	0.5320	0.5738	-0.8184
	X_2	0.5939	0.7933	0.0318	0.5939	0.5613	0.4177
	X_3	0.5781	0.5781	-0.6368	0.6497	-0.5464	-0.4241

Tabela 6.2.15: CÍRCULOS DAS CORRELAÇÕES para os casos 4-5-6

CASO 4	CASO 5	CASO 6
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

- **Casos 7, 8 e 9**

Estes três casos têm a mesma matriz de correlação dentro dos grupos, i.e., $\rho_{12}^{D} = +0{,}9$, $\rho_{13}^{D} = -0{,}9$ e $\rho_{23}^{D} = -0{,}9$. Queremos ver o efeito da mudança na matriz de correlações ENTRE os grupos.

Tabela 6.2.16: Coeficientes (Casos 7-8-9)

Variação	Variáveis	CASO 7		CASO 8		CASO 9	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	1.5550	-2.5326	-4.6217	1.5659	1.2501	-5.6117
	X_2	1.4490	2.6813	-0.0004	0.6363	1.1447	10.1611
	X_3	2.9391	-0.1398	-5.5377	-0.1291	-1.8656	4.5430
Dentro	X_1	1.2738	-2.0746	-2.0319	0.6884	0.3058	-1.3728
	X_2	1.1887	2.1995	-0.0002	0.2860	0.2811	2.4958
	X_3	2.3966	-0.1140	-2.1920	-0.0511	-0.4505	1.0970
Bruto	X_1	1.3216	-2.1525	-2.1081	0.7143	0.3173	-1.4244
	X_2	1.2281	2.2724	-0.0002	0.2955	0.2905	2.5785
	X_3	2.5212	-0.1199	-2.3059	-0.0537	-0.4739	1.1541

Ao respeito destes casos podemos comentar o seguinte:

⇒ Os coeficientes *Bruto* e *Dentro* são semelhantes devido as variâncias iguais a um.

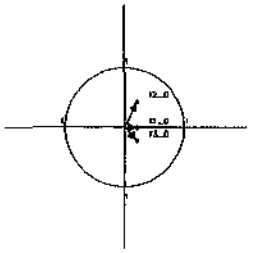
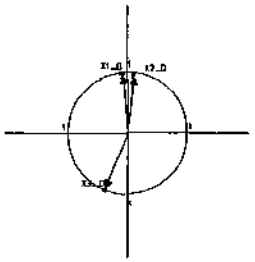
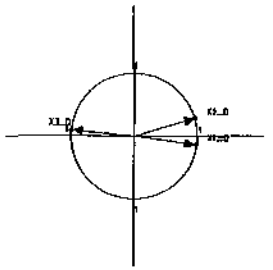
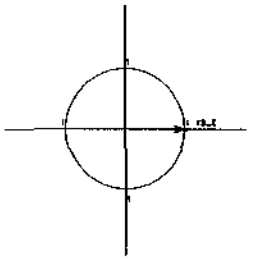
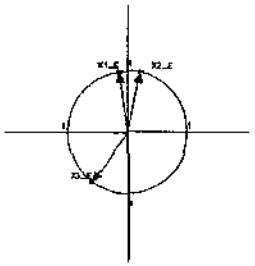
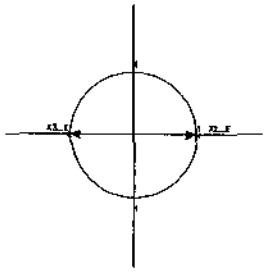
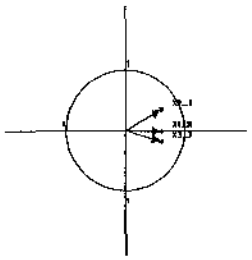
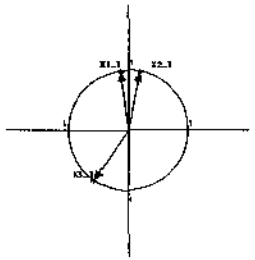
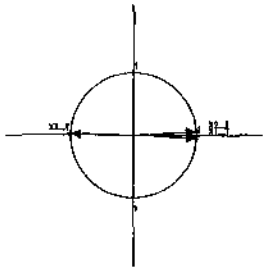
⇒ Quando temos pelo menos alguma correlação ENTRE negativa ou zero, como nos casos 8 e 9, observam-se contraste nos coeficientes.

⇒ Observamos os Círculos das Correlações na Tabela 6.2.18 vemos como as três variáveis têm altíssima correlação *Entre* os grupos com a variável **can1** (ver os valores na Tabela 6.2.17), e portanto uma influência grande na separação dos grupos.

Tabela 6.2.17: Correlações entre originais e discriminantes (Casos 7-8-9)

Estrutura de Correlação	Variáveis	CASO 7		CASO 8		CASO 9	
		Can1	Can2	Can1	Can2	Can1	Can2
Total	X_1	0.5978	-0.0015	-0.1302	0.9902	0.9976	-0.0277
	X_2	0.5961	0.3620	0.2065	0.9695	0.9973	0.0717
	X_3	0.6034	-0.1776	-0.5613	-0.8265	-0.9984	0.0254
Entre	X_1	1.0000	0.0000	-0.1396	0.9902	1.0000	-0.0000
	X_2	1.0000	0.0000	0.2226	0.9748	1.0000	0.0000
	X_3	1.0000	-0.0000	-0.5886	-0.8083	-1.0000	0.0000
Dentro	X_1	0.2046	-0.0019	-0.0798	0.9903	0.9590	-0.1133
	X_2	0.2037	0.4415	0.1238	0.9483	0.9550	0.2920
	X_3	0.2074	-0.2179	-0.3822	-0.9180	-0.9724	0.1053

Tabela 6.2.18: CÍRCULOS DAS CORRELAÇÕES para os casos 7-8-9

CASO 7	CASO 8	CASO 9
a) ESTRUTURA DE CORRELAÇÃO DENTRO DOS GRUPOS		
		
b) ESTRUTURA DE CORRELAÇÃO ENTRE OS GRUPOS		
		
c) ESTRUTURA DE CORRELAÇÃO PARA A AMOSTRA TOTAL		
		

⇒ Quando as correlações "ENTRE" e "DENTRO" dos grupos coincidem, a estrutura de correlação *Dentro* acompanha muito bem a orientação das médias, como podemos ver na Tabela 6.2.18 a), caso 9. No caso 7 não podemos dizer nada, pois as correlações *Dentro* são pequenas com as duas discriminantes e no caso 8 a *Dentro* não acompanhará a direção de $\mathbf{can}_1$.

6.3 Conclusões derivadas dos resultados das simulações

6.3.1 Conclusões sobre o comportamento dos coeficientes.

1. A magnitude dos coeficientes em si é pouco informativa.
2. Os coeficientes chamados de *Bruto* quando comparadas as magnitude dos coeficientes das distintas variáveis observou-se que:
   - Uma mesma ordem de magnitude é indicativa em geral de variâncias iguais.
   - Contraste podem ser indicativos de padrão de correlação negativa ou nula ao longo dos grupos.
   - Distintas ordens de magnitude refletem variâncias distintas das variáveis dentro ou entre os grupos.
3. Quanto aos coeficientes chamados *Dentro*, quando comparadas as magnitudes dos coeficientes das distintas variáveis, observou-se que:
   - Uma mesma ordem de magnitude é indicativa de variâncias iguais dentro dos grupos.
   - Contrastes são dominados por correlação entre os grupos negativa e nula ao longo dos grupos.
   - Distintas ordens de magnitude refletem variâncias distintas dentro dos grupos.
4. As diferenças entre magnitude dos coeficientes *Bruto* e *Dentro* referem-se apenas às diferenças de variâncias das variáveis originais dentro dos grupos.
   - Uma mudança de coeficiente na mesma variável do *Bruto* para o *Dentro* deve-se à variância dentro dos grupos distinta de um.

Grandes mudanças devem-se a grande variabilidade dentro dos grupos.

- Se as magnitudes dos coeficientes *Dentro* são aproximadamente múltiplas dos coeficientes *Bruto* então as variâncias são aproximadamente iguais ao longo das variáveis.
- Se as contribuições de cada variável aparecem relativamente alteradas do *Dentro* para o *Bruto*, então as variâncias das variáveis originais dentro dos grupos são possivelmente distintas e as maiores contribuições ocorrem pelas variáveis de maior variância.

5. As magnitudes dos coeficientes chamados *Total* devem ser lidas em comparação com os coeficientes *Bruto* e *Dentro*. As diferenças melhor se explicam pela contribuição da variabilidade entre relativamente à variabilidade dentro dos grupos.
   - Os sinais podem ficar mantidos ou alterarem-se: no primeiro caso indica que a variável dentro do grupo define uma direção de alongamento que concorda com a direção de alongamento ao longo dos grupos, no segundo caso, se o sinal do coeficiente de uma variável se altera do *Total* para o *Dentro* e *Bruto*, há indicação que a direção de alongamento da variável dentro do grupo é diferente ou está em oposição à direção de alongamento através dos grupos.

6.3.2 Conclusões sobre o comportamento das correlações entre as variáveis originais e as discriminantes.

1. Se as correlações entre as originais e a primeira canônica apresentarem direções concordantes e boa representação nos Círculos de Correlações, então as direções de alongamento das nuvens dentro dos grupos e ao longo dos grupos tendem a ser coincidentes.

2. Se as correlações *Entre* forem altas e *Dentro* baixas para a primeira direção canônica então há indicação que o alongamento das nuvens dentro dos grupos tem direção oposta ao alongamento definido pela nuvem das médias dos grupos.

3. Variâncias distintas com mesma variabilidade total tem maior impacto nas correlações quando a estrutura dentro dos grupos é de correlações fracas, por exemplo comparação do caso 7 com o caso 10, ambos bivariados.

4. Contraste em $\mathbf{can}_1$ pode acontecer quando a orientação ao longo dos grupos tem estrutura de correlação negativa, como nos casos 3, 9 e 15, bivariados.

5. Outras situações podem gerar o aparecimento de contrastes na primeira direção canônica, portanto (4) é condição necessária mas não suficiente.

6. Correlações baixas entre as variáveis originais e a primeira canônica podem estar presentes mesmo quando o poder discriminante é alto. Podem estar associados à ausência de correlação ao longo dos grupos (exemplo casos 5, 11 e 17) acentuadas se a nuvem dentro dos grupos apresentar direção preferencial de alongamento (casos 5 e 17)

6.3.3 Procedimento para análise dos Círculos das Correlações

Círculo Dentro

⇒ Revisar primeiro como estão representadas as variáveis no Círculo das Correlações

- Se estiverem bem representadas:
  - Observar como se comporta a correlação entre as variáveis originais e recuperar a estrutura de correlação entre as variáveis originais.
  - Observar o comportamento ao longo das direções dadas pelas variáveis discriminantes. As correlações com as primeira e segunda direções discriminantes dão elementos para a interpretação.
- Se não estiverem bem representadas:
  - Não dá para recuperar a estrutura de correlação das variáveis originais. Aconselha-se recorrer á Análise de Componentes Principais para cada grupo, caso a hipótese de igualdade das matrizes de variâncias e covariâncias seja refutada. No caso de não haver evidência de desigualdade das matrizes de variâncias e covariâncias dentro dos grupos, então desenvolver a Análise de Componentes Principais na S_{pooled} pode ser simultaneamente aplicada aos grupos.

Círculo Entre

⇒ Revisar primeiro como estão representadas as variáveis no Círculo das Correlações

- Se estiverem bem representadas:
  - Observar se as correlações com a primeira direção discriminante são altas. Em caso afirmativo podem-se ler as correlações entre as variáveis originais e as canônicas de forma a recuperar a estrutura de correlação entre.
  - Observar se as correlações se distribuem entre as direções discriminantes e/ou são altas com a segunda direção discriminante. Neste caso há indicações de correlações baixas entre as variáveis originais na estrutura de correlação entre os grupos. Não se recupera a estrutura entre a partir dos ângulos entre as variáveis originais. Exemplo caso 5.
- Se não estiverem bem representadas:
  - Não é possível recuperar as correlações das variáveis originais na estrutura entre. As baixas correlações com as direções discriminantes são indicativas da não participação destas variáveis na separação dos grupos nas direções nas que isto acontece.

6.3.4 Conclusões sobre o significado dos Círculos de Correlações.

Círculo para a estrutura de correlação Dentro: Orientar na explicação das diferenças entre indivíduos dentro de cada grupo. No caso de distância ao centro do círculo pequena para todas as variáveis, fica necessário voltar aos dados originais para explicar as diferenças dos indivíduos dentro dos grupos.

Círculo para a estrutura de Correlação Entre: Orientar na explicação das diferenças entre indivíduos de grupos distintos.

Círculo para a estrutura de Correlação Total: Orientar na explicação das diferenças entre indivíduos quando ignora a estrutura de grupos.

Capítulo 7: Aplicação em um exemplo com dados reais

Para a aplicação dos resultados obtidos foram usados dados de uma pesquisa a cargo do Prof. Dr. George Shepherd, orientando o trabalho de doutorado de Fabio Vitta, ambos do Instituto de Biologia da UNICAMP. Trata-se dos dados morfométricos de 4 amostras de populações diferentes de *Lagenocarpus* (familia *Cyperaceae*) coletadas na Serra do Cipó, MG, elas são: *Lagenocarpus tenuifolius* azul, *Lagenocarpus tenuifolius* amarelo, *Lagenocarpus rigidus* verde e *Lagenocarpus rigidus* azul. Cada uma com 24, 24, 37 e 19 indivíduos respectivamente. Além disso temos 29 plantas, híbridas, das quais não se conhece o grupo de origem.

Descrição do Problema

Os dois grupos principais, i.e., *Lagenocarpus tenuifolius* e *Lagenocarpus rigidus*, podem ser facilmente identificados pelo seu tamanho, mas nem tanto os indivíduos dentro destes dois grupos, pois é difícil saber se eles pertencem ao grupo "azul" ou "amarelo", no caso dos *tenuifolius*, ou ao "verde" ou "azul", no caso dos *rigidus*. Pesquisar a importância que desempenham as variáveis medidas na diferenciação destes grupos e conseguir enxergar o comportamento dos híbridos, respeito ao resto dos indivíduos são dois dos aspectos de interesse que deverão ser resolvidos. Outra coisa é classificar cada híbrido em um dos grupos conhecidos com o que tenha características comuns.

Descrição das variáveis

Foram medidas 14 variáveis. Os nomes, tipos e unidades utilizadas estão apresentados na Tabela a seguir:

Variável[1]	Descrição	Unidade	Tipo
X_1	Escapo comprimento	cm	Contínua
X_2	Escapo largura	cm	Contínua
X_4	Nós estéreis		Discreta
X_5	Nós masculinos		Discreta
X_6	Parte masculina a partir de	cm	Contínua
X_7	Comprimento parte feminina	cm	Contínua
X_8	Folha comprimento	cm	Contínua
X_9	Folha largura	cm	Contínua
X_{10}	Bainha folha	cm	Contínua
X_{11}	Biomassa "caule"	g	Contínua
X_{12}	Biomassa parte feminina	g	Contínua
X_{13}	Biomassa parte masculina	g	Contínua
X_{14}	Fruto altura	mm	Contínua
X_{15}	Fruto largura	mm	Contínua

Metodologia eststística proposta

1. Análise descritiva univariada: construção dos gráficos de Box-Plot[2], matriz de correlações.

2. Uso do *Stepdisc* para reduzir o número de variáveis, separando as que aportam mais a separação dos grupos.

3. Desenvolver a Análise Discriminante de Fisher para estudar as inter-relações entre as variáveis. Desenhar os planos discriminantes e os Círculos das

[1] A variável X_3 foi tirada da análise por ser soma de outras duas.
[2] Conhecido também com o nome de Desenho Esquemático

Correlações para ver o comportamento dos indivíduos e das variáveis, respectivamente. Interpretação da importância das variáveis na separação dos grupos, usando estes resultados.

4. Tratamento dos híbridos como indivíduos suplementares. Classificação em um dos grupos. Caso particular de dados faltantes.

5. Localização dos indivíduos suplementares no gráfico discriminante. Interpretação dos resultados.

Resultados

1. Estatísticas descritivas

As estatísticas univariadas básicas para as variáveis medidas encontram-se a seguir, primeiro por grupo e depois para a amostra geral.

Estatísticas para o Grupo 1: *L. tenuifolius azul*

GRUPOS=1

Variable	N	Mean	Std Dev	Minimum	Maximum
X1	24	66.604	12.399	46.000	87.000
X2	24	1.377	0.205	1.000	1.800
X4	24	0.000	0.000	0.000	0.000
X5	24	2.792	0.658	2.000	4.000
X6	24	33.417	8.493	18.000	48.000
X7	24	10.525	3.172	5.500	18.000
X8	24	38.500	9.690	24.000	60.000
X9	24	2.024	0.369	1.200	2.800
X10	24	6.550	1.985	3.000	10.500
X11	24	0.465	0.197	0.190	0.975
X12	24	0.086	0.046	0.022	0.255
X13	24	0.136	0.078	0.032	0.382
X14	24	2.549	0.164	2.200	2.900
X15	24	1.467	0.131	1.100	1.800

Estatísticas para o Grupo 2: *L. tenuifolius amarelo*

GRUPOS=2

Variable	N	Mean	Std Dev	Minimum	Maximum
X1	24	49.242	7.520	34.000	65.000
X2	24	0.957	0.131	0.700	1.160
X4	24	0.000	0.000	0.000	0.000
X5	24	2.250	0.737	1.000	4.000
X6	24	26.888	6.823	12.000	39.000
X7	24	7.133	2.387	3.500	12.000
X8	24	24.700	4.221	18.600	34.000
X9	24	1.922	0.359	1.100	2.510
X10	24	5.350	0.882	3.800	7.000
X11	24	0.186	0.049	0.110	0.264
X12	24	0.050	0.033	0.015	0.180
X13	24	0.098	0.048	0.024	0.214
X14	24	2.436	0.548	0.000	2.840
X15	24	1.331	0.356	0.000	2.260

Estatísticas para o Grupo 3: *L. rigidus verde*

GRUPOS=3

Variable	N	Mean	Std Dev	Minimum	Maximum
X1	38	198.921	37.241	146.000	285.000
X2	38	5.014	0.844	3.700	7.500
X4	38	3.974	0.944	2.000	6.000
X5	38	3.158	0.855	2.000	5.000
X6	38	118.566	34.072	62.000	215.000
X7	38	35.605	10.697	20.000	67.000
X8	38	76.308	20.514	42.500	123.000
X9	38	10.536	2.085	6.300	14.500
X10	38	8.205	1.740	5.100	13.000
X11	38	23.333	12.012	9.200	60.550
X12	38	3.237	2.516	1.080	14.360
X13	38	3.686	2.298	0.450	9.330
X14	38	2.333	0.240	2.000	2.900
X15	38	1.272	0.121	1.000	1.500

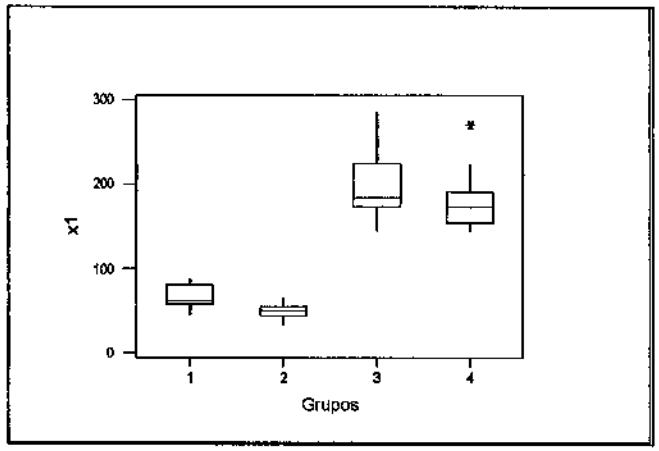
Estatísticas para o Grupo 4: *L. rigidus azul*

GRUPOS=4

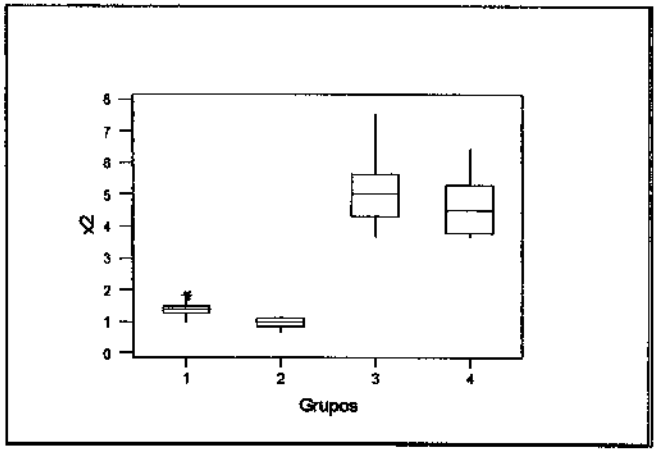
Variable	N	Mean	Std Dev	Minimum	Maximum
X1	19	178.947	31.026	146.000	270.000
X2	19	4.621	0.815	3.700	6.400
X4	19	4.421	0.961	3.000	6.000
X5	19	3.579	1.170	2.000	6.000
X6	19	97.895	27.039	53.000	168.000
X7	19	39.263	7.971	28.000	56.000
X8	19	61.053	14.890	40.000	90.000
X9	19	11.326	2.235	7.300	14.800
X10	19	5.316	0.506	5.000	7.000
X11	19	16.058	4.757	9.280	26.000
X12	19	2.435	1.067	0.920	4.300
X13	19	2.508	1.080	0.750	4.360
X14	19	3.153	0.262	2.700	3.550
X15	19	1.899	0.119	1.700	2.150

Nos gráficos de Box-Plot, os quais podem ser vistos a seguir, algumas variáveis exibem a simples vista a separação dos indivíduos *L.tenuifolius*, grupos 1 e 2, dos *L.rigidus*, grupos 3 e 4, elas são X_1, X_2, X_4, X_6, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_14 e X_{15}, outras variáveis, no entanto, parecem mostrar facilidade para separar especificamente os indivíduos dos grupos 3 e 4, elas são X_{14}, X_{15} e X_{10}. Apesar disto, só a análise multivariada dos dados, onde também entrarão as inter-relações entre variáveis, poderá dar uma resposta certa ao poder discriminante das variáveis na presença de todas.

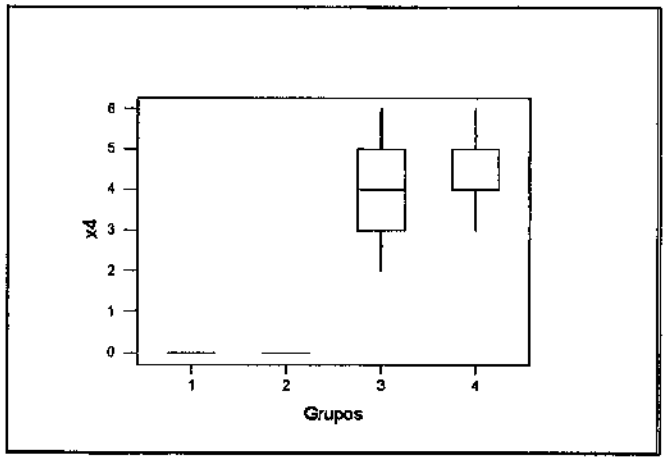
Box-Plot



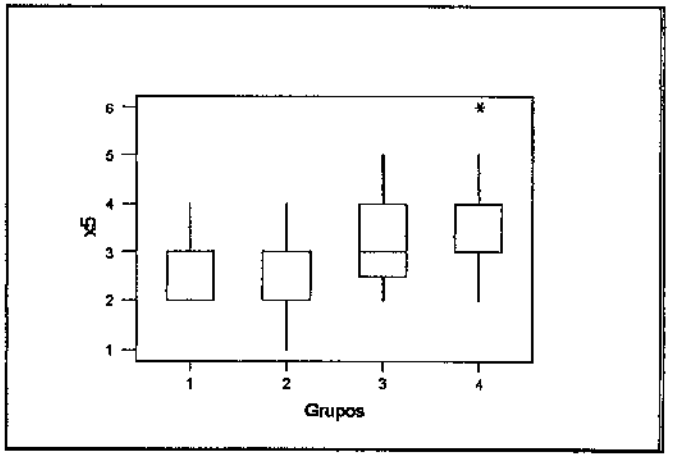
Box-Plot da variável X_1: Escapo Comprimento, para os 4 grupos.



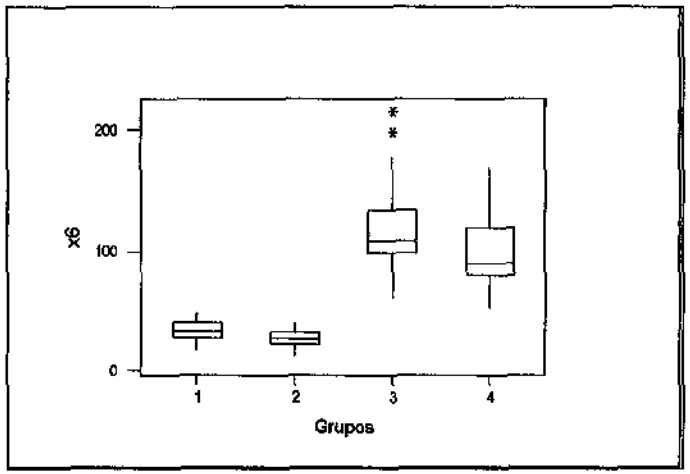
Box-Plot da variável X_2: Escapo Largura, para os 4 grupos.



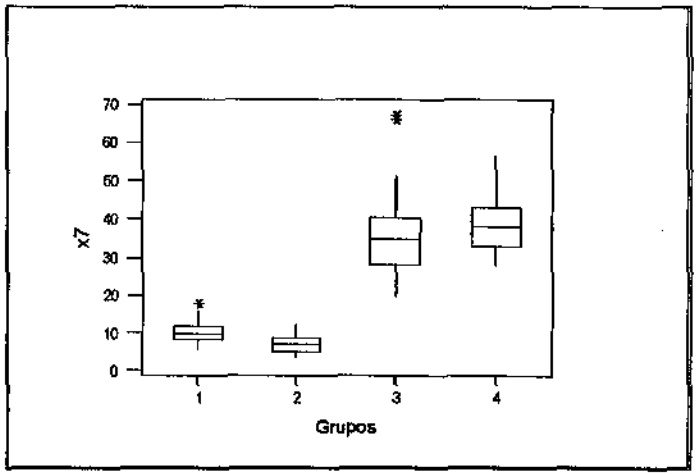
Box-Plot da variável X_4: Nós estéreis, para os 4 grupos.



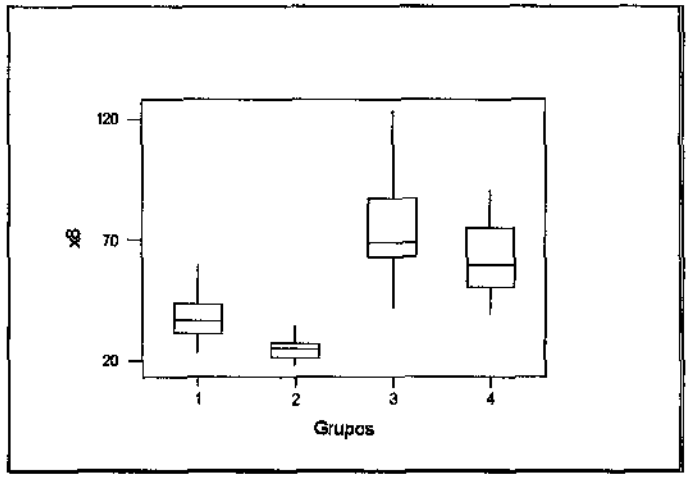
Box-Plot da variável X_5: Nós Masculinos, para os 4 grupos.



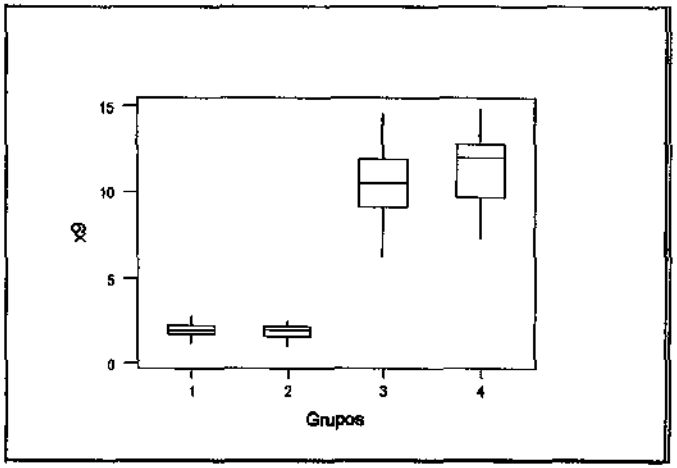
Box-Plot da variável X_6: Parte masculina a partir de, para os 4 grupos.



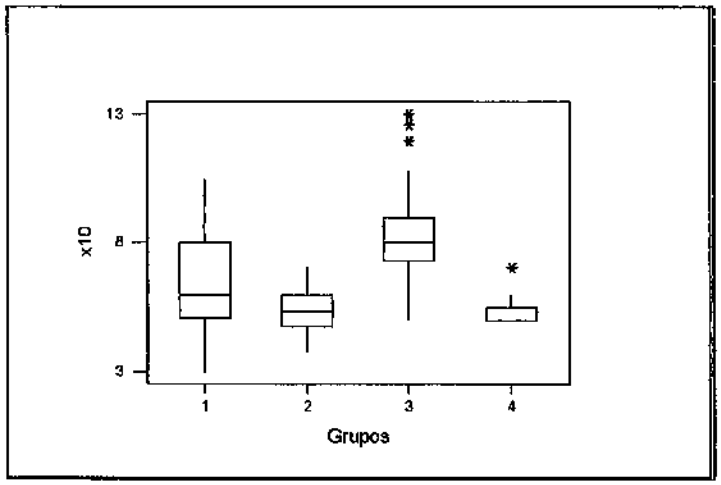
Box-Plot da variável X_7, Comprimento parte feminina, para os 4 grupos.



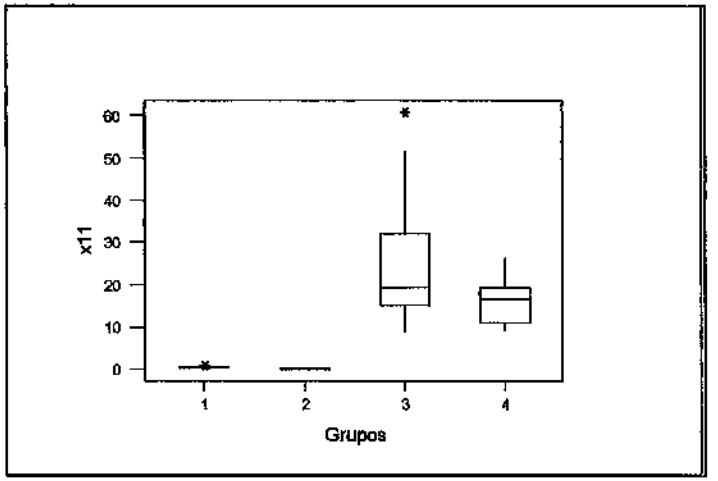
Box-Plot da variável X_8: Folha Comprimento, para os 4 grupos.



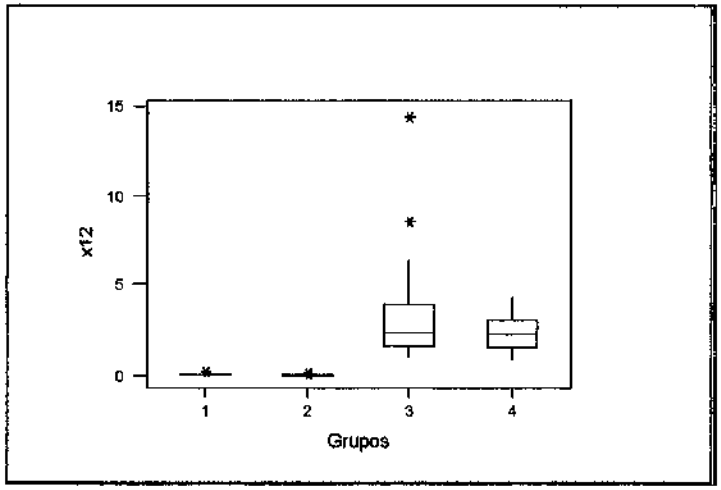
Box-Plot da variável X_9: Folha Largura, para os 4 grupos.



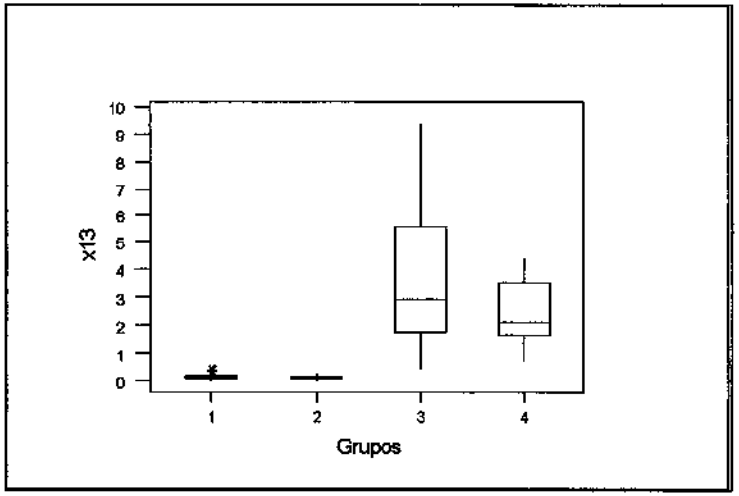
Box-Plot da variável X_{10}: Bainha Folha, para os 4 grupos.



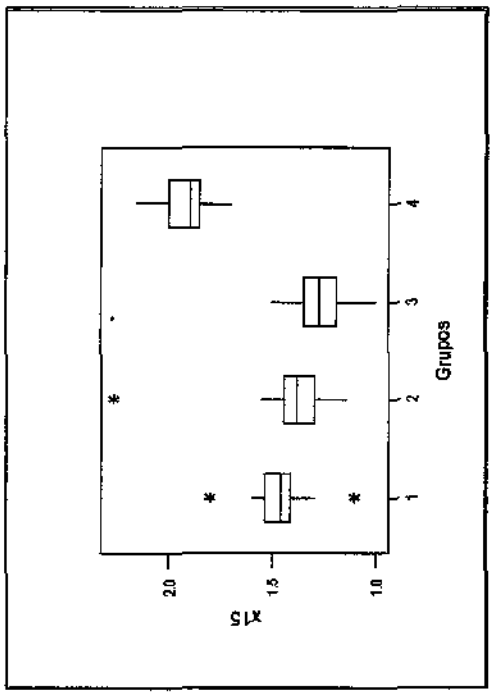
Box-Plot da variável X_{11}: Biomassa "caule" para os 4 grupos.



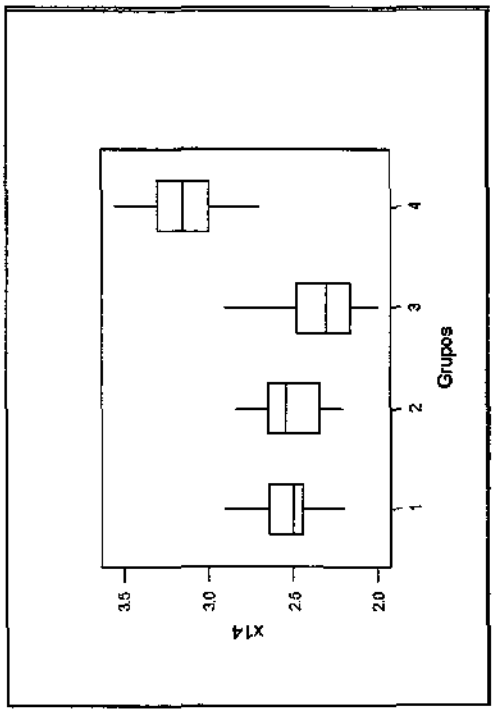
Box-Plot da variável X_{12}: Biomassa parte feminina, para os 4 grupos.



Box-Plot da variável X_{13}: Biomassa parte masculina, para os 4 grupos.



Box-Plot da variável X_{15}: Fruto largura, para os 4 grupos.



Box-Plot da variável X_{14}: Fruto Altura, para os 4 grupos.

Construiu-se a matriz de correlações para pesquisar o comportamento da correlação linear entre as variáveis medidas. Na figura 7.15, a seguir, mostram-se os valores tabelados:

	X_1	X_2	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
X_1	1.000	0.947	0.918	0.383	0.958	0.903	0.885	0.913	0.516	0.920	0.755	0.766	-0.005	0.060
X_2	0.947	1.000	0.912	0.441	0.882	0.881	0.874	0.953	0.446	0.909	0.794	0.828	0.035	0.070
X_4	0.918	0.912	1.000	0.319	0.892	0.854	0.744	0.910	0.315	0.804	0.645	0.667	0.120	0.165
X_5	0.383	0.441	0.319	1.000	0.232	0.290	0.416	0.427	0.139	0.320	0.188	0.424	0.199	0.196
X_6	0.958	0.882	0.892	0.232	1.000	0.803	0.854	0.847	0.526	0.895	0.697	0.690	-0.083	0.008
X_7	0.903	0.881	0.854	0.290	0.803	1.000	0.758	0.878	0.346	0.807	0.795	0.706	0.161	0.184
X_8	0.885	0.874	0.744	0.416	0.854	0.758	1.000	0.831	0.705	0.884	0.750	0.751	-0.028	-0.026
X_9	0.913	0.953	0.910	0.427	0.847	0.878	0.831	1.000	0.358	0.835	0.743	0.743	0.138	0.170
X_{10}	0.516	0.446	0.315	0.139	0.526	0.346	0.705	0.358	1.000	0.563	0.418	0.448	-0.322	-0.352
X_{11}	0.920	0.909	0.804	0.320	0.895	0.807	0.884	0.835	0.563	1.000	0.816	0.818	-0.092	-0.011
X_{12}	0.755	0.794	0.645	0.188	0.697	0.795	0.750	0.743	0.418	0.816	1.000	0.715	-0.006	-0.003
X_{13}	0.766	0.828	0.667	0.424	0.690	0.706	0.751	0.743	0.448	0.818	0.715	1.000	-0.068	-0.059
X_{14}	-0.005	0.035	0.120	0.199	-0.083	0.161	-0.028	0.138	-0.322	-0.092	-0.006	-0.068	1.000	0.691
X_{15}	0.060	0.070	0.165	0.196	0.008	0.184	-0.026	0.170	-0.352	-0.011	-0.003	-0.059	0.691	1.000

Matriz de Correlações para os dados das plantas.

A matriz de correlações na forma dos diagramas de dispersão e descrevendo os histogramas correspondentes, para cada variável, é mostrada na seguinte figura.

Matriz de Correlações para os dados das plantas na forma de diagramas de dispersão e histogramas.

Observou-se que várias variáveis estavam fortemente correlacionadas entre si. Por isto a matriz foi reorganizada para formar um bloco com estas variáveis, resultando na matriz a seguir, onde 10, das 13 variáveis trabalhadas, estão fortemente correlacionadas entre sim.

	X_1	X_6	X_2	X_{11}	X_4	X_9	X_7	X_8	X_{13}	X_{12}	X_{10}	X_5	X_{15}	X_{14}
X_1	1.000	0.958	0.947	0.920	0.918	0.913	0.903	0.885	0.766	0.755	0.516	0.383	0.060	-0.005
X_6	0.958	1.000	0.882	0.895	0.892	0.847	0.803	0.854	0.690	0.697	0.526	0.232	0.008	-0.083
X_2	0.947	0.882	1.000	0.909	0.912	0.953	0.881	0.874	0.828	0.794	0.446	0.441	0.070	0.035
X_{11}	0.920	0.895	0.909	1.000	0.804	0.835	0.807	0.884	0.818	0.816	0.563	0.320	-0.011	-0.092
X_4	0.918	0.892	0.912	0.804	1.000	0.910	0.854	0.744	0.667	0.645	0.315	0.319	0.165	0.120
X_9	0.913	0.847	0.953	0.835	0.910	1.000	0.878	0.831	0.743	0.743	0.358	0.427	0.170	0.138
X_7	0.903	0.803	0.881	0.807	0.854	0.878	1.000	0.758	0.706	0.795	0.346	0.290	0.184	0.161
X_8	0.885	0.854	0.874	0.884	0.744	0.831	0.758	1.000	0.751	0.750	0.705	0.416	-0.026	-0.028
X_{13}	0.766	0.690	0.828	0.818	0.667	0.743	0.706	0.751	1.000	0.715	0.448	0.424	-0.059	-0.068
X_{12}	0.755	0.697	0.794	0.816	0.645	0.743	0.795	0.750	0.715	1.000	0.418	0.188	-0.003	-0.006
X_{10}	0.516	0.526	0.446	0.563	0.315	0.358	0.346	0.705	0.448	0.418	1.000	0.139	-0.352	-0.322
X_5	0.383	0.232	0.441	0.320	0.319	0.427	0.290	0.416	0.424	0.188	0.139	1.000	0.196	0.199
X_{15}	0.060	0.008	0.070	-0.011	0.165	0.170	0.184	-0.026	-0.059	-0.003	-0.352	0.196	1.000	0.691
X_{14}	-0.005	-0.083	0.035	-0.092	0.120	0.138	0.161	-0.028	-0.068	-0.006	-0.322	0.199	0.691	1.000

Matriz de Correlações reorganizada para destacar os blocos de variáveis fortemente correlacionadas

Por outro lado, X_{10}, X_5, X_{15} e X_{14} têm correlações baixas com o resto das variáveis, sendo as de valores menores X_{15} e X_{14}, elas têm a ver com altura e largura dos frutos, respectivamente. Além disto, duas correlações altas aparecem isoladas, é o caso de X_{10} com X_8 e X_{14} com X_{15}, com valores de 0,701 e 0,691, respectivamente.

2. Step Disc

Para reduzir o número de variáveis decidiu-se fazer inicialmente um stepdisc para ficar com aquelas de maior poder discriminatório. Após esta análise, ficaram 9 variáveis das 14 iniciais. As variáveis selecionadas foram: X_1, X_2, X_4, X_6, X_8, X_9, X_{10}, X_{11} e X_{15}, e é com elas que continuaremos a análise. Os resultados do stepdisc podem ser vistos a seguir:

Stepwise Discriminant Analysis

Stepwise Selection: Summary

Step	Variable Entered	Variable Removed	Number In	Partial R**2	F Statistic	Prob > F	Wilks' Lambda	Prob < Lambda
1	X2		1	0.9017	308.692	0.0001	0.09833732	0.0001
2	X15		2	0.5648	43.265	0.0001	0.04279377	0.0001
3	X4		3	0.4102	22.956	0.0001	0.02523776	0.0001
4	X11		4	0.3636	18.661	0.0001	0.01606202	0.0001
5	X1		5	0.2765	12.354	0.0001	0.01162152	0.0001
6	X10		6	0.1555	5.894	0.0010	0.00981401	0.0001
7	X9		7	0.1447	5.356	0.0019	0.00839433	0.0001
8	X6		8	0.0978	3.398	0.0210	0.00757296	0.0001
9	X8		9	0.0600	1.978	0.1227	0.00711877	0.0001

Step	Variable Entered	Variable Removed	Number In	Average Squared Canonical Correlation	Prob > ASCC
1	X2		1	0.30055423	0.0001
2	X15		2	0.48870885	0.0001
3	X4		3	0.52542350	0.0001
4	X11		4	0.54725413	0.0001
5	X1		5	0.57310943	0.0001
6	X10		6	0.60661855	0.0001
7	X9		7	0.63367381	0.0001
8	X6		8	0.63826756	0.0001
9	X8		9	0.65112815	0.0001

Sobre esta seleção podemos comentar que as variáveis escolhidas são as que parecem mostrar melhor, até aqui somente com os resultados univariados, a separação dos grupos. Quanto as não escolhidas, X_5, X_7, X_{12}, X_{13} e X_{14}, tudo parece indicar que na presença das outras elas não contribuem com o mesmo aparente poder

discriminatório mostrado nas análises univariados a exceção de X_5, a qual já mostrava seu baixo poder discriminatório na figura 7.4.

Resultados da Análise Discriminante de Fisher trabalhando com 4 grupos e 9 variáveis, escolhidas usando o procedimento de *stepdisc*.

Se observamos os autovalores da matriz $\mathbf{S}_D^{-1}\mathbf{S}_E$, nos resultados a seguir, vemos que o primeiro acumula 89,3% da variância entre grupos e o segundo 9,46%. Somando os dois primeiros eles acumulam 98,77% da variância, ou seja, quase toda a variância do sistema pode ser explicada pelas duas primeiras variáveis discriminantes. Em termos do problema botânico que nos ocupa as variáveis discriminantes serão investigadas para saber a informação que carregará cada uma.

Eigenvalues of INV(E)*H

	Eigenvalue	Difference	Proportion	Cumulative
1	30.6100	27.3663	0.8931	0.8931
2	3.2436	2.8228	0.0946	0.9877
3	0.4209	.	0.0123	1.0000

Neste ponto precisaremos dos gráficos dos planos discriminantes, os quais são contruídos a partir dos coeficientes brutos, centralizando as variáveis originais para uma melhor visualização dos indivíduos.

As médias dos quatro grupos nos planos canônicos estão dadas pelas seguintes coordenadas.

Class Means on Canonical Variables

GRUPOS	CAN1	CAN2	CAN3
1	-5.291880697	0.236407802	-0.986927364
2	-6.437982249	0.173903973	0.890231781
3	4.389650240	-1.858200641	0.028821609
4	6.037368505	3.198112724	0.064498571

No plano discriminante correspondente a $\mathbf{can_1}$ - $\mathbf{can_2}$ observa-se uma grande separação dos *L. tenuifolius* dos *L. rigidus*, quando projetados os pontos em $\mathbf{can_1}$. Segundo a descrição dos especialistas em botânica estas duas espécies são muito

diferenciadas fisicamente pelo tamanho das plantas em geral, isto indica que da interpretação da primeira variável discriminante devem destacar-se as variáveis de tamanho como as de maior influência.

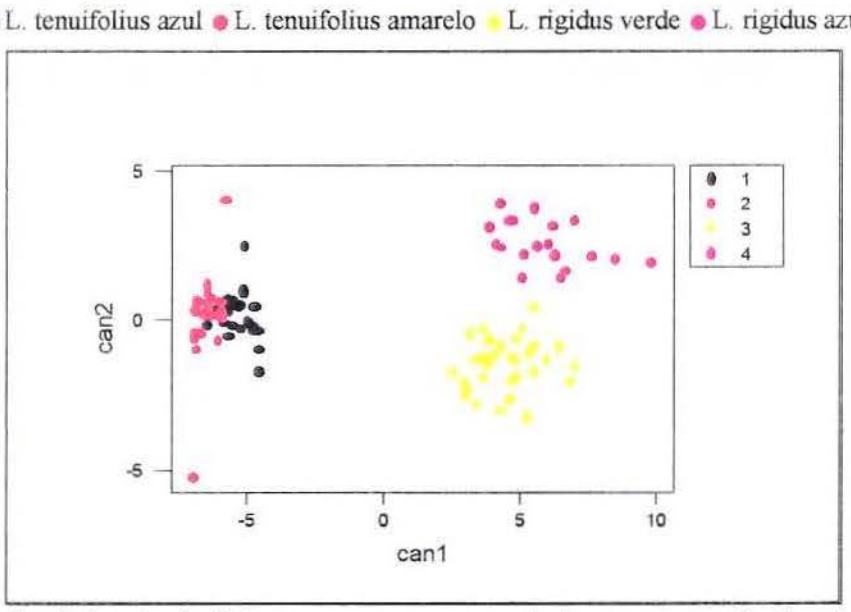


Figura 7.19: Plano discriminante de Fisher $\mathbf{can_1}$ - $\mathbf{can_2}$, caso de 4 grupos e variáveis selecionadas usando stepdisc

Mesmo sendo o nosso interesse separar os quatro grupos, a $\mathbf{can_1}$ não consegue discriminar os *L. tenuifolius* "azuis" dos "amarelos" nem os *L. rigidus* "verdes" dos "azuis". As variações existentes entre estes grupos passam a ocupar um segundo lugar de importância na presença de uma separação marcada entre os *L. tenuifolius* dos *L. rigidus*.

A segunda fonte de variação entre grupos de importância resulta das diferenças entre os *L.rigidus* "verdes" e *L.rigidus* "azuis", e é $\mathbf{can_2}$ quem carregará esta diferenciação. Isto pode ser visto no gráfico anterior se projetamos imaginariamente os dados sobre $\mathbf{can_2}$.

Projetando os indivíduos no plano $\mathbf{can_2}$ - $\mathbf{can_3}$ é obtido o seguinte gráfico, mostrado na figura 7.20,

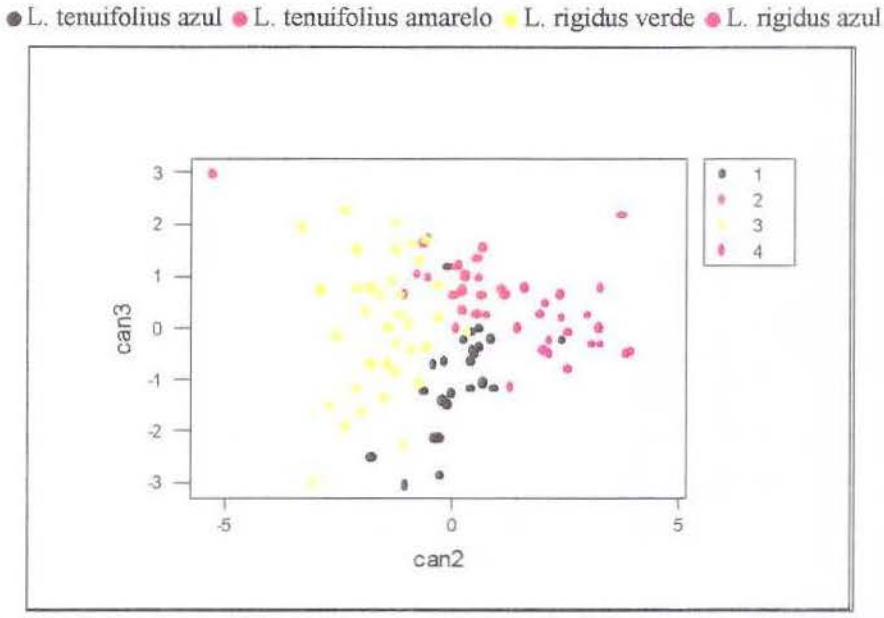


Figura 7.20: Plano discriminante de Fisher $\mathbf{can_2}$ - $\mathbf{can_3}$, caso de 4 grupos e variáveis selecionadas usando *stepdisc*

A melhor separação possível entre os *L. tenuifolius* "azuis" e os *L. tenuifolius* "amarelos" só será captada pela $\mathbf{can_3}$, como mostra o gráfico, sendo esta separação a mais sutil e difícil de conseguir. Para facilitar a visualização gráfica da separação destes dois grupos apresentamos a seguir o plano $\mathbf{can_1}$ - $\mathbf{can_3}$.

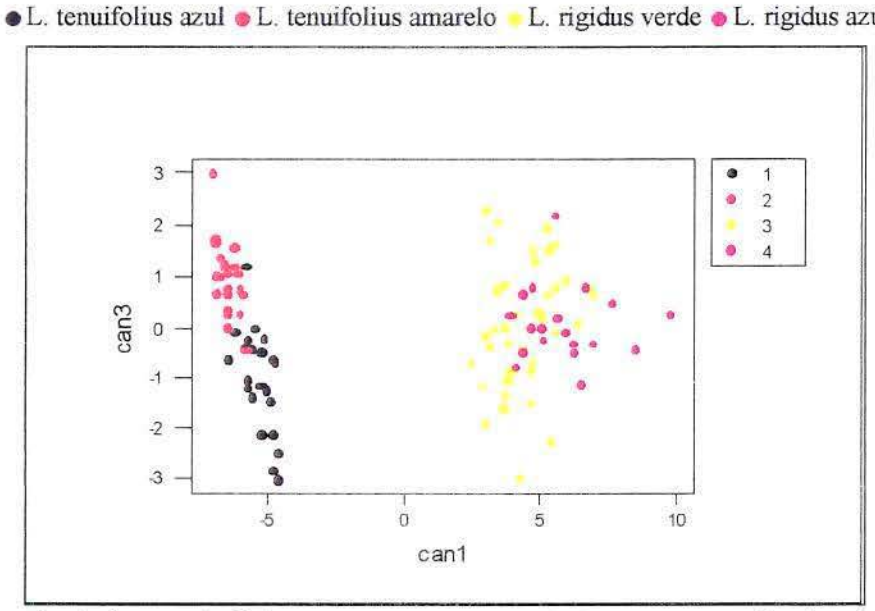


Figura 7.21: Plano discriminante de Fisher $\mathbf{can_1}$ - $\mathbf{can_3}$, caso de 4 grupos e variáveis selecionadas usando *stepdisc*

O problema seguinte será descobrir quais as variáveis que influem na aparição de cada uma das fontes de variação explicadas anteriormente. Para isto analisaremos os valores dos coeficientes e os valores das correlações entre originais e discriminantes.

A seguir temos os valores dos coeficientes *Bruto* (`Raw Canonical Coefficients`), *Dentro* (`Pooled Within-Class Standardized Canonical Coefficients`) e *Total* (`Total-Sample Standardized Canonical Coefficients`). Primeiramente analisaremos a ordem de importância das variáveis segundo a magnitude dos coeficientes[3]. Após a análise dos coeficientes nos auxiliaremos das estruturas de correlação *Entre*, *Dentro* e *Total* para concluir a interpretação de quais variáveis são mais importantes na separação dos quatro grupos, i.é, *L. tenuifolius azul*, *L. tenuifolius amarelo*, *L. rigidus verde* e *L. rigidus azul*.

[3] Na listagem da ordem de importância das variáveis apareceram em **negrito** os valores maiores, em *itálico* valores também grandes, embora em segundo lugar de importância. Não serão destacados os coeficientes com valores muito pequenos.

Coeficientes Discriminantes de Fisher

Coeficiente *Bruto*:

Raw Canonical Coefficients

	CAN1	CAN2	CAN3
X1	0.045281395	0.003490701	-0.019184962
X2	1.557095868	-0.577138754	-1.374454139
X4	0.861853495	0.361597355	0.395632821
X6	-0.023137850	-0.016788101	0.001296037
X8	-0.004604977	0.012877571	-0.055389331
X9	0.155254935	0.111503107	0.542053084
X10	-0.088894190	-0.253915054	-0.119005900
X11	-0.186980561	-0.015130143	0.180680424
X15	0.782056305	5.830572807	-2.139368113

Ordem de importância das variáveis segundo a magnitude do coeficiente *Bruto*	
Can1	**X_2, X_4, X_{15}**, X_{11}, X_9, X_{10}, X_1, X_6, X_8
Can2	**X_{15}**, X_2, X_4, X_{10}, X_9, X_6, X_{11}, X_8, X_1
Can3	**X_{15}**, **X_2**, X_9, X_4, X_{11}, X_{10}, X_8, X_1, X_6

Coeficiente *Dentro*:

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2	CAN3
X1	1.221346679	0.094152491	-0.517463937
X2	0.976078626	-0.361784276	-0.861588123
X4	0.603999583	0.253412736	0.277265290
X6	-0.558475071	-0.405212066	0.031282260
X8	-0.068166219	0.190623170	-0.819913171
X9	0.247593721	0.177820235	0.864442342
X10	-0.132716156	-0.379086976	-0.177671965
X11	-1.410482073	-0.114133766	1.362957183
X15	0.117060137	0.872734668	-0.320225951

Ordem de importância das variáveis segundo a magnitude do coeficiente *Dentro*	
Can1	**X_{11}**, **X_1**, **X_2**, X_4, X_6, X_9, X_{10}, X_{15}, X_8
Can2	**X_{15}**, X_6, X_{10}, X_2, X_4, X_8, X_9, X_{11}, X_1
Can3	**X_{11}**, X_9, X_2, X_8, X_1, X_{15}, X_4, X_{10}, X_6

Coeficiente *Total*:

Total-Sample Standardized Canonical Coefficients

	CAN1	CAN2	CAN3
X1	3.300152561	0.254405723	-1.398218841
X2	3.067394440	-1.136932055	-2.707600136
X4	1.880356802	0.788918361	0.863175551
X6	-1.106731896	-0.803010093	0.061992159
X8	-0.117987435	0.329945526	-1.419169993
X9	0.729309186	0.523785217	2.546291313
X10	-0.172509329	-0.492751162	-0.230944540
X11	-2.422571737	-0.196030308	2.340945420
X15	0.208584066	1.555085709	-0.570595873

Ordem de importância das variáveis segundo a magnitude do coeficiente *Total*	
Can1	**X_1, X_2, X_{11}**, X_4, X_6, X_9, X_{15}, X_{10}, X_8
Can2	**X_{15}, X_2**, X_6, X_4, X_9, X_{10}, X_8, X_1, X_{11}
Can3	**X_2, X_9, X_{11}**, X_8, X_1, X_4, X_{15}, X_{10}, X_6

Estruturas de Correlação

Correlação *Entre*:

Segundo a estrutura de correlação *Entre*, para as correlações entre originais e discriminantes, as variáveis X_4, X_9, X_2, X_1, X_6, X_{11}, X_8[4] têm altas correlações positivas com **can1** , como é exibido no Círculo das Correlações para o plano **can1 - can2** .

Estas variáveis: Nós estéreis, Largura folha, Largura escapo, Comprimento escapo, Parte masculina, Biomassa "caule" e Comprimento folha são as mais fortemente correlacionadas com a direção de máxima separação dos *L.tenuifolius* e os *L.rigidus*. Isto quer dizer que estas características das plantas aumentarão na direção de aumento de **can1** .

Voltando ao plano discriminante, figura 7.21, vemos que os indivíduos pertencentes aos dois grupos do tipo *L.rigidus* apresentam os maiores valores na variável **can1** , portanto eles apresentaram magnitude maior de Nós estéreis, Largura folha, Largura escapo, Comprimento escapo, Parte masculina, Biomassa "caule" e

[4] Neste caso a ordem não é de importância.

Comprimento folha. Sendo que para os indivíduos pertencentes aos grupos do tipo *L.tenuifolius* os valores estas variáveis serão menores respeito aos valores para *L.rigidus*.

Tal e como era já suspeitado, pelas características físicas das plantas, estas variáveis dizem respeito a diferenças no tamanho e esta análise corrobora que os *L.rigidus* são maiores que os *L.tenuifolius*.

Estrutura de Correlação *Entre* os grupos (**CAN1-CAN2**)

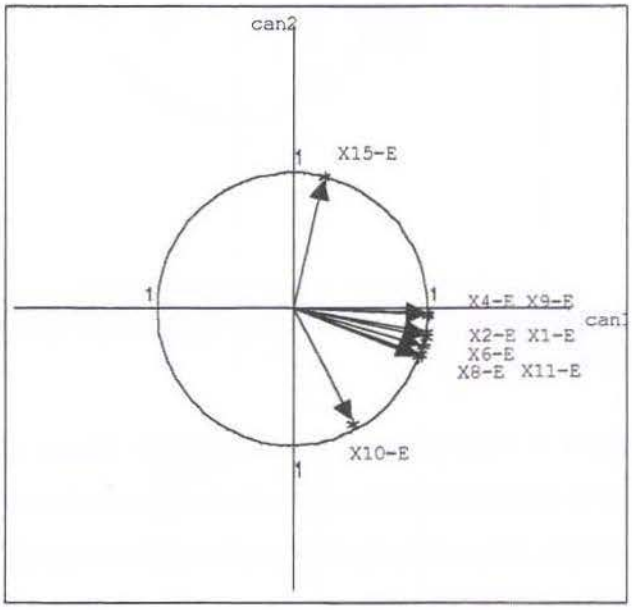


Figura 7.22: Círculo das Correlações *Entre*, plano $\mathbf{can_1}$ - $\mathbf{can_2}$, caso de 4 grupos e variáveis selecionadas usando stepdisc

Quando analisamos os coeficientes *Brutos*, construídos para os dados padronizados, tínhamos visto que os maiores coeficientes eram precisamente os das variáveis X_1, X_2, X_{11}, X_4, X_6, X_9, reafirmando que são elas (a menos de X_8) que aportam as características fundamentais que diferenciam os *L.tenuifolius* dos *L.rigidus*.

Para analisar quais as variáveis que determinam a separação dos grupos na segunda dimensão, **can2**, vejamos como se comporta a estrutura de correlações *Entre*, entre originais e discriminantes, agora no plano **can2** - **can3**.

Estrutura de Correlação *Entre* os grupos (**CAN2-CAN3**)

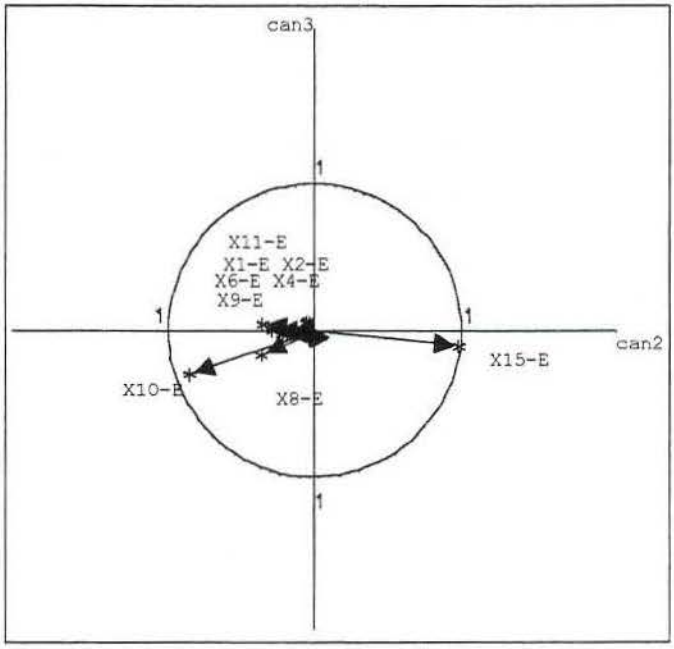


Figura 7.23: Círculo das Correlações *Entre*, plano $\mathbf{can_2}$-$\mathbf{can_3}$, caso de 4 grupos e variáveis selecionadas usando *stepdisc*

No Círculo para a estrutura de correlações *Entre* os grupos, vemos que as variáveis X_{15} e X_{10} são as mais destacadas com correlações altas com **can2**. Além disso elas estão em contraste, X_{15} com sinal positivo e X_{10} negativo. Isto indica que valores maiores da Largura do Fruto, X_{15}, e menores para a Bainha da Folha, X_{10}, se corresponderão com os maiores valores de **can2**. Da mesma forma valores grandes para a Bainha da Folha, X_{10}, e pequenos da Largura do Fruto, X_{15}, estarão relacionados aos menores valores de **can2**.

Voltando no plano discriminante, figura 7.22, onde **can2** separa os *L. rigidus* "verdes" dos *L. rigidus* "azuis", observamos que os *L.rigidus* "azuis" correspondem com valores altos de **can2** e portanto maior Largura do Fruto e menor Bainha da Folha. Os *L.rigidus* "verdes", por outro lado, com valores menores de **can2**, terão menor Largura do Fruto e maior Bainha da Folha.

A análise dos coeficientes *Brutos*, padronizados, tinha nos levado a dizer que, quando referentes a **can2**, as variáveis com maiores magnitude eram X_{15} e X_2, ou seja, a própria Largura do Fruto, X_{15}, em contraste com a largura do escapo, X_2 tiveram

magnitude maior, no caso dos coeficientes *Brutos*, para separar os *L.rigidus* "verdes" dos *L.rigidus* "azuis".

Os valores numéricos destas correlações podem ser vistos a seguir:

Between Canonical Structure

	CAN1	CAN2	CAN3
X1	0.978208	-0.206700	-0.019585
X2	0.983997	-0.177999	-0.008158
X4	0.996986	-0.031274	0.071001
X6	0.960796	-0.277001	0.011878
X8	0.923152	-0.348993	-0.161228
X9	0.997030	-0.044295	0.062994
X10	0.445891	-0.842905	-0.301153
X11	0.938404	-0.341274	0.054133
X15	0.235610	0.967088	-0.096064

Neste mesmo Círculo das Correlações, analisaremos a terceira dimensão **can3**, que separa os *L.tenuifolius* "azuis" dos *L.tenuifolius* "verdes". Aqui observamos que são as variáveis X_{10}, X_8 e X_{15} as que têm as correlações maiores, as duas com sinal negativo; com **can3**.

Voltando no plano discriminante, figura 7.22, vemos que os indivíduos do grupo *L.tenuifolius azul* apresentam os menores valores em **can3**. Sendo assim, este grupo apresentará valores para a Bainha da Folha, X_{10}, Folha Comprimento, X_8, e a Largura do Fruto, X_{15}, grandes respeito a estas mesmas características no grupo *L.tenuifolius verde*, que tem os valores maiores de **can3**.

Quando foi visto o coeficientes *Bruto*, para os dados padronizados, ficava X_2, X_9, X_{11}, X_8, X_1.

Correlação *Dentro*:

Referindo-nos à estrutura de correlação *Dentro* podemos dizer que as variáveis Nós estéreis, X_4, Folha Largura, X_9, Escapo Comprimento, X_1 e Escapo Largura, X_2, estão positiva e fracamente correlacionadas com **can1**, isto significa que, dentro dos grupos, a direção de alongamento das nuvens não acompanha nestas variáveis a direção da máxima separação entre *L.tenuifolius* e *L.rigidus*.

Estrutura de Correlação DENTRO dos grupos (**CAN1-CAN2**)

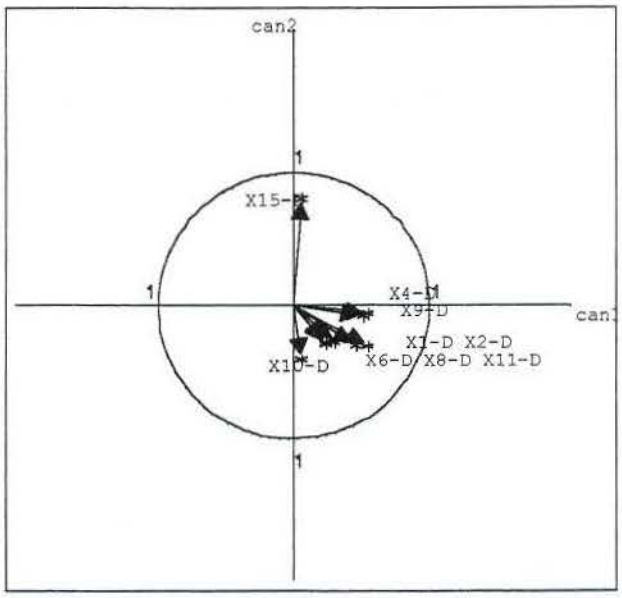


Figura 7.24: Círculo das Correlações *Dentro*, plano $\mathbf{can}_1$ - $\mathbf{can}_2$, caso de 4 grupos e variáveis selecionadas usando stepdisc

Relativamente a **can2**, podemos dizer que a Largura do Fruto, X_{15}, destaca-se com um valor de correlação de 0.81, coincidindo com o resultado dado pelos coeficientes *Dentro*. Isto quer dizer que é a Largura do Fruto quem acompanha dentro dos grupos a direção de máxima separação entre os *L.rigidus* "verdes" e os *L.rigidus* "azuis".

O Círculo de Correlação para **can2** - **can3** é mostrado a seguir. Em **can3** destacam-se a largura do Fruto, X_{15}, a Bainha da Folha, X_{10} e a Folha Comprimento, X_8, cujos valores no plano discriminante diferenciaram os *L.tenuifolius* "azuis" dos "amarelos", e vemos que também respondem pelas diferenças dentro destes grupos. Isto em contraste com os Nós estéreis, X_4, e a Folha Largura, X_9 que não acompanharam a direção de **can3**.

Estrutura de Correlação Dentro (**CAN2-CAN3**)

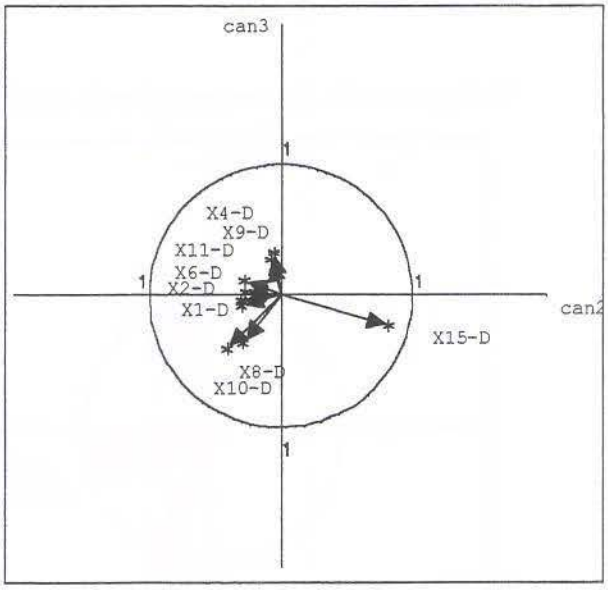


Figura 7.25: Círculo das Correlações *Dentro*, plano $\mathbf{can_2}$ - $\mathbf{can_3}$, caso de 4 grupos e variáveis selecionadas usando stepdisc

Os valores numéricos para estas correlações *Dentro* estão a seguir:

Pooled Within Canonical Structure

	CAN1	CAN2	CAN3
X1	0.451395	-0.293009	-0.077076
X2	0.538549	-0.299271	-0.038080
X4	0.539994	-0.052035	0.327962
X6	0.302976	-0.268331	0.031943
X8	0.240928	-0.279799	-0.358851
X9	0.507608	-0.069277	0.273514
X10	0.069317	-0.402538	-0.399265
X11	0.242113	-0.270486	0.119110
X15	0.064152	0.808903	-0.223067

Correlação *Total*:

As correlações *Totais* são uma espécie de resumo do que foi visto para a estrutura de correlação *Entre* e *Dentro*, desta vez vendo o resultado para a amostra total, sem levar em conta a separação em grupos.

Segundo este resultado as variáveis de maiores correlações com a direção de **can1** são X_4, X_9, X_2, X_1, X_6, X_{11}, X_8, todas com correlações altas e positivas. A seguir o

Círculo das Correlações para este caso mostra mais claramente a magnitude destas correlações.

Estrutura de Correlação Total (**CAN1-CAN2**)

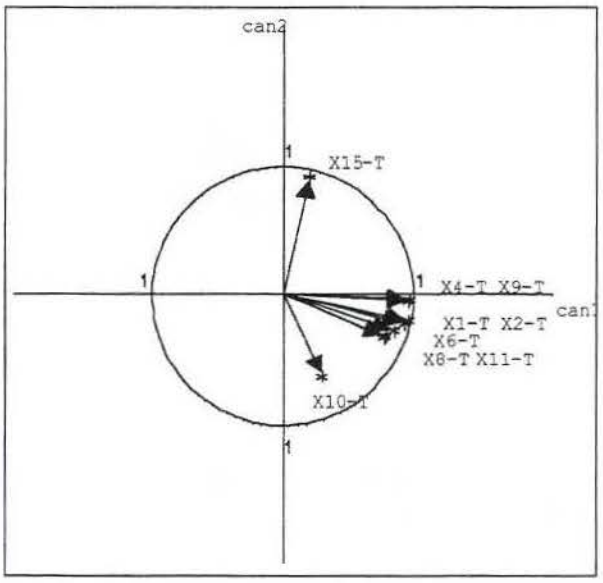


Figura 7.26: Círculo das Correlações *Total*, plano $\mathbf{can_1}$-$\mathbf{can_3}$, caso de 4 grupos e variáveis selecionadas usando stepdisc

Para a segunda discriminante, a variável de maior valor de correlação, e muito bem representada neste plano, é X_{15} com 0,92, o mesmo resultado para o caso do coeficiente *Total*, onde X_{15} é a de maior valor. No caso de **can3** o padrão *Entre* está bem aproximado do *Total*. Isto indica pouca influência das correlações *Dentro* que compareceram baixas, indicando perda da estrutura de correlação *Dentro*, que é forte e, como consequência, as diferenças entre plantas dentro dos grupos ficam sem explicação no plano discriminante.

Estrutura de Correlação Total (**CAN2-CAN3**)

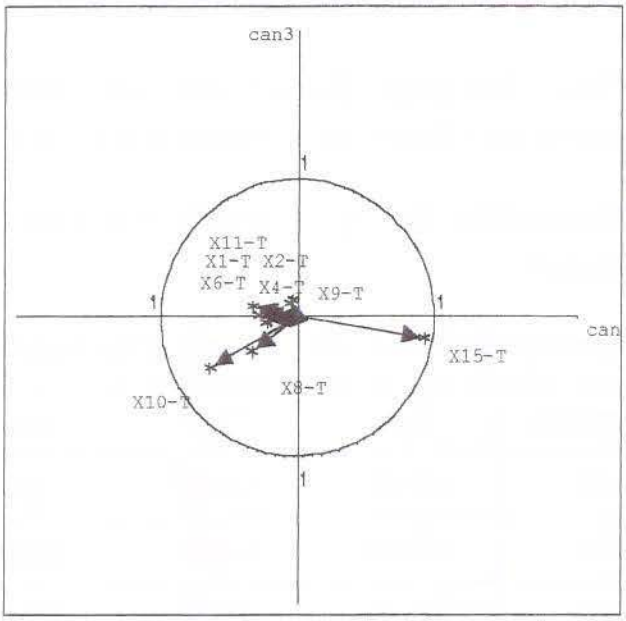


Figura 7.27: Círculo das Correlações *Total*, plano $\mathbf{can}_2$-$\mathbf{can}_3$, caso de 4 grupos e variáveis selecionadas usando stepdisc

Os resultados numéricos para estas correlações foram:

Total Canonical Structure

	CAN1	CAN2	CAN3
X1	0.925588	-0.220140	-0.033508
X2	0.949503	-0.193327	-0.014234
X4	0.961041	-0.033932	0.123749
X6	0.847081	-0.274882	0.018935
X8	0.771217	-0.328165	-0.243539
X9	0.954801	-0.047745	0.109076
X10	0.295468	-0.628684	-0.360822
X11	0.781024	-0.319705	0.081463
X15	0.199477	0.921590	-0.147056

Classificação dos indivíduos suplementares para o caso de 4 grupos e 9 variáveis.

Temos 29 indivíduos, chamados híbridos, dos quais desconhecemos o grupo a que pertencem; eles deverão ser classificados em um dos quatro grupos.

Desta forma poremos estas plantas na análise como indivíduos suplementares. Para isto faremos o seguinte:

1. Calcular as coordenadas dos híbridos nas discriminantes.

Indivíduos	Can1	Can2	Can3
171	1.41270	0.41338	-3.88329
172	0.86020	0.43034	-2.61074
173	1.03126	-0.59085	2.55811
175	2.16681	0.81310	-3.00624
176	1.11530	-0.72507	-3.21612
177	1.20004	-0.73622	-4.15119
178	0.06692	-1.18337	-2.47365
179	-0.77531	1.07344	-2.23692
220	1.06449	1.37872	0.24553
221	-0.02912	1.67645	-0.15244
222	0.01454	1.56065	-1.13831
223	0.07388	1.02482	-0.95116
224	-0.03535	1.36880	-0.64662
225	-0.52370	1.61655	0.59442
226	-1.15359	1.56269	-0.52886
227	0.32911	0.83002	0.10451
228	-1.57463	0.98847	-0.23537

Indivíduos	Can1	Can2	Can3
229	0.48738	0.21240	0.12084
233	-1.95671	3.55916	-2.72796
234	-2.57115	3.21886	-2.16550
235	-1.82854	3.72998	-1.28358
236	-1.49025	3.57156	0.05362
237	-1.55114	4.41086	-1.59206
238	-1.72576	3.43981	-1.19204
239	-0.73111	3.79531	-0.68434
Mirante A	-1.37933	4.46428	-2.47468
Mirante B	-1.38852	1.35954	-1.26267
Mirante C	-2.10589	1.08196	-3.12120
Mirante D	-2.46464	3.39948	-1.85754

2. Colocar estes indivíduos nos planos discriminantes, usando as coordenadas calculadas anteriormente, junto aos indivíduos dos grupos 1, 2, 3 e 4, para observar seu comportamento.

Gráficos dos Planos Discriminantes adicionando os Híbridos

Neste caso precisamos fazer algumas observações:

✓ No plano **can1 - can2** observamos que os híbridos formam um grupo independente dos outros quatro, relativamente às projeções sobre **can1**. Eles se concentram justamente no centro das projeções sobre **can1**. Como sabemos que esta direção é a que melhor exibe diferenças de tamanho entre as plantas, isto nos leva a pensar que estes indivíduos em geral terão um tamanho intermediário entre o tamanho natural dos grupos pertencentes aos *L.tenuifolius* e aos *L.rigidus*.

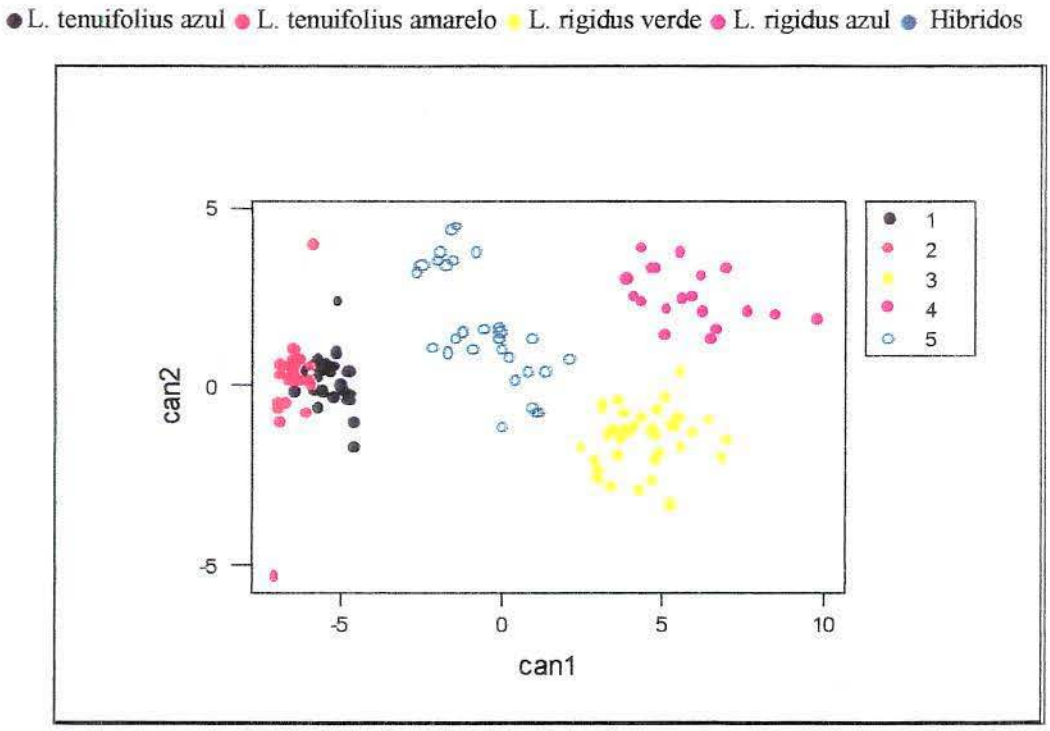


Figura 7.28: Plano discriminante $\mathbf{can}_1$ - $\mathbf{can}_2$, caso de 4 grupos e variáveis selecionadas usando stepdisc, após a alocação dos híbridos.

✓ Analisando a posição com respeito a cada um dos grupos observamos que se projetássemos os híbridos sobre **can2**, que é a direção que consegue separar os *L.rigidus* "verdes" dos *L.rigidus* "azuis", podemos dizer que o grupo com o que se mistura menos é com o *L.rigidus verde.*

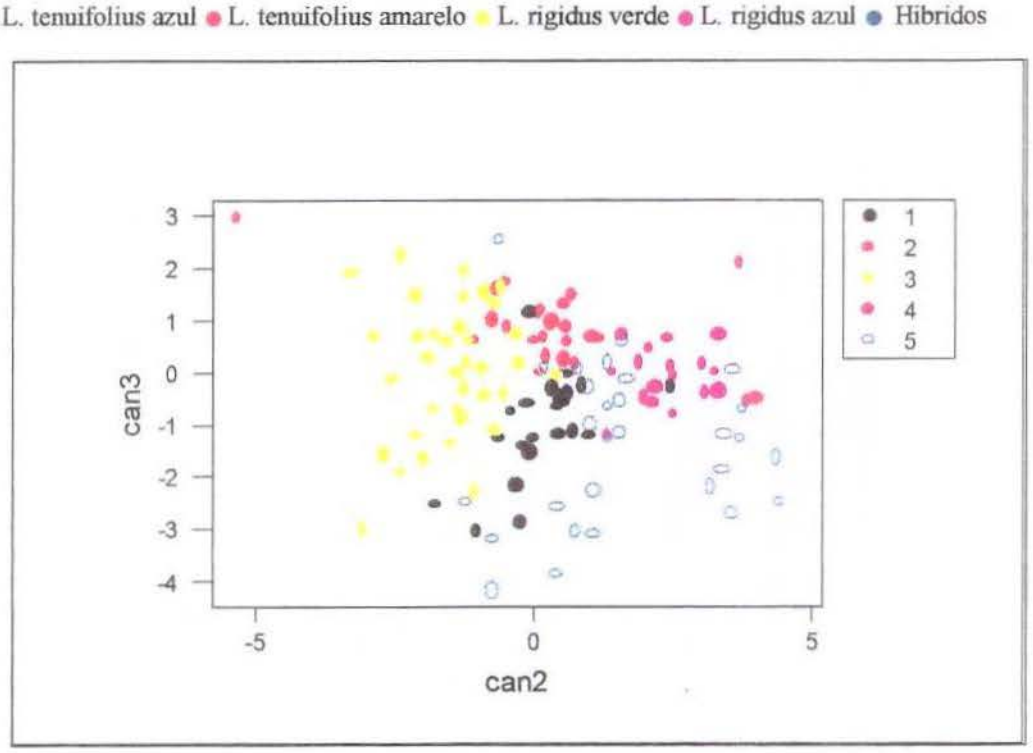


Figura 7.29: Plano discriminante $\mathbf{can_2}$ - $\mathbf{can_3}$, caso de 4 grupos e variáveis selecionadas usando stepdisc, após a alocação dos híbridos.

✓ Com respeito às projeções em **can3** ele só não se mistura muito com o *L.tenuifolius amarelo.*

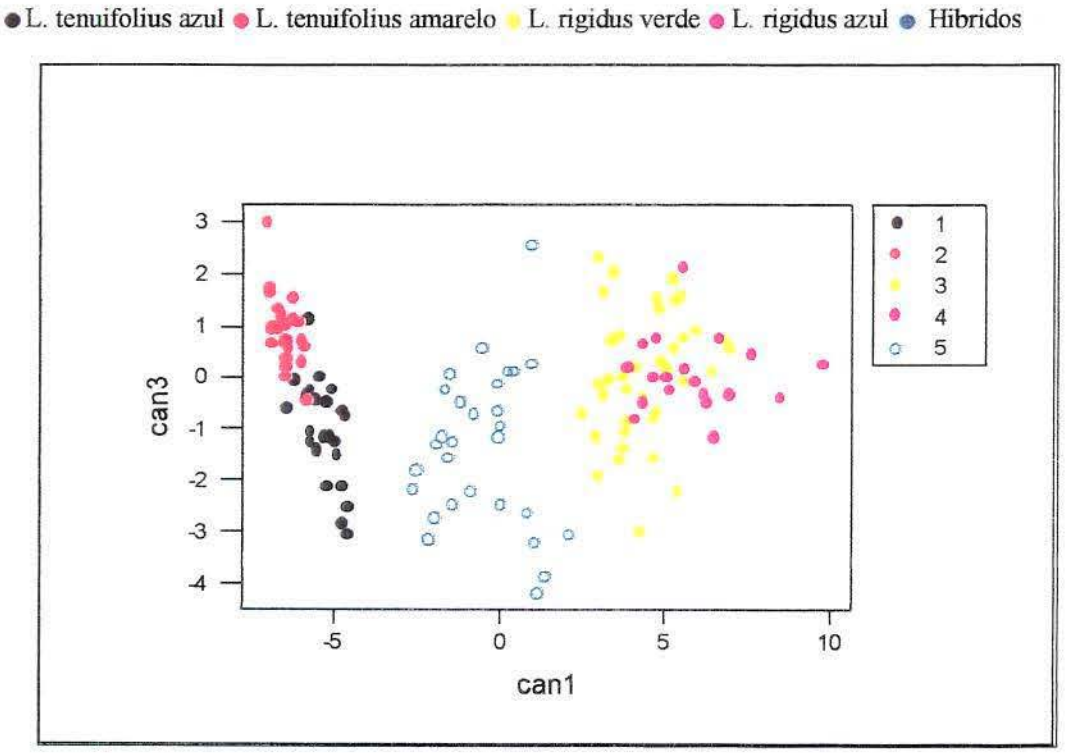


Figura 7.30: Plano discriminante $\mathbf{can_1}$ - $\mathbf{can_3}$, caso de 4 grupos e variáveis selecionadas usando stepdisc, após a alocação dos híbridos.

Resultados da Análise Discriminante de Fisher trabalhando com 2 grupos de cada vez e com 6 e 4 variáveis (escolhidas usando stepdisc) para os *L. tenuifolius* e *L. rigidus*, respectivamente.

Visando o esclarecimento do nível discriminatório das variáveis medidas, decidiu-se mudar o cenário e trabalhar com 2 grupos de cada vez, os *L.tenuifolius* por um lado e os *L. rigidus* por outro.

Análise para o caso de dois grupos: *L. tenuifolius amarelo e azul*

A seguir vemos as variáveis selecionada pelo stepdisc, neste caso:

Stepwise Discriminant Analysis

Stepwise Selection: Summary

Step	Variable Entered	Variable Removed	Number In	Partial R**2	F Statistic	Prob > F	Wilks' Lambda	Prob < Lambda
1	X2		1	0.6082	71.420	0.0001	0.39175764	0.0001
2	X9		2	0.1416	7.424	0.0091	0.33627912	0.0001
3	X1		3	0.1651	8.698	0.0051	0.28077562	0.0001
4	X13		4	0.1734	9.020	0.0044	0.23209190	0.0001
5	X5		5	0.0865	3.979	0.0526	0.21200626	0.0001
6	X7		6	0.1355	6.425	0.0152	0.18328568	0.0001
7	X8		7	0.0589	2.503	0.1215	0.17249144	0.0001
8		X1	6	0.0000	0.001	0.9696	0.17249778	0.0001

Step	Variable Entered	Variable Removed	Number In	Average Squared Canonical Correlation	Prob > ASCC
1	X2		1	0.60824236	0.0001
2	X9		2	0.66372088	0.0001
3	X1		3	0.71922438	0.0001
4	X13		4	0.76790810	0.0001
5	X5		5	0.78799374	0.0001
6	X7		6	0.81671432	0.0001
7	X8		7	0.82750856	0.0001
8		X1	6	0.82750222	0.0001

Aplicando a Análise Discriminante de Fisher termos neste caso uma única variável discriminante de Fisher pois estamos trabalhando com apenas 2 grupos.

Construindo o gráfico discriminante, observa-se uma clara separação entre os *L.tenuifolius* "azuis" dos "amarelos".

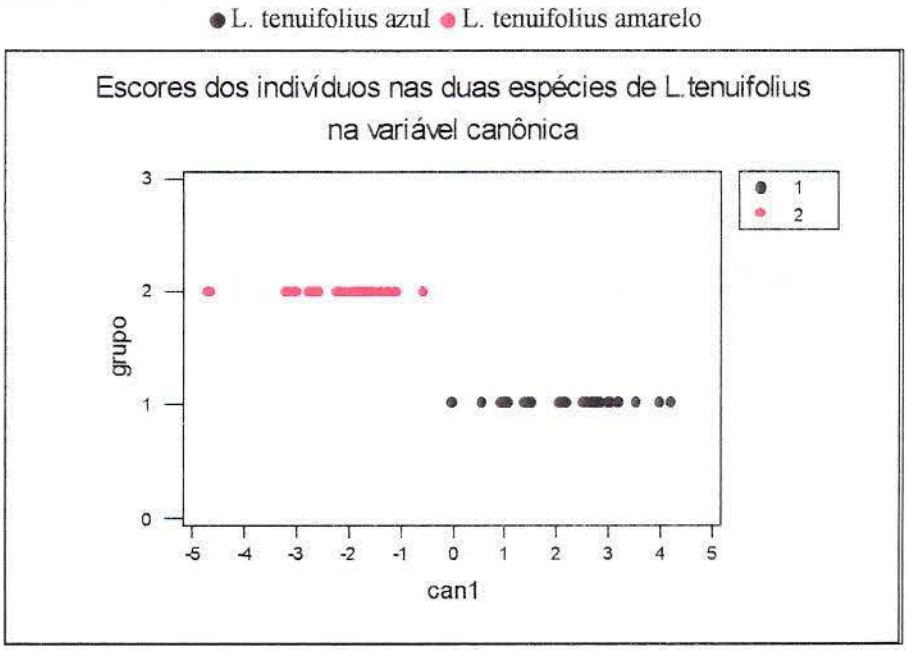


Figura 7.31: Eixo discriminante $\mathbf{can}_1$, caso de 2 grupos, *L.tenuifolius*, e variáveis selecionadas usando stepdisc.

Neste caso, por ter uma dimensão só, não conseguiremos construir o Círculo das Correlações, por tanto trabalharemos com a saída digital das correlações entre as variáveis originais e as discriminantes.

Between Canonical Structure

	CAN1
X2	1.000000
X9	1.000000
X13	1.000000
X5	1.000000
X7	1.000000
X8	1.000000

Como pode ser visto, as correlações *Entre* os grupos, são todas iguais a um, o qual é obvio pois neste caso temos somente duas médias. Além disto, as correlações são todas positivas, o qual indica que o grupo *L.tenuifolius azul*, com valores maiores para **can1** terá maiores valores nas variáveis: Escapo Largura, X_2, Folha largura, X_9,

Biomassa parte masculina, X_{13}, Nós masculinos, X_5, Comprimento parte feminina, X_7 e Folha comprimento, X_8. Pelo contrário estas variáveis no grupo *L.tenuifolius amarelo* terão magnitude menor.

As correlações positivas *Dentro* dos grupos, as quais podem ser vistas a seguir, acompanham a orientação das médias.

Pooled Within Canonical Structure

	CAN1
X2	0.568901
X9	0.065408
X13	0.138144
X5	0.180776
X7	0.281762
X8	0.430589

Colocando os híbridos no gráfico, juntos aos *tenuifolius*, com o objetivo de estudar as semelhanças com os parentais, vemos que nem todos tem as mesmas caraterísticas, uns parecem ser mais parecidos com os *tenuifolius* "azuis" e outros com os "amarelos".

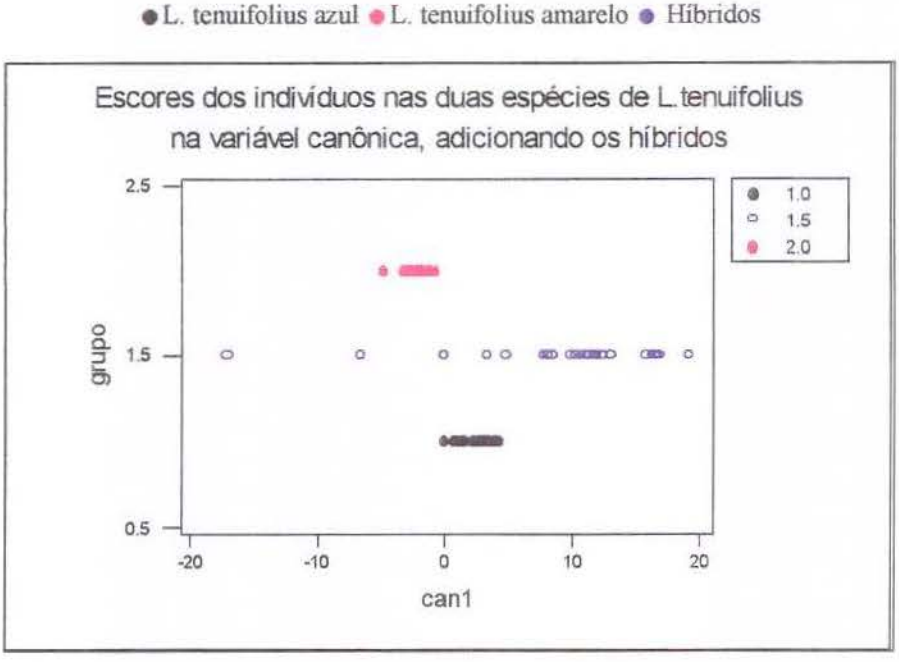


Figura 7.32: Alocação dos híbridos no eixo discriminante $\mathbf{can}_1$ para o caso de 2 grupos, *L.tenuifolius*, e variáveis selecionadas usando stepdisc.

No entanto, observa-se uma grande variabilidade destes indivíduos híbridos com os *tenuifolius*. A estratégia para classificar os indivíduos suplementares, neste caso os híbridos, será a seguinte:

⇒ Calcular as coordenadas dos hibridos na variável discriminante Can1, usadas no gráfico anterior

⇒ Calcular a distância de cada indivíduo suplementar ao centroide de cada grupo

⇒ Alocar o indivíduo no grupo do qual fique mais perto.

Na tabela a seguir mostram-se os cálculos e o grupo onde cada hibrido foi classificado:

L. tenuifolius						
Planta	**Can1**	**Centroide G1**	**Centroide G2**	**d(X0, G1)**	**d(X0, G2)**	**Alocação**
171	8.2927	2.14413	-2.14413	37.8049	108.9274	1
172	11.8284			93.7851	195.2316	1
173	-0.0382			4.7626	4.4349	2
175	12.5247			107.7562	215.1746	1
176	4.9946			8.1252	50.9615	1
177	-17.0756			369.3980	222.9488	2
178	11.2364			82.6694	179.0386	1
179	16.6130			209.3482	351.8299	1
220	11.9599			96.3493	198.9237	1
221	12.3894			104.9656	211.2235	1
222	10.3717			67.6929	156.6460	1
223	11.4012			85.6933	183.4760	1
224	15.8262			187.1990	322.9328	1
225	10.8778			76.2770	169.5707	1
226	9.9545			61.0019	146.3768	1
227	15.8028			186.5593	322.0923	1
228	9.9206			60.4735	145.5577	1
229	16.5532			207.6213	349.5901	1
233	3.3715			1.5064	30.4222	1
234	19.2784			293.5832	458.9248	1
235	-6.6191			76.7942	20.0254	2
236	13.1547			121.2327	234.0542	1
237	7.6644			30.4734	96.2073	1
238	16.5951			208.8305	351.1587	1
239	16.8537			216.3714	360.9175	1
Mirante A	11.9696			96.5399	199.1974	1
Mirante B	16.3608			202.1137	342.4324	1
Mirante C	10.8100			75.0973	167.8095	1
Mirante D	8.5644			41.2199	114.6726	1

Convertendo cada hibrido a membro do grupo no qual foi alocado, o gráfico discriminante fica da seguinte forma:

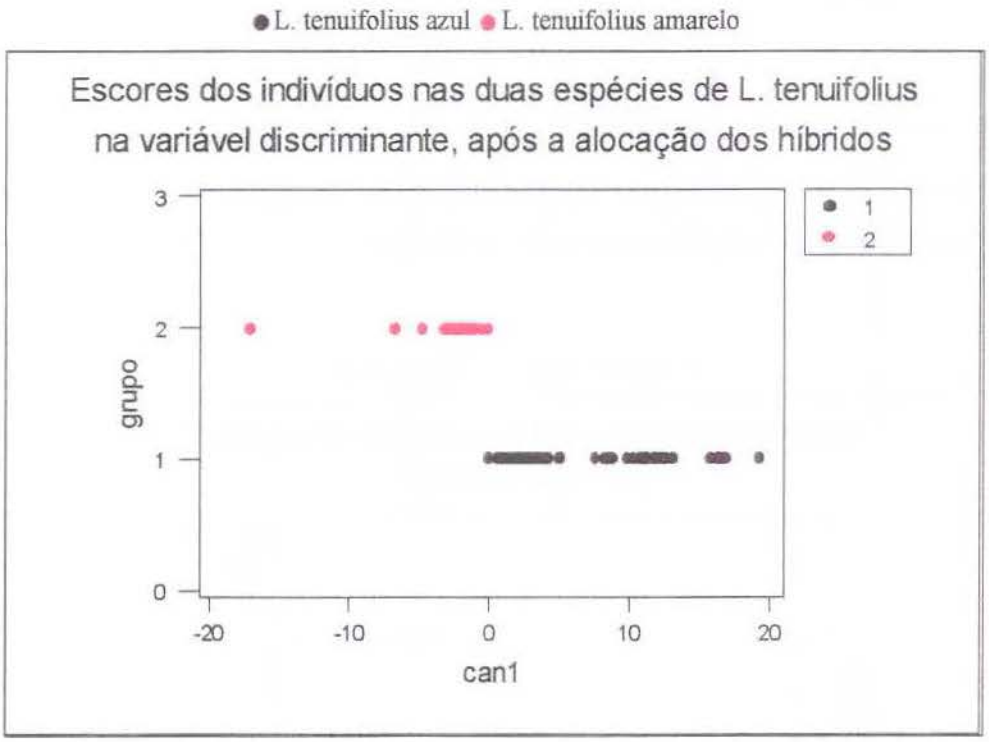


Figura 7.33: Classificação dos híbridos no eixo discriminante $\mathbf{can}_1$ para o caso de 2 grupos, *L.tenuifolius*, e variáveis selecionadas usando stepdisc.

Análise para o caso de dois grupos: *L. rigidus verde e azul*

Após a aplicação do procedimento de *stepdisc*, para reduzir o número de variáveis, quando trabalhamos somente com os grupos *L rigidus verde* e *azul*, foram selecionadas as seguintes:

Stepwise Selection: Summary

Step	Variable Entered	Variable Removed	Number In	Partial R**2	F Statistic	Prob > F	Wilks' Lambda	Prob < Lambda
1	X15		1	0.8621	343.892	0.0001	0.13788180	0.0001
2	X10		2	0.3814	33.299	0.0001	0.08528848	0.0001
3	X14		3	0.1145	6.852	0.0115	0.07552476	0.0001
4	X4		4	0.0451	2.457	0.1230	0.07211667	0.0001

Stepwise Selection: Summary

Step	Variable Entered	Removed	Number In	Average Squared Canonical Correlation	Prob > ASCC
1	X15		1	0.86211820	0.0001
2	X10		2	0.91471152	0.0001
3	X14		3	0.92447524	0.0001
4	X4		4	0.92788333	0.0001

No gráfico a seguir mostram-se as plantas dos 2 grupos com suas respectivas coordenadas na única variável discriminante de Fisher.

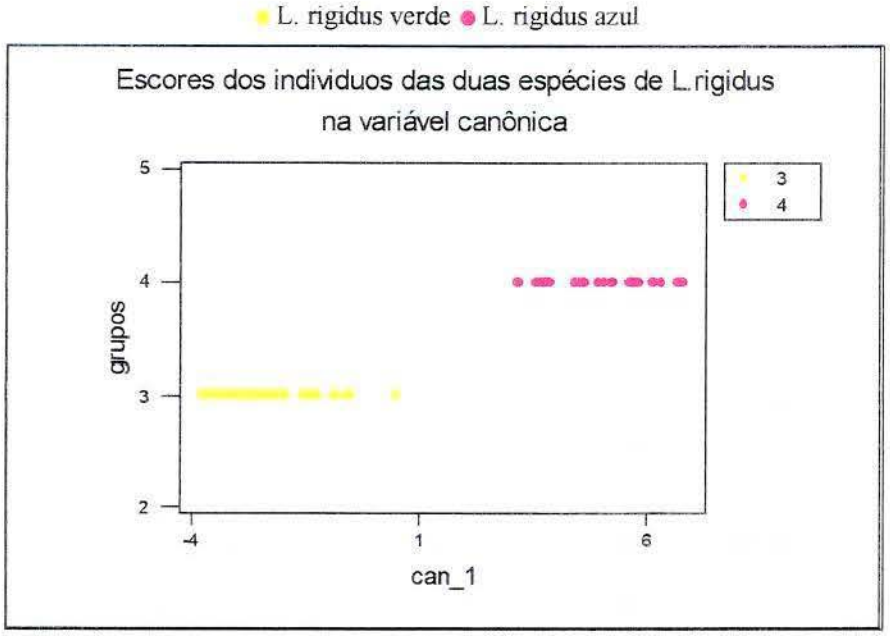


Figura 7.34: Eixo discriminante $\mathbf{can}_1$ para o caso de 2 grupos, *L.rigidus*, e variáveis selecionadas usando *stepdisc*.

A estrutura de correlação *Entre* os grupos mostra um contraste entre as variáveis X_4, X_{14}, X_{15} e X_{10}, esta última com correlação negativa. Isto indica que o grupo *L.rigidus azul* terá valores maiores de Nós estéreis, Fruto Altura, Fruto Largura e menor bainha da Folha. Por outro lado o grupo *L.rigidus verde* observa o comportamento inverso.

Between Canonical Structure

	CAN1
X4	1.000000
X10	-1.000000
X14	1.000000
X15	1.000000

A estrutura de correlação *Dentro* dos grupos mostra que X_{10} é a única das quanto variáveis selecionadas que não acompanha a direção de máxima variabilidade entre os grupos.

Pooled Within Canonical Structure

	CAN1
X4	0.063024
X10	-0.265477
X14	0.443946
X15	0.697110

Para colocar a coordenada de cada hibrido neste gráfico substituimos os valores observados, para cada um destes indivíduos, na combinação linear com o Bruto como coeficiente. Desta forma o gráfico fica assim:

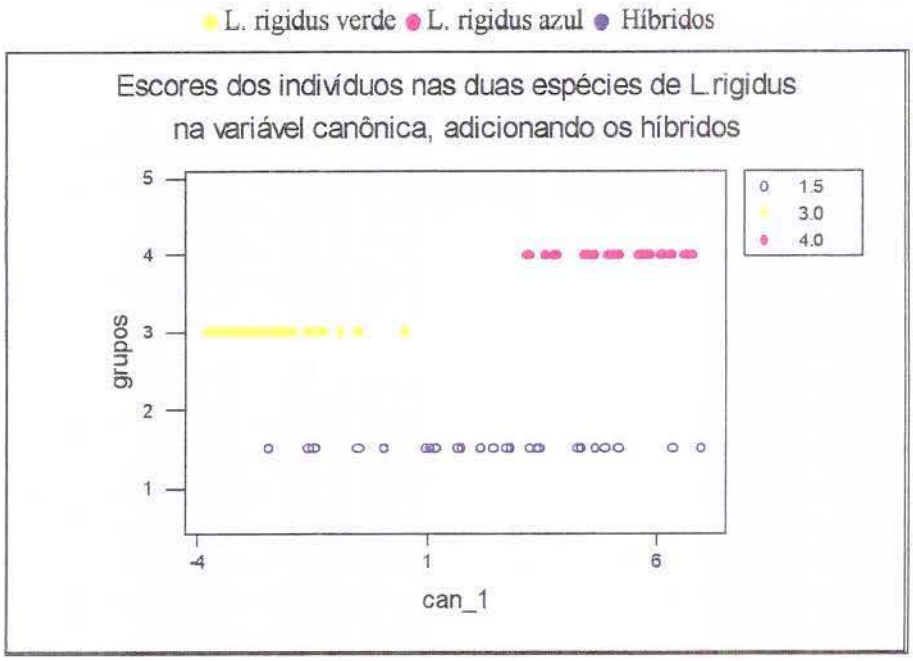


Figura 7.35: Alocação dos híbridos no eixo discriminante $\mathbf{can}_1$ para o caso de 2 grupos, *L.rigidus*, e variáveis selecionadas usando stepdisc.

A estratégia a seguir para a clasificação dos híbridos ou indivíduos suplementares em um dos dois grupos, neste caso onde temos indivíduos faltantes na variável X_{10}, será a seguinte:

1. Colocar valor zero nos lugar dos dados faltantes
2. Calcular a distância de cada um dos híbridos ao centroide dos grupos, usando as coordenadas nas variáveis canônicas. Isto quer dizer que se chamamos x_0 a um indivíduo suplementar genérico, cuja coordenada na variável discriminante de Fisher é $can\,x_0 = (can_1 x_0)$, a distância desta coordenada ao centroide de cada uma dos grupos, C_{G_j}, $j = 1,2$, pode ser calculada como:

$$d(x_0, C_{G_j}) = (can_1 x_0 - C_{G_j})^2, \text{ com } j = 1,2.$$

3. Alocar o indivíduo ao grupo com a menor distância.
4. Imputar no indivíduos a média da variável X_{10} do grupo no qual foi alocado.

Os resultados até o passo 3 anterior estão apresentados a seguir, eles são as coordenadas dos indivíduos sumplementares, centroide dos grupos e distância de cada híbrido aos 2 grupos, assim como o grupo do qual ficaram mais perto.

L. rigidus						
Planta	**Can1**	**Centroide G3**	**Centroide G4**	**d(X0, G3)**	**d(X0, G4)**	**Alocação**
171	0.0988	-2.491482	4.982964	6.7094	23.8554	**3**
172	-0.4515			4.1614	29.5336	**3**
173	-0.5131			3.9140	30.2066	**3**
175	1.1858			13.5226	14.4182	**3**
176	-1.5879			0.8165	43.1763	**3**
177	-2.4415			0.0025	55.1228	**3**
178	-1.4805			1.0221	41.7764	**3**
179	0.9966			12.1665	15.8913	**3**
220	2.7576			27.5529	4.9522	**4**
221	3.4080			34.8042	2.4804	**4**
222	3.4080			34.8042	2.4804	**4**
223	3.2457			32.9153	3.0181	**4**
224	2.7587			27.5646	4.9473	**4**
225	3.2457			32.9153	3.0181	**4**
226	2.8410			28.4348	4.5882	**4**
227	2.1905			21.9211	7.7977	**4**
228	1.7858			18.2950	10.2220	**4**

L. rigidus						
Planta	Can1	Centroide G3	Centroide G4	d(X0, G3)	d(X0, G4)	Alocação
229	1.1354			13.1539	14.8041	3
233	3.4373			35.1503	2.3891	4
234	4.2836			45.9015	0.4891	4
235	5.1921			59.0377	0.0437	4
236	4.3794			47.2085	0.3643	4
237	6.9862			89.8259	4.0128	4
238	4.7135			51.9116	0.0726	4
239	4.8832			54.3852	0.0100	4
Mirante A	6.4135			79.2982	2.0463	4
Mirante B	2.4324			24.2442	6.5056	4
Mirante C	1.6862			17.4526	10.8690	4
Mirante D	4.3087			46.2430	0.4546	4

Se juntamos os hibridos ao grupo onde foi clasificado ficará da seguinte forma:

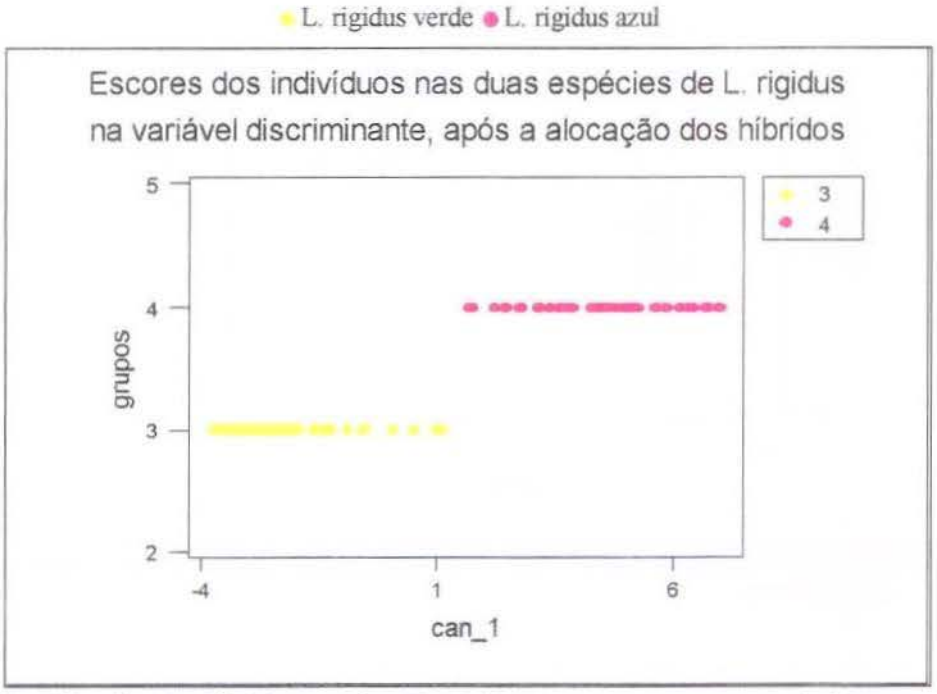


Figura 7.36: Classificação dos híbridos no eixo discriminante $\mathbf{can}_1$ para o caso de 2 grupos, *L.rigidus*, e variáveis selecionadas usando stepdisc.

Resultados da Análise Discriminante de Fisher para os dados das plantas, trabalhando com 4 grupos e todas as variáveis.

Trabalhando com todas as variáveis originais, as coordenadas dos indivíduos no plano discriminante de Fisher podem ser vistas na Figura 7.39. Como pode se observar $\mathbf{can}_1$ exibe como maior separação os *L.tenuifolius* dos *L.rigidus* mas também os *L.tenuifolius* "amarelos" dos *L.tenuifolius* "azuis" dos . Quando analisada a segunda dimensão, $\mathbf{can}_2$, é possível discriminar os *L.rigidus* "verdes" dos *L.rigidus* "azuis".

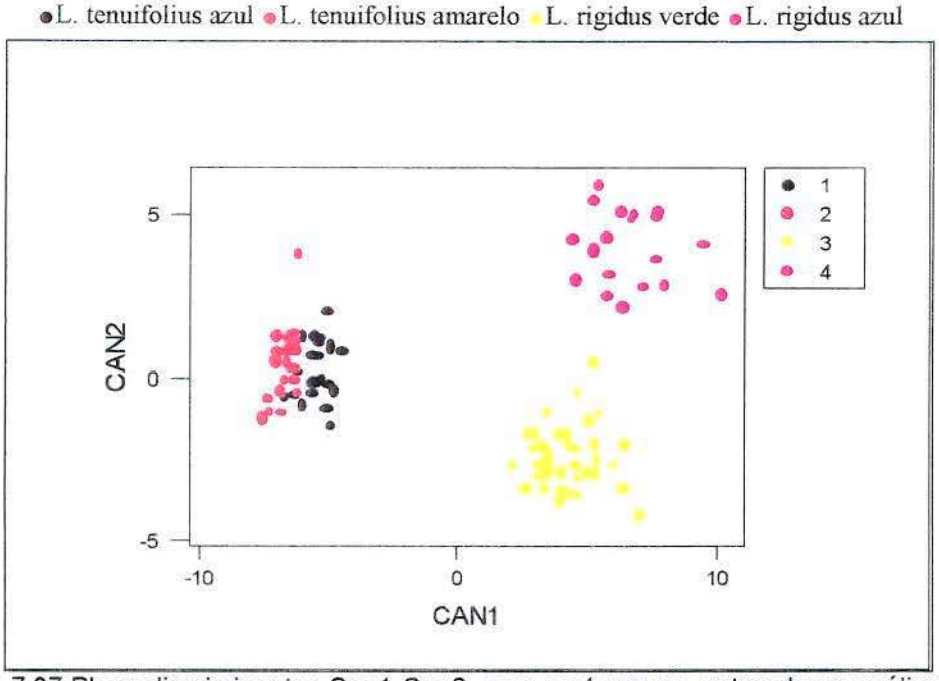


Figura 7.37 Plano discriminantes Can1-Can2, para os 4 grupos, entrando na análise com todas as variáveis.

A separação dos *L.rigidus* "verdes" dos *L.rigidus* "azuis" em $\mathbf{can}_2$ pode ser vista desde outro ângulo na Figura 7.38.

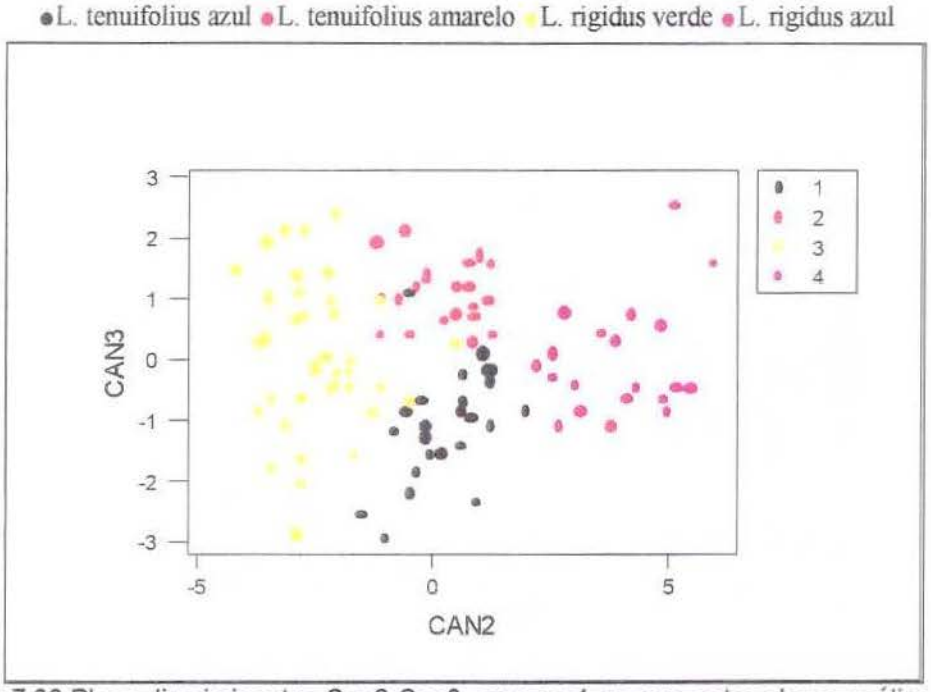


Figura 7.38 Plano discriminantes Can2-Can3, para os 4 grupos, entrando na análise com todas as variáveis.

A terceira dimensão discriminante mostra sua utilidade na Figura 7.38, aqui é visto como ela consegue separar os *L.tenuifolius* "azuis" dos "amarelos".

Os resultados para as diferentes estruturas de correlação são mostrados a seguir nos Círculos das Correlações, Figura 7.40. Observam-se altas correlações positivas *Entre* os grupos, para as variáveis X_4, X_9, X_2, X_1, X_6, X_{11}, X_8, X_7, X_5, X_{12} e X_{13} com **can1**, indicando que estas variáveis: Nós estéreis, Largura folha, Largura escapo, Comprimento escapo, Parte masculina, Biomassa "caule", Comprimento folha, Comprimento parte feminina, Nós masculinos, Biomassa parte feminina e Biomassa parte masculina são as mais fortemente correlacionadas com a direção de máxima separação entre os *L.tenuifolius* e os *L.rigidus*. Isto quer dizer que estas características das plantas aumentarão na direção de aumento de **can1**.

Repetindo a Figura 7.38 para dar uma melhor visualização de **can3** obtemos a Figura 7.39, mostrada a seguir:

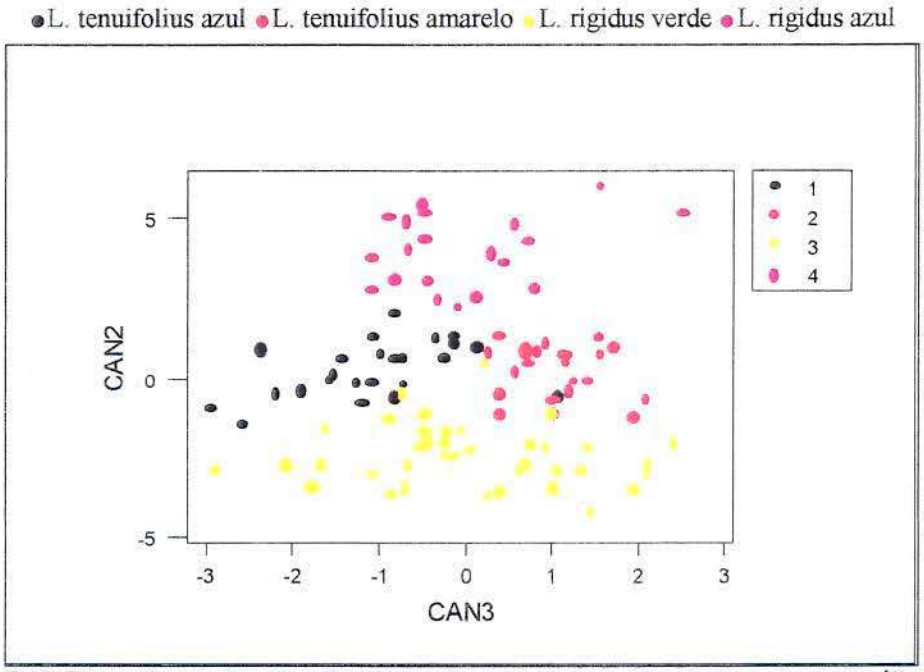


Figura 7.39 Plano discriminantes Can3-Can2, para os 4 grupos, entrando na análise com todas as variáveis.

Voltando ao plano discriminante, figura 7.37, vemos que os indivíduos pertencentes aos dois grupos do tipo *L.rigidus* apresentam os maiores valores na variável $\mathbf{can}_1$, portanto eles apresentaram maior número de Nós estéreis, Largura folha, Largura escapo, Comprimento escapo, Parte masculina, Biomassa "caule" e Comprimento folha. Sendo que, para os indivíduos pertencentes aos grupos do tipo *L.tenuifolius,* os valores destas características serão menores respeito aos valores para *L.rigidus.*

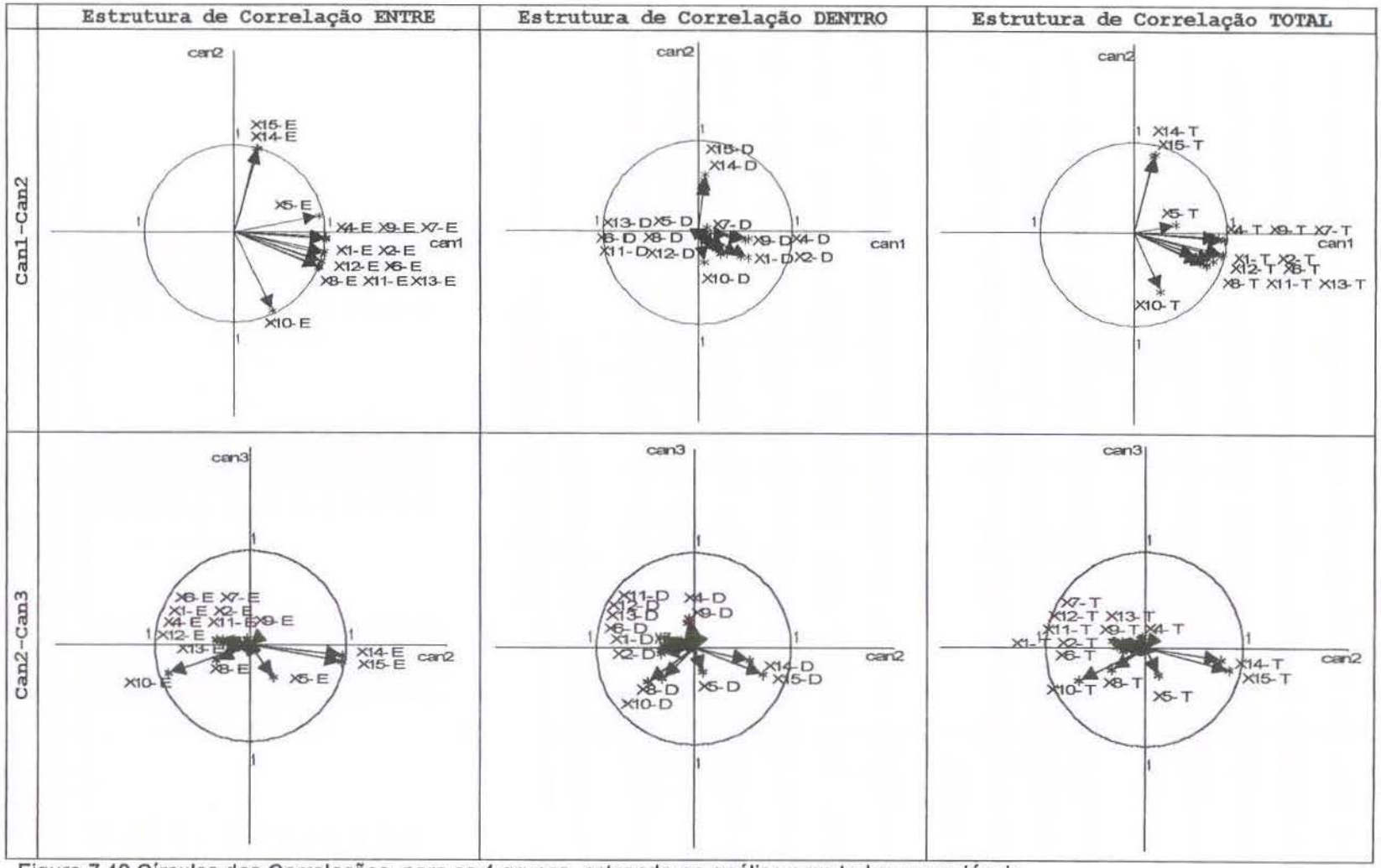


Figura 7.40 Círculos das Correlações, para os 4 grupos, entrando na análise com todas as variáveis.

Com $\mathbf{can}_2$, por outro lado, três variáveis apresentam altas correlações, positivas com Fruto Largura, X_{14} e Fruto Altura, X_{15} e negativa no caso da Bainha da Folha, X_{10}. Portanto, os indivíduos com valores maiores de $\mathbf{can}_2$ terão maior largura e altura do fruto e menor Bainha da Folha, entanto que os de coordenadas menores apresentaras frutos menores en Largura e Altura e Bainha da Folha maior.

Voltando nos planos discriminantes, Figura 7.38, vemos que serão os *L.rigidus* "azuis" os de frutos maiores em largura e altura e Bainha da Folha menor. Não sendo assim com os *L.rigidus* "verdes" que terão frutos menores em largura e altura e bainha da Folha maior.

Com respeito a $\mathbf{can}_3$, se destacam as variáveis Bainha da Folha, X_{10} e Nós masculinos, X_5, ambas com correlações negativas com a terceira variável discriminante, indicando que os indivíduos com valores menores em $\mathbf{can}_3$, que neste caso são os pertencentes ao grupo *L.tenuifolius azul* (ver Figura 7.39), apresentam Bainha da Folha e número de Nós masculinos maior que os *L.tenuifolius* "amarelos", os quais, segundo esta análise tem Bainha da Folha e número de Nós masculinos menor.

Os valores numéricos destas correlações são:

Between Canonical Structure

	CAN1	CAN2	CAN3
X1	0.971090	-0.238440	-0.011459
X2	0.977800	-0.209536	-0.001097
X4	0.995526	-0.060478	0.072595
X5	0.918353	0.192983	-0.345522
X6	0.951462	-0.306938	0.022571
X7	0.998840	-0.047235	-0.009370
X8	0.911416	-0.384035	-0.147776
X9	0.995149	-0.073780	0.065075
X10	0.418518	-0.867272	-0.269595
X11	0.927056	-0.368858	0.067160
X12	0.944772	-0.320619	0.067890
X13	0.922747	-0.379336	0.068136
X14	0.249878	0.967850	-0.028767
X15	0.266327	0.954856	-0.131605

Quanto às correlações *Dentro* dos grupos, olhando nos Círculos da Figura 7.41, podemos dizer que nenhuma variável acompanha fortemente a direção de $\mathbf{can}_1$. Já no

caso de $\mathbf{can}_2$ as variáveis Fruto Largura, X_{14} e Fruto Altura, X_{15} apresentam as correlações *Dentro* mais destacadas e acompanhando a direção de máxima separação dos *L.rigidus* "verdes" e "azuis".

Analisando agora $\mathbf{can}_3$, podemos observar que esta direção de máxima separação dos *L.tenuifolius* "azuis" dos "amarelos" não é acompanhada com grande intensidade por nenhuma das variáveis, somente as variáveis X_5 e X_{10} se destacam do resto, com correlações negativas com $\mathbf{can}_3$.

Os valores numéricos destas correlações podem ser visto a seguir:

Pooled Within Canonical Structure

	CAN1	CAN2	CAN3
X1	0.432960	-0.268251	-0.042062
X2	0.517064	-0.279594	-0.004775
X4	0.520974	-0.079860	0.312776
X5	0.086392	0.045810	-0.267611
X6	0.289888	-0.235974	0.056619
X7	0.331481	-0.039555	-0.025602
X8	0.229823	-0.244355	-0.306793
X9	0.489521	-0.091579	0.263548
X10	0.062863	-0.328705	-0.333390
X11	0.231098	-0.232019	0.137838
X12	0.155904	-0.133504	0.092236
X13	0.183401	-0.190247	0.111497
X14	0.058894	0.575607	-0.055823
X15	0.070064	0.633855	-0.285046

Total Canonical Structure

	CAN1	CAN2	CAN3
X1	0.917886	-0.242617	-0.018686
X2	0.942531	-0.217429	-0.001824
X4	0.958624	-0.062691	0.120602
X5	0.442121	0.100014	-0.286984
X6	0.837969	-0.291004	0.034296
X7	0.897303	-0.045679	-0.014523
X8	0.760612	-0.345008	-0.212765
X9	0.951997	-0.075980	0.107401
X10	0.277038	-0.618005	-0.307884
X11	0.770768	-0.330133	0.096335
X12	0.658700	-0.240636	0.081661
X13	0.703676	-0.311406	0.089644
X14	0.203810	0.849801	-0.040481
X15	0.225246	0.869345	-0.192028

A estrutura de correlação total dá uma idéia da nuvem de indivíduos sem estrutura de grupo. Os valores estão a seguir e as representações gráficas podem ser vistas na Figura 7.40.

Outros resultados para a saída da ADF no SAS para o caso de 4 grupos, com todas as variáveis podem ser vistos a seguir.

Canonical Discriminant Analysis

105	Observations	104	DF Total
14	Variables	101	DF Within Classes
4	Classes	3	DF Between Classes

Class Level Information

GRUPOS	Frequency	Weight	Proportion
1	24	24.0000	0.228571
2	24	24.0000	0.228571
3	38	38.0000	0.361905
4	19	19.0000	0.180952

Eigenvalues of INV(E)*H

	Eigenvalue	Difference	Proportion	Cumulative
1	32.7897	27.6399	0.8534	0.8534
2	5.1498	4.6661	0.1340	0.9874
3	0.4837	.	0.0126	1.0000

Total-Sample Standardized Canonical Coefficients

	CAN1	CAN2	CAN3
X1	2.972652643	-1.706167771	0.018820115
X2	3.468070591	-1.125442112	-3.076180233
X4	1.642950749	0.398988789	0.885599415
X5	-0.088428487	0.380667038	-0.184923698
X6	-0.957858676	0.849433383	-0.292983305
X7	0.378730351	0.746260170	-0.828860131
X8	-0.110059650	-0.483503085	-1.560716436
X9	0.784850138	0.548649601	2.541083991
X10	-0.220527766	-0.272822591	-0.185559716
X11	-2.215205300	0.264038238	1.909717382
X12	-0.523687844	-0.009313345	0.306379614
X13	0.025855370	-0.055211788	0.480383286
X14	0.236741793	0.998646594	0.329325134
X15	0.184721073	1.125867708	-0.689642064

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2	CAN3
X1	1.100142908	-0.631432124	0.006965098
X2	1.103578181	-0.358128050	-0.978874361
X4	0.527741101	0.128161348	0.284468180
X5	-0.078999295	0.340076243	-0.165205153
X6	-0.483351202	0.428638020	-0.147844182
X7	0.178967792	0.352642809	-0.391675152
X8	-0.063586009	-0.279339720	-0.901690402
X9	0.266449360	0.186261463	0.862674504
X10	-0.169658056	-0.209889898	-0.142756176
X11	-1.289748129	0.153729690	1.111885395
X12	-0.386247192	-0.006869079	0.225971000
X13	0.017317498	-0.036979939	0.321752755
X14	0.143018401	0.603293727	0.198949046
X15	0.103667909	0.631851849	-0.387036248

Raw Canonical Coefficients

	CAN1	CAN2	CAN3
X1	0.040787769	-0.023410329	0.000258231
X2	1.760490374	-0.571306135	-1.561555784
X4	0.753039446	0.182874804	0.405910703
X5	-0.092204084	0.396920235	-0.192819315
X6	-0.020025437	0.017758648	-0.006125245
X7	0.023739482	0.046776895	-0.051954406
X8	-0.004295560	-0.018870828	-0.060913801
X9	0.167078463	0.116796223	0.540944552
X10	-0.113638128	-0.140585691	-0.095619065
X11	-0.170975465	0.020379177	0.147397091
X12	-0.243191295	-0.004324951	0.142277229
X13	0.011827491	-0.025256529	0.219750436
X14	0.660259329	2.785168264	0.918468973
X15	0.692585404	4.221280926	-2.585714883

Class Means on Canonical Variables

GRUPOS	CAN1	CAN2	CAN3
1	-5.456264378	0.239386368	-1.061052594
2	-6.636745830	0.418271899	0.950882908
3	4.352307905	-2.392398408	0.057654417
4	6.570765505	3.954070584	0.023852874

Comparando os resultados obtidos para quatro grupos com as 9 variáveis selecionadas usando o *stepdisc* e sem usar *stepdisc*, com todas as 14 variáveis.

O aspecto mais importante a ser tomado em consideração aqui é o destaque da variável X_5 em alguns itens da análise. O comportamento em geral é muito parecido com o analisado quando usamos as 9 variáveis obtidas trabalhando com o *stepdisc*.

Os Círculos das Correlações dados a seguir permitirão comparar as semelhanças e diferenças da mudança no número de variáveis.

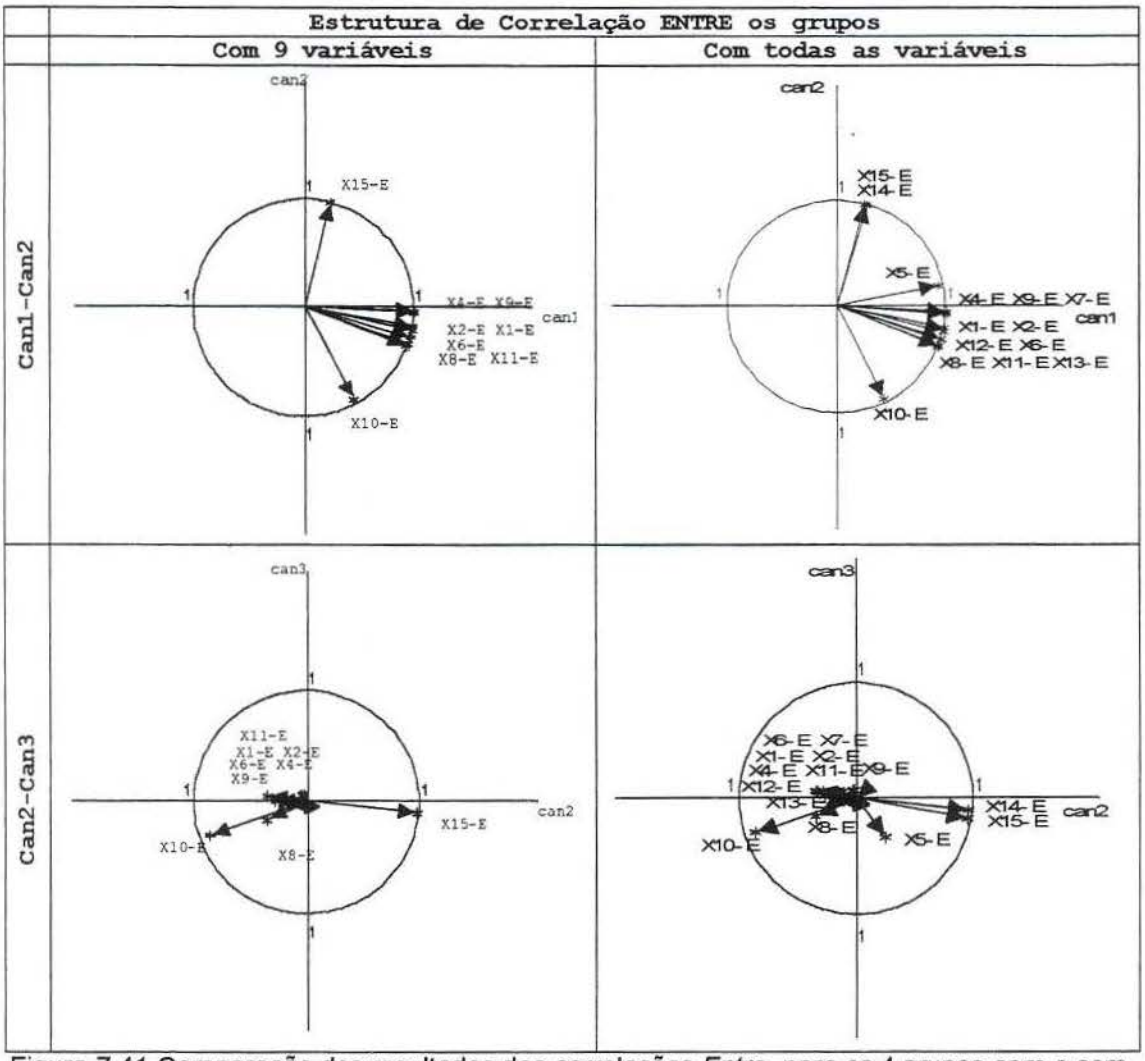


Figura 7.41 Comparação dos resultados das correlações *Entre*, para os 4 grupos com e sem *stepdisc*.

Portanto, a redução de 14 para 9 variáveis manteve o mesmo perfil de informação, com menor redundância. Isto não implica que o *stepdisc* deve ser adotado pois o conhecimento substantivo pode considerar como necessário o uso de todas as variáveis.

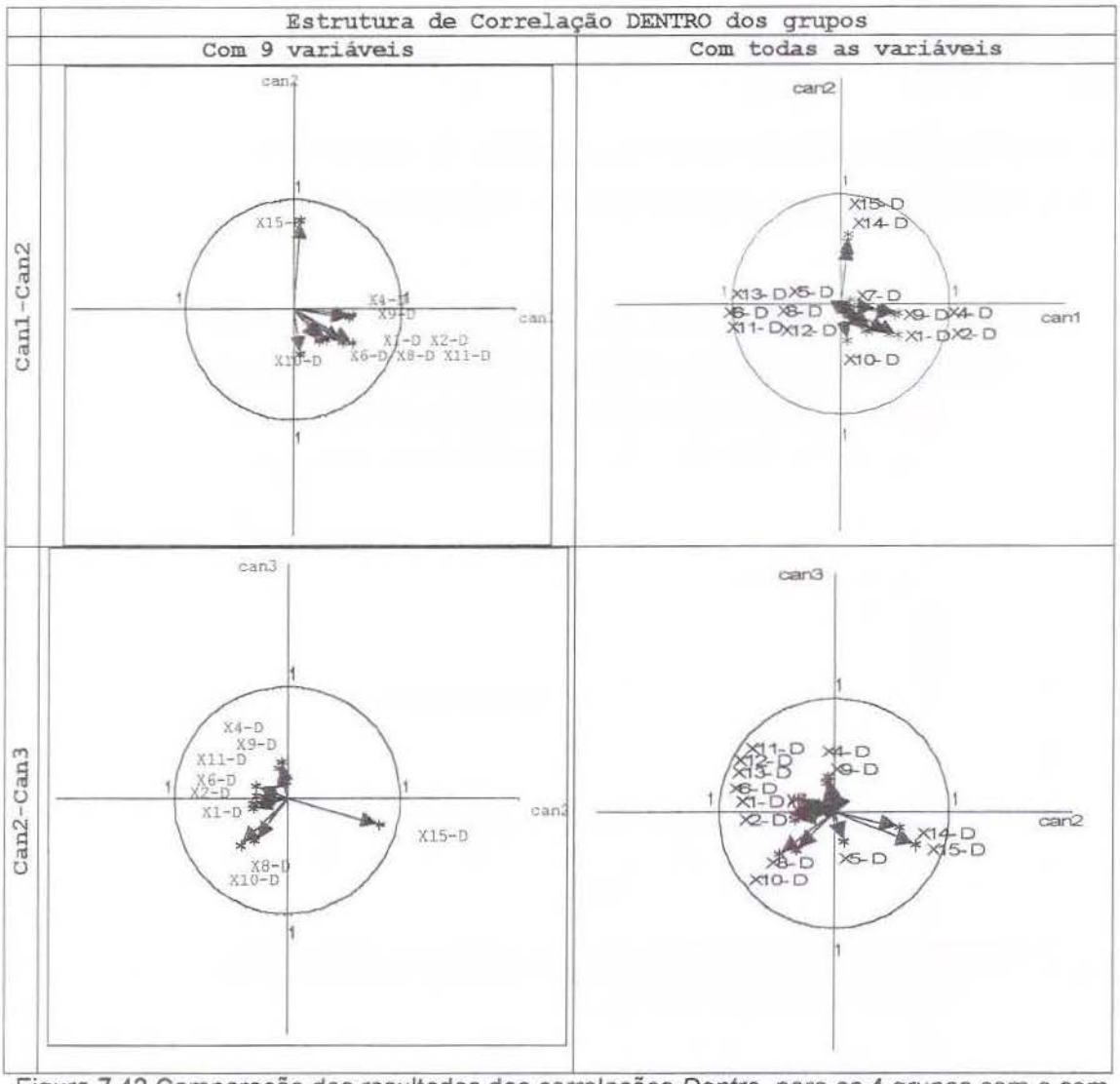


Figura 7.42 Comparação dos resultados das correlações *Dentro*, para os 4 grupos com e sem *stepdisc*.

Aqui também os padrões mantém-se semelhantes. É interessante notar que neste caso na presença de correlações altas e positivas dentro dos grupos a representação das variáveis próximas no centro do círculo indica perda destas direções no plano discriminante. Ou seja, descrever diferenças entre plantas dentro de cada grupo seria melhor, por exemplo, fazendo Componentes Principais dentro de cada grupo.

Resultados da Análise Discriminante de Fisher trabalhando com 2 grupos de cada vez, *L. tenuifolius* e *L. rigidus*, e com todas as variáveis.

Análise para o caso de dois grupos: *L. tenuifolius amarelo e azul* com todas as variáveis

Trabalhando com todas as variáveis, o gráfico da única variável discriminante de Fisher para o caso de 2 grupos, para este caso *L. tenuifolius amarelo e azul*, fica:

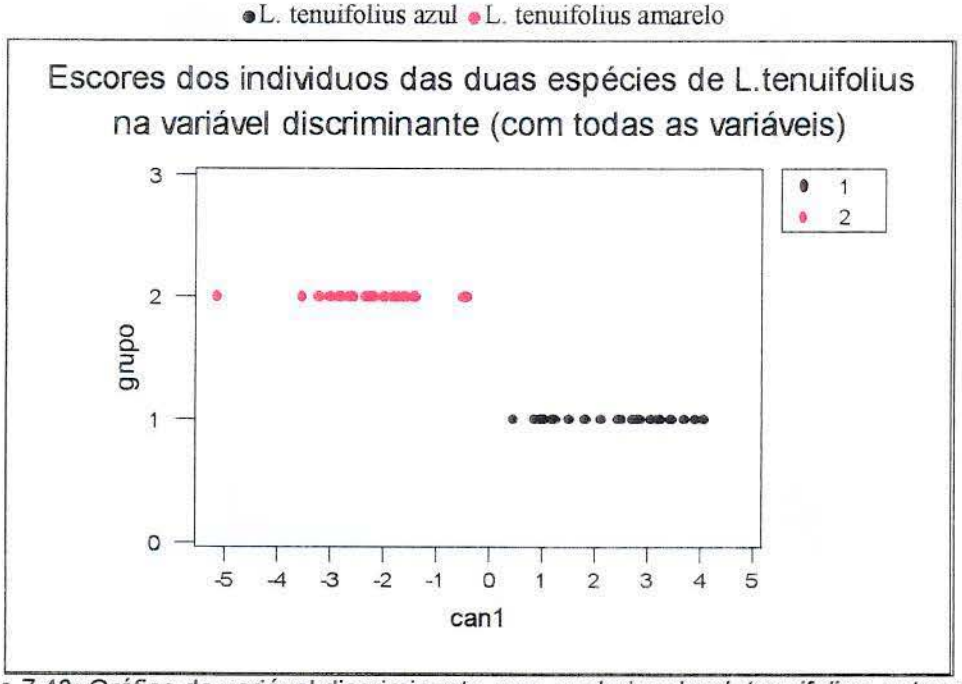


Figura 7.43: Gráfico da variável discriminante para os dados dos *L.tenuifolius*, entrando na análise com todas as variáveis.

O aspecto deste gráfico respeito ao que foi feito usando só as variáveis escolhidas pelo *stepdisc* é basicamente o mesmo.

Quanto as correlações *Entre* grupos são todas positivas, sendo zero no caso particular de X_4, pois não foram observados nós estéreis nos *L.tenuifolius*. Isto indica que, observando no gráfico discriminante, os *L.tenuifolius* "azuis" apresentam valores

maiores que es *L.tenuifolius* "amarelos" para todas as características medidas, a exceção de X_4, pois os *tenuifolius* não exibiram presença de Nós estéreis.

Between Canonical Structure

	CAN1
X1	1.000000
X2	1.000000
X4	0.000000
X5	1.000000
X6	1.000000
X7	1.000000
X8	1.000000
X9	1.000000
X10	1.000000
X11	1.000000
X12	1.000000
X13	1.000000
X14	1.000000
X15	1.000000

Quanto à estrutura de correlação *Dentro* dos grupos, somente se destaca a variável X_2 indo na direção da separação dos grupos, o resto é indiferente à direção de separação dos grupos.

Pooled Within Canonical Structure

	CAN1
X1	0.359387
X2	0.517802
X4	0.000000
X5	0.164539
X6	0.179903
X7	0.256453
X8	0.391913
X9	0.059533
X10	0.165841
X11	0.412197
X12	0.188382
X13	0.125735
X14	0.020933
X15	0.096572

Outras saídas da Análise Discriminante de Fisher neste caso podem ser vistas a seguir:

Canonical Discriminant Analysis

48 Observations	47 DF Total
14 Variables	46 DF Within Classes
2 Classes	1 DF Between Classes

Class Level Information

GRUPOS	Frequency	Weight	Proportion
1	24	24.0000	0.500000
2	24	24.0000	0.500000

Multivariate Statistics and Exact F Statistics

S=1 M=5.5 N=16

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.14725981	15.1450	13	34	0.0001
Pillai's Trace	0.85274019	15.1450	13	34	0.0001
Hotelling-Lawley Trace	5.79071920	15.1450	13	34	0.0001
Roy's Greatest Root	5.79071920	15.1450	13	34	0.0001

Eigenvalues of INV(E)*H

	Eigenvalue	Difference	Proportion	Cumulative
1	5.7907	.	1.0000	1.0000

Total Canonical Structure

	CAN1
X1	0.708368
X2	0.844559
X4	0.000000
X5	0.398659
X6	0.430225
X7	0.568714
X8	0.742988
X9	0.153570
X10	0.401382
X11	0.762614
X12	0.447108
X13	0.313613
X14	0.054480
X15	0.245125

Total-Sample Standardized Canonical Coefficients

	CAN1
X1	-1.827462501
X2	1.836420355
X4	0.000000000
X5	1.687852037
X6	1.522564303
X7	1.592549996
X8	1.098788562
X9	-0.708484097
X10	-0.706124669
X11	0.115612182
X12	-0.228733569
X13	-0.889441078
X14	-0.029833309
X15	-0.169461604

Pooled Within-Class Standardized Canonical Coefficients

	CAN1
X1	-1.397195460
X2	1.161852100
X4	0.000000000
X5	1.586282521
X6	1.412355099
X7	1.369905009
X8	0.808014032
X9	-0.708905914
X10	-0.662919007
X11	0.082969012
X12	-0.210579711
X13	-0.860529688
X14	-0.030117654
X15	-0.166847613

Raw Canonical Coefficients

	CAN1
X1	-0.13625801
X2	6.74868947
X4	0.00000000
X5	2.27023436
X6	0.18334949
X7	0.48802304
X8	0.10811443
X9	-1.94781861
X10	-0.43164282
X11	0.57743912
X12	-5.26479084
X13	-13.28401105
X14	-0.17387307
X15	-0.93433312

Class Means on Canonical Variables

GRUPOS	CAN1
1	2.355724778
2	-2.355724778

Análise para o caso de dois grupos: *L. rigidus azul* e *verde* com todas as variáveis

O gráfico discriminante não se diferencia praticamente do obtido com as variáveis selecionadas pelo *stepdisc*. Embora os grupos se observam um pouco mais separados, no caso com menos variáveis já tinha se obtido uma clara separação.

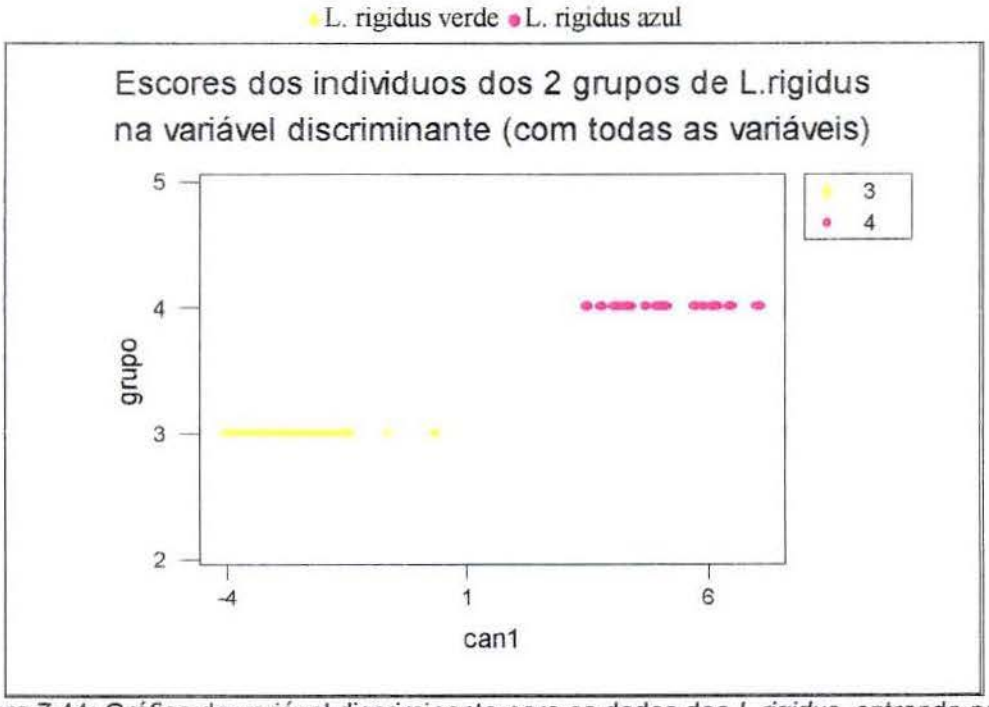


Figura 7.44: Gráfico da variável discriminante para os dados dos *L.rigidus*, entrando na análise com todas as variáveis.

Da estrutura de correlação *Entre* podemos dizer que os indivíduos do grupo *L.rigidus azul*, com maiores valores em $\mathbf{can}_1$, apresentarão maiores valores quanto a: número de Nós estéreis, X_4, Nós masculinos, X_5, Comprimento da parte feminina, X_7, Folha largura, X_9, Fruto altura, X_{14} e Fruto largura, X_{15}; e menores valores para

Escapo comprimento, X_1, Escapo largura, X_2, Parte masculina, X_6, Folha comprimento, X_8, Bainha da Folha, X_{10}, Biomassa "caule", X_{11}, Biomassa parte feminina, X_{12} e Biomassa parte masculina, X_{13}.

Pelo contrário, os indivíduos do grupo *L.rigidus verde*, com menores valores em $\mathbf{can}_1$, apresentarão maiores valores quanto a Escapo comprimento, X_1, Escapo largura, X_2, Parte masculina, X_6, Folha comprimento, X_8, Bainha da Folha, X_{10}, Biomassa "caule", X_{11}, Biomassa parte feminina, X_{12} e Biomassa parte masculina, X_{13} e menores para número de Nós estéreis, X_4, Nós masculinos, X_5, Comprimento da parte feminina, X_7, Folha largura, X_9, Fruto altura, X_{14} e Fruto largura, X_{15}.

Os valores de correlação Entre os grupos estão a seguir:

Between Canonical Structure

	CAN1
X1	-1.000000
X2	-1.000000
X4	1.000000
X5	1.000000
X6	-1.000000
X7	1.000000
X8	-1.000000
X9	1.000000
X10	-1.000000
X11	-1.000000
X12	-1.000000
X13	-1.000000
X14	1.000000
X15	1.000000

A estrutura de correlação *Dentro*, por outro lado, evidencia que a Largura do Fruto, X_{15}, acompanha à direção de máxima separação entre os *L.rigidus* "azuis" e "verdes".

Pooled Within Canonical Structure

	CAN1
X1	-0.070932
X2	-0.059089
X4	0.059099
X5	0.054492
X6	-0.081192
X7	0.046410
X8	-0.101483
X9	0.046462
X10	-0.248944
X11	-0.089299
X12	-0.046716
X13	-0.074495
X14	0.416299
X15	0.653696

Outros aspectos da saída da Análise Discriminante de Fisher estão resumidos a seguir:

Canonical Discriminant Analysis

57 Observations	56 DF Total
14 Variables	55 DF Within Classes
2 Classes	1 DF Between Classes

Class Level Information

GRUPOS	Frequency	Weight	Proportion
3	38	38.0000	0.666667
4	19	19.0000	0.333333

Multivariate Statistics and Exact F Statistics

S=1 M=6 N=20

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.06397081	43.8964	14	42	0.0001
Pillai's Trace	0.93602919	43.8964	14	42	0.0001
Hotelling-Lawley Trace	14.63212983	43.8964	14	42	0.0001
Roy's Greatest Root	14.63212983	43.8964	14	42	0.0001

Eigenvalues of INV(E)*H

Eigenvalue	Difference	Proportion	Cumulative
14.6321	.	1.0000	1.0000

Total Canonical Structure

	CAN1
X1	-0.270661
X2	-0.227874
X4	0.227911
X5	0.210915
X6	-0.306567
X7	0.180667
X8	-0.374043
X9	0.180864
X10	-0.712785
X11	-0.334111
X12	-0.181823
X13	-0.283259
X14	0.875325
X15	0.959707

Total-Sample Standardized Canonical Coefficients

	CAN1
X1	-0.248671115
X2	0.364999632
X4	0.307592279
X5	-0.054308379
X6	-0.077748326
X7	0.483606342
X8	0.193446139
X9	0.006922010
X10	-1.030720081
X11	-0.271405689
X12	-0.342815216
X13	0.016000551
X14	0.587665284
X15	2.540199432

Pooled Within-Class Standardized Canonical Coefficients

	CAN1
X1	-.2421658466
X2	0.3592407462
X4	0.3027366516
X5	-.0536468159
X6	-.0749217573
X7	0.4804705731
X8	0.1819671658
X9	0.0068768904
X10	-.7531837287
X11	-.2591594765
X12	-.3405234431
X13	0.0155272433
X14	0.3153535523
X15	0.9517742760

Raw Canonical Coefficients

	CAN1
X1	-0.006854885
X2	0.430355664
X4	0.318773689
X5	-0.055340917
X6	-0.002345643
X7	0.048590045
X8	0.009648722
X9	0.003220561
X10	-0.517235013
X11	-0.025354352
X12	-0.158238842
X13	0.007829178
X14	1.276124730
X15	7.911446619

Class Means on Canonical Variables

GRUPOS	CAN1
3	-2.656945781
4	5.313891563

Resumo das conclusões e interpretação para o caso de 4 grupos

Fonte	**Considerando as variáveis escolhidas pelo *stepdisc***	**Considerando todas as variáveis**
can_1	Em geral as plantas do tipo *L.tenuifolius* se diferenciam facilmente das do tipo *L.rigidus* principalmente pelas seguintes características: número de Nós estéreis, Largura folha, Largura escapo, Comprimento escapo, Parte masculina e Biomassa "caule" e Folha Comprimento.	Em geral os indivíduos do tipo *L.rigidus* apresentaram maior número de Nós estéreis, Largura folha, Largura escapo, Comprimento escapo, Parte masculina, Biomassa "caule" e Comprimento folha, respeito aos indivíduos tipo *L.tenuifolius* para os quais se observarão valores menores nestas características.
can_2	Os *L.rigidus* "verdes" podem ser diferenciados dos *L.rigidus* "azuis" observando principalmente o contraste entre a Largura do Fruto e a Bainha da Folha, i.é, os *L.rigidus* "azuis" terão maior Largura do Fruto e menor Bainha da Folha enquanto os *L.rigidus* "verdes" terão menor Largura do Fruto e maior Bainha da Folha.	Os *L.rigidus* "azuis" serão os de frutos maiores em largura e altura e Bainha da Folha menor, portanto os *L.rigidus* "verdes" terão frutos menores em largura e altura e Bainha da Folha maior.

Fonte	Considerando as variáveis escolhidas pelo *stepdisc*	Considerando todas as variáveis
can$_3$	Os *L. tenuifolius* "azuis" e os *L.tenuifolius* "amarelos" são os mais difíceis de serem distinguidos. Apesar de existir diferenças entre eles, esta só poderá é captada na terceira variável discriminante. Os indivíduos do grupo *L.tenuifolius azul* apresentarão valores maiores para a Bainha da Folha, X_{10}, Folha Comprimento, X_8, e a Largura do Fruto, X_{15}, sendo menores no grupo *L.tenuifolius verde.*	Os indivíduos pertencentes ao grupo *L.tenuifolius azul* apresentam Bainha da Folha maior e também maior número de Nós masculinos, respeito aos *L.tenuifolius* "amarelos" que tem Bainha da Folha e número de Nós masculinos menor.

Conclusões e interpretação para o caso de 2 grupos: *L. tenuifolius amarelo* e *azul*

Fonte	Considerando as variáveis escolhidas pelo *stepdisc*	Considerando todas as variáveis
can$_1$	Os indivíduos do grupo *L.tenuifolius azul* terão maiores valores nas variáveis: Escapo Largura, X_2, Folha largura, X_9, Biomassa parte masculina, X_{13}, Nós masculinos, X_5, Comprimento parte feminina, X_7 e Folha comprimento, X_8, ao contrário do grupo *L.tenuifolius amarelo* com magnitude menor nestas características.	Os *L.tenuifolius* "azuis" apresentam valores maiores para todas as características medidas comparados com os *L.tenuifolius* "amarelos". A exceção de X_4, pois nos *tenuifolius* não se observaram Nós estéreis.

Conclusões e interpretação para o caso de 2 grupos: *L. rigidus verde* e *azul*

Fonte	Considerando as variáveis escolhidas pelo *stepdisc*	Considerando todas as variáveis
can1	O grupo *L.rigidus azul* terá valores maiores de Nós estéreis, Fruto Altura, Fruto Largura e menores para a Bainha da Folha. Os *L.rigidus verde* terão maior Bainha da Folha e menor quantidade de Nós estéreis e menor Fruto Altura e Fruto Largura	O grupo *L.rigidus azul* apresenta valores grandes quanto ao número de Nós estéreis e masculinos, Comprimento da parte feminina, Folha largura, Fruto altura, e Fruto largura e valores pequenos para Escapo comprimento Escapo largura, Parte masculina, Folha comprimento, Bainha da Folha, Biomassa "caule", Biomassa parte feminina e Biomassa parte masculina. Isto relativo aos indivíduos do grupos *L.rigidus verde* que apresentaram o comportamento contrário a este.

Capítulo 8: Roteiro para a interpretação da Análise Discriminante de Fisher

Após a análise dos resultados deste trabalho propõe-se o seguinte roteiro para auxiliar na interpretação dos resultados na Análise Discriminante de Fisher:

1. Faça análise exploratória univariada dos seus dados visando examinar a separação marginal dos grupos.

2. Faça análise exploratória bivariada dos seus dados visando a separação duas a duas das variáveis.

3. Desenvolva a Análise Discriminante de Fisher a seus dados, não sendo necessário padronizar.

4. Construa os gráficos discriminantes calculando as coordenadas dos indivíduos em cada variável discriminante usando o coeficiente *Bruto* e as variáveis originais centradas, para uma melhor visualização.

5. Calcule os coeficientes de correlação *Entre, Dentro* e *Total* para as correlações entre originais e discriminantes, usando as expressões (2.4.6), (2.4.4) e (2.4.2), respectivamente.

6. Construa os Círculos das Correlações para visualizar melhor as estruturas de correlação (programa disponível no anexo).

7. Observe, a partir das correlações das variáveis originais com as variáveis discriminantes, as correlações mais destacadas em cada caso, conforme 3 acima, e analise o sinal, a magnitude e a qualidade da representação (ver as conclusões do capítulo 6, página).

8. Volte nos gráficos discriminantes e procure a posição dos grupos com respeito aos valores das variáveis discriminantes, utilizando a interpretação do passo 5.

9. Usando o passo anterior relacione o aumento ou diminuição do valor das variáveis com cada grupo em particular.

10. Finalmente volte na análise descritiva inicial e veja se suas interpretações são consistentes com o padrão observado.

Capítulo 9: Conclusões

Nas técnicas multivariadas a comparação entre indivíduos e o entendimento das semelhanças e diferenças, é um interesse central. Na análise discriminante a esferização das nuvens, a presença de duas métricas gerando vários coeficientes e correspondentes correlações, dificulta a tarefa acima mencionada.

Neste trabalho vimos que é possível analisar melhor quando utilizamos os coeficientes *Dentro, Bruto* e *Total* de forma comparativa; as correlações entre originais e discriminantes para orientar as interpretações das diferenças e semelhanças entre indivíduos no plano discriminante e ainda possível complementação com círculos de correlação nos casos de interesse.

Na Análise Discriminante de Fisher através das funções lineares discriminantes, obtem-se as direções que melhor permitem visualizar as diferenças entre grupos. Os círculos de correlação entre as funções lineares discriminantes e as originais, quando tomados Entre os grupos, revelam se quais e como as variáveis contribuem para estas diferenças. Também as diferenças entre indivíduos de diferentes grupos pode aí ser melhor entendida.

Os círculos de correlação entre as funções lineares discriminante e as originais, quando tomadas *Dentro*, revelam se, quais e como as variáveis contribuem para as diferenças entre indivíduos dentro de cada grupo na presença das demais.

Os círculos de correlação entre as funções lineares discriminante e as originais, quando tomadas *Total*, revelam por comparação com os anteriores as semelhanças e diferenças e influências dominantes das variáveis nas direções preferenciais de alongamento total da nuvem.

Capítulo 10: Referências Bibliográficas

[1] FISHER, R.A. The Use of Multiple Measurements in Taxonomic Problems, **Annals of Eugenics**, 7 , 179-188. 1936

[2] HAND, D.J. et al, **Handbook of Small Data Sets**. London: Chapman & Hall, 1994. p. 254-255.

[3] JOHNSON, R.A. e WICHERN, D.W. **Applied Multivariate Statistical Analysis**. New York: Prentice-Hall International, Inc, 1992. 642 p.

[4] KRZANOWSKI, W.J. **Principles of Multivariate Analysis; a users perspective**. New York: Clarendon Press Oxford, 1988. 563 p. (Oxford Statistical Science Series).

[5] LEBART, L., MORINEAU, A. E PIRON, M. **Statistique Exploratoire Multidimensionnelle**. Paris: DUNOD, 1995.

[6] LINDSEY, J.C., HERZBERG, A.M. e WATTS, D.G. A method for cluster analysis based on projections and quantile-quantile plots, **Biometrics**, 43, 327-341. 1987.

[7] LUBISCHEW, A.A. On the use of discriminant functions in taxonomy, **Biometrics**, December, 1962.

[8] MARDIA, K.V., KENT, J.T. e BIBBY, J.M. **Multivariate Analysis**. London: Academic Press, 1979. 521 p.

[9] CRIVISQUI,.E., Análise em Componentes Principais, Programa **PRESTA**, 1999.

[10] STEVENS,.J. **Applied Multivariate Statistic in Social Sciences.** Lawrence Erlbaum Associates. 1992.

[11] WATKINS, D.S. **Fundamentals of Matrix Computations**. John Wiley & Sons. 1991

[12] LACHENBRUCH, P.A. **Discriminant Analysis**. Collier Macmillan Canada Ltd. 1975.

[13] GITTINS, R. **Canonical Analysis:** A Review with Applications in Ecology. Springer-Verlag Berlin Heidelberg. 1985.

[14] GILBERT, E. The Effect of Unequal Variance-Covariance Matrices on Fisher´s Linear Discriminant Function. **Biometrics**, September 1969..

Capítulo 11: Bibliografia

[1] ANDERSON, R.A., R.A. Fisher and Multivariate Analysis, **Statistical Sciences**, v. 11, No. 1, 1996, p.20-34.

[2] KSHIRSAGAR, A.M. **Multivariate Analysis**. Marcel Dekker, Inc. 1972

[3] LAWLEY, D.N., Test of Significance in Canonical Analysis. **Biometrika**, 46, 59-66. 1959

[4] RAO, C. R. Linear **Statistical Inference and Its Applications**. Jonh Wiley & Sons. 1965.

[5] SHARMA, S. **Applied Multivariate Techniques**. Jonh Wiley & Sons. 1996

Anexo 1: Saída completa da Análise Discriminante de Fisher para os dados dos Insetos

Canonical Discriminant Analysis

30	Observations	29	DF Total
3	Variables	27	DF Within Classes
3	Classes	2	DF Between Classes

Class Level Information

ESPECIE	Frequency	Weight	Proportion
1	10	10.0000	0.333333
2	10	10.0000	0.333333
3	10	10.0000	0.333333

Within-Class SSCP Matrices

ESPECIE = 1

Variable	X1	X2	X3
X1	1492.900000	710.600000	223.500000
X2	710.600000	440.400000	125.000000
X3	223.500000	125.000000	50.500000

ESPECIE = 2

Variable	X1	X2	X3
X1	2669.600000	861.400000	391.200000
X2	861.400000	1033.600000	84.800000
X3	391.200000	84.800000	82.900000

ESPECIE = 3

Variable	X1	X2	X3
X1	1201.600000	672.600000	271.400000
X2	672.600000	597.600000	206.400000
X3	271.400000	206.400000	102.100000

Pooled Within-Class SSCP Matrix

Variable	X1	X2	X3
X1	5364.100000	2244.600000	886.100000
X2	2244.600000	2071.600000	416.200000
X3	886.100000	416.200000	235.500000

Between-Class SSCP Matrix

Variable	X1	X2	X3
X1	25780.20000	246.40000	-611.80000
X2	246.40000	209.06667	26.13333
X3	-611.80000	26.13333	19.46667

Total-Sample SSCP Matrix

Variable	X1	X2	X3
X1	31144.30000	2491.00000	274.30000
X2	2491.00000	2280.66667	442.33333
X3	274.30000	442.33333	254.96667

Within-Class Covariance Matrices

ESPECIE = 1 DF = 9

Variable	X1	X2	X3
X1	165.8777778	78.9555556	24.8333333
X2	78.9555556	48.9333333	13.8888889
X3	24.8333333	13.8888889	5.6111111

ESPECIE = 2 DF = 9

Variable	X1	X2	X3
X1	296.6222222	95.7111111	43.4666667
X2	95.7111111	114.8444444	9.4222222
X3	43.4666667	9.4222222	9.2111111

ESPECIE = 3 DF = 9

Variable	X1	X2	X3
X1	133.5111111	74.7333333	30.1555556
X2	74.7333333	66.4000000	22.9333333
X3	30.1555556	22.9333333	11.3444444

Pooled Within-Class Covariance Matrix DF = 27

Variable	X1	X2	X3
X1	198.6703704	83.1333333	32.8185185
X2	83.1333333	76.7259259	15.4148148
X3	32.8185185	15.4148148	8.7222222

Between-Class Covariance Matrix DF = 2

Variable	X1	X2	X3
X1	1289.010000	12.320000	-30.590000
X2	12.320000	10.453333	1.306667
X3	-30.590000	1.306667	0.973333

Total-Sample Covariance Matrix DF = 29

Variable	X1	X2	X3
X1	1073.941379	85.896552	9.458621
X2	85.896552	78.643678	15.252874
X3	9.458621	15.252874	8.791954

Pooled Within-Class Correlation Coefficients / Prob > |R|

Variable	X1	X2	X3
X1	1.00000	0.67334	0.78839
	0.0	0.0001	0.0001
X2	0.67334	1.00000	0.59587
	0.0001	0.0	0.0008
X3	0.78839	0.59587	1.00000
	0.0001	0.0008	0.0

Between-Class Correlation Coefficients / Prob > |R|

Variable	X1	X2	X3
X1	1.00000	0.10613	-0.86362
	0.0	0.9323	0.3364
X2	0.10613	1.00000	0.40964
	0.9323	0.0	0.7313
X3	-0.86362	0.40964	1.00000
	0.3364	0.7313	0.0

Total-Sample Correlation Coefficients / Prob > |R|

Variable	X1	X2	X3
X1	1.00000	0.29557	0.09734
	0.0	0.1128	0.6088
X2	0.29557	1.00000	0.58007
	0.1128	0.0	0.0008
X3	0.09734	0.58007	1.00000
	0.6088	0.0008	0.0

Pairwise Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Squared Distance to ESPECIE

From ESPECIE	1	2	3
1	0	21.82490	26.39006
2	21.82490	0	94.47023
3	26.39006	94.47023	0

F Statistics, NDF=3, DDF=25 for Squared Distance to ESPECIE

From ESPECIE	1	2	3
1	0	33.68041	40.72540
2	33.68041	0	145.78739
3	40.72540	145.78739	0

Prob > Mahalanobis Distance for Squared Distance to ESPECIE

From ESPECIE	1	2	3
1	1.0000	0.0001	0.0001
2	0.0001	1.0000	0.0001
3	0.0001	0.0001	1.0000

Univariate Test Statistics

F Statistics, Num DF= 2 Den DF= 27

Variable	Total STD	Pooled STD	Between STD	R-Squared	RSQ/ (1-RSQ)
X1	32.7710	14.0950	35.9028	0.827766	4.8061
X2	8.8681	8.7593	3.2332	0.091669	0.1009
X3	2.9651	2.9533	0.9866	0.076350	0.0827

Variable	F	Pr > F
X1	64.8818	0.0001
X2	1.3624	0.2731
X3	1.1159	0.3423

Average R-Squared: Unweighted = 0.3319284
Weighted by Variance = 0.7722323

Multivariate Statistics and F Approximations

S=2 M=0 N=11.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.04879594	29.3915	6	50	0.0001
Pillai's Trace	1.04284532	9.4426	6	52	0.0001
Hotelling-Lawley Trace	17.61545551	70.4618	6	48	0.0001
Roy's Greatest Root	17.50818851	151.7376	3	26	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.972610	0.970631	0.010033	0.945970
2	0.311248	0.266684	0.167706	0.096875

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	17.5082	17.4009	0.9939	0.9939
2	0.1073	.	0.0061	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.04879594	29.3915	6	50	0.0001
2	0.90312454	1.3945	2	26	0.2659

Estrutura de correlação para a amostra total

	CAN1	CAN2
X1	0.932575	0.228559
X2	0.008735	0.972374
X3	-0.255798	0.386249

Estrutura de correlação entre os grupos

	CAN1	CAN2
X1	0.996938	0.078190
X2	0.028061	0.999606
X3	-0.900391	0.435081

Estrutura de correlação dentro dos grupos

	CAN1	CAN2
X1	0.522327	0.523375
X2	0.002130	0.969584
X3	-0.061867	0.381933

Total-Sample Standardized Canonical Coefficients

	CAN1	CAN2
X1	4.135019571	-0.088354638
X2	-0.507121902	1.176936385
X3	-1.170188187	-0.281927862

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2
X1	1.778499846	-0.038001926
X2	-0.500900573	1.162497827
X3	-1.165538380	-0.280807606

Raw Canonical Coefficients

	CAN1	CAN2
X1	0.1261790581	-.0026961190
X2	-.0571847806	0.1327153267
X3	-.3946508161	-.0950813401

Class Means on Canonical Variables

ESPECIE	CAN1	CAN2
1	0.157527627	0.439236337
2	4.781014296	-0.230296412
3	-4.938541923	-0.208939925

Anexo 2: Exemplos dos programas para os casos 1 e 4 bivariados[1]

```
libname tese "c:\users\celeste";

*       Programa para a configuração POSITIVA-POSITIVA. Análise Discriminante de Fisher;
*       construção dos graficos das correlações entre as variáveis originais e as canônicas;

                    /*  CASO 1. VARIÂNCIAS IGUAIS */

title;

  /* calcula o valor da chi-quadrado de interesse */

        data chi;
           n=3; alfa=0.05;
             do gl=n-1;      * graus de libertade da chi-quadrado;
               chi=cinv(1-alfa, gl);
              output;
             end;
    proc print data=chi; run;

/* calcula os valores do L e o k para depois poder localizar corretamente as médias */
  /* trabalhando com rho=.9 */

    data eleka;
       set chi;
         retain chi;
         var1=1; var2=1; rho=.9;
         std1=sqrt(var1);  std2=sqrt(var2); cov=rho*std1*std2;
         c=sqrt(chi);
         L=c*sqrt(var1+cov);
         k=L/sqrt(2);
    run;

proc print data=eleka;
     var c L k;
run;

                         /* geração dos dados*/
data tese.geranorm;
  set eleka;
  keep amostra xorig yorig;

  n=3;                * numero de amostras a serem geradas = numero de amostras;
  tamanho=1000;       * tamanho das amostras;
  teta=30;            * angulo teta escolhido para rotar os pontos;

  c=sqrt(1-rho**2);

    do j=1 to n;
       amostra=j;

      do i = 1 to tamanho;

        /*  gera variaveis N(0,1) com coeficiente de correlacao rho  */

            xorig = rannor(123);
            yorig = rho*xorig+c*rannor(123);

        /*  transforma estas variaveis para ter a media e a variancia precisada  */
```

[1] programas em SAS feitos para desenvolver a simulação dos dados, a Análise Discriminante de Fisher e a construção dos Círculos das Correlações

Casos 1 e 4 Bivariado[1]

```
libname tese "c:\users\celeste";


*       Programa para a configuração POSITIVA-POSITIVA. Análise Discriminante de Fisher;
*       construção dos graficos das correlações entre as variáveis originais e as canônicas;


                    /*  CASO 1. VARIÂNCIAS IGUAIS */

title;

  /* calcula o valor da chi-quadrado de interesse */

        data chi;
           n=3; alfa=0.05;
             do gl=n-1;     * graus de libertade da chi-quadrado;
               chi=cinv(1-alfa, gl);
              output;
             end;
    proc print data=chi; run;


 /* calcula os valores do L e o k para depois poder localizar corretamente as médias */
  /* trabalhando com rho=.9 */

    data eleka;
       set chi;
         retain chi;
         var1=1; var2=1; rho=.9;
         std1=sqrt(var1);  std2=sqrt(var2); cov=rho*std1*std2;
         c=sqrt(chi);
         L=c*sqrt(var1+cov);
         k=L/sqrt(2);
    run;

proc print data=eleka;
     var c L k;
run;

                         /* geração dos dados*/
data tese.geranorm;
  set eleka;
  keep amostra xorig yorig;

  n=3;                * numero de amostras a serem geradas = numero de amostras;
  tamanho=1000;       * tamanho das amostras;
  teta=30;            * angulo teta escolhido para rotar os pontos;

  c=sqrt(1-rho**2);

    do j=1 to n;

        amostra=j;

       do i = 1 to tamanho;

         /*  gera variaveis N(0,1) com coeficiente de correlacao rho  */

             xorig = rannor(123);
             yorig = rho*xorig+c*rannor(123);

         /*  transforma estas variaveis para ter a media e a variancia precisada  */
```

[1] programas em SAS feitos para desenvolver a simulação dos dados, a Análise Discriminante de Fisher e a construção dos Círculos das Correlações

```
        /* calcula as posições das médias dos amostras (exceto do primeiro),
            segundo o calculo feito para calcular o k anteriormente */

            xorig = (2*j - 2)*k  + sqrt(var1)*xorig;
            yorig = (2*j - 2)*k  + sqrt(var2)*yorig;

         output;
      end;
  end;
run;

proc corr data=tese.geranorm noprob outp=tese.pearson1;
     var xorig yorig;
     by amostra;
run;

proc print data=tese.pearson1;
run;

title;
            /*********************************/
            /*       DADOS PERTURBADOS       */
            /*********************************/

data tese.perturba;
     set tese.geranorm;
              var1=1; var2=1;
             desvio1=sqrt(var1);
             desvio2=sqrt(var2); rmil=sqrt(1000);

       /* Perturbação aleatória das coordenadas no eixo x e no y */

              /* gerar U(-2/sqrt(1000),+2/sqrt(1000)) */

       if amostra=1 then do;

              x=xorig;
              y=yorig;
       end;

       if amostra>1 then do;

              uni01X=ranuni(12);
              uni01Y=ranuni(23);

              uniabX=2*(2/rmil)*uni01X-(2/rmil);
              uniabY=2*(2/rmil)*uni01Y-(2/rmil);

       /* perturbando as coordenadas x e y */

              x=xorig+uniabX;
              y=yorig+uniabY;
       end;
run;

proc corr data=tese.perturba noprob outp=tese.pearson2;
     var x y;
     by amostra;
run;

proc print data=tese.pearson2;
run;

proc gplot data=tese.perturba;
       plot y*x;
run;
quit;
```

```
options ps=30 ls=78 nodate nonumber;

  /*****************************************************/
  /*     Analise Discriminante Canonica de Fisher      */
  /*****************************************************/

proc candisc data=tese.perturba out=tese.out11 outstat=tese.out12 distance anova
             wsscp wcov psscp pcov tsscp tcov bsscp bcov;
     var x y ;
     class amostra;
run;


title;
        /*****************************************/
        /*    Gráfico das Variáveis Canônicas    */
        /*****************************************/

proc gplot data=tese.out11;
     plot can2*can1=amostra;
     format amostra;
     title2 "Gráfico das Variáveis Canônicas com tres grupos";
     title3 "Posição dos grupos no sentido de correlação positiva entre eles";
     title4 "Coeficiente de correlação dentro dos grupos: rho=+0,9";
     title5 " Variâncias Iguais";
run;
quit;

title;

        /*****************************************************************/
        /* A seguir converte-se a saida do proc candisc num data set SAS */
        /*   conveniente para fazer os graficos                          */
        /*****************************************************************/

title;
data teste;
     set tese.out12;
     if _type_ in ('STRUCTUR', 'BSTRUCT', 'PSTRUCT', 'SCORE', 'PSCORE', 'RAWSCORE');
run;

proc sort data=teste;
     by _NAME_;
run;

data teste1;
     set teste;
     retain x_st y_st x_bs y_bs
            x_ps y_ps  x_sc y_sc
            x_psc y_psc  x_rs y_rs;
     drop x y amostra _type_;

     if _type_='STRUCTUR' then do;
        x_st=x;
        y_st=y;

     end;
     if _type_='BSTRUCT' then do;
        x_bs=x;
        y_bs=y;

     end;
     if _type_='PSTRUCT' then do;
        x_ps=x;
        y_ps=y;

     end;
     if _type_='SCORE' then do;
```

```
        x_sc=x;
        y_sc=y;

     end;
     if _type_='PSCORE' then do;
        x_psc=x;
        y_psc=y;

     end;
     if _type_='RAWSCORE' then do;
        x_rs=x;
        y_rs=y;

        output;
     end;
run;

title;
*proc print data=teste1;
*     id _NAME_;
*run;

proc transpose data=teste1 out=saida;
run;

        /*************************/
        /*    Macro - Grafico    */
        /*************************/

%macro grafico(nvar,amostra);
data saida1;
  set saida;
  drop n i;
  n=&nvar;
  do i=0 to 5;
    if (i*n+1 le _n_ le (i+1)*n) then classe=i+1;
  end;
run;

data auxiliar;
       set saida1;
         if classe=&amostra;
         xvar=can1; yvar=can2; name=_name_;

data anno;
       set auxiliar;
        retain when 'a';
        length function text style $8;

/* Determina o valor de pi */
        pi=arcos(-1);

/* Determina a longitude do lado mais longe e dos lados adjacentes do triangulo reto */
        if xvar>0 and yvar>0 then do;
                newx=xvar;
                newy=yvar;
           end;
        if xvar<0 and yvar>0 then do;
                newx=abs(xvar - 0);
                newy=yvar-0;
           end;
        if xvar<0 and yvar<0 then do;
                newx=abs(xvar-0);
                newy=abs(yvar-0);
           end;
        if xvar>0 and yvar<0 then do;
                newx=xvar;
                newy=abs(yvar-0);
```

```
          end;
/* Determina o angulo adjacente do triangulo reto em radianes  */
        radA=atan(newy/newx);
/* Determina o angulo adjacente em graus */
        A=(radA*180)/pi;
/* Determina o angulo oposto do triangulo reto em graus */
        B=90-A;
/* Determina a longitude da hipotenusa do triangulo reto   */
        c=sqrt(newx**2 + newy**2);
/* Ajusta as cordenadas Y tal que as cabeças das setas nao sobre-escrevam simbolos */
        adjy=sin (radA)*(c-.1);
/* Ajusta as cordenadas X tal que as cabeças das setas nao sobre-escrevam simbolos */
        adjx=cos (radA)*(c-.1);
/* Especifica condicionalmente o angulo e as coordenadas X/Y  */
        if xvar>0 and yvar>0 then do; adjx=adjx; ang=-B; end;
        if xvar<0 and yvar>0 then do; adjx=-adjx; ang=B; end;
        if xvar<0 and yvar<0 then do; adjx=-adjx; adjy=-adjy; ang=90+A; end;
        if xvar>0 and yvar<0 then do; adjx=adjx; adjy=-adjy; ang=-90-A; end;
        if xvar=0 and yvar>0 then do; adjx=xvar; adjy=yvar-.1; ang=0; end;
        if xvar=0 and yvar<0 then do; adjx=xvar; adjy=yvar+.1; ang=180; end;
        if xvar>0 and yvar=0 then do; adjx=xvar-.1; adjy=yvar; ang=-90; end;
        if xvar<0 and yvar=0 then do; adjx=xvar+.1; adjy=yvar; ang=90; end;
/* Gera ramas de setas começando do (0,0) */
        function='move'; xsys='2'; ysys='2'; x=0; y=0; output;
        function='draw'; xsys='2'; ysys='2'; x=adjx; y=adjy; output;
/* Gera as cabeças das setas*/
        function='label'; xsys='2'; ysys='2'; x=adjx; y=adjy; output;
                style='marker'; text='C';
                angle=ang; position='5'; output;
/* Disenha a circunferencia  */
        function='pie'; xsys='2'; ysys='2'; hsys='2'; x=0; y=0;
                angle=0; rotate=360; size=1;
                style='empty'; when='a'; output;
/* Label da circunferencia */
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
           style='simplex';
                x=1; y=0; text='1'; position='3'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=-1; y=0; text='1'; position='1'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=0; y=1; text='1'; position='3'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=0; y=-1; text='1'; position='9'; output;
/* Pontos Labels  */
```

```
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=xvar; y=yvar; text=right(name);
                if xvar>0 then do;
                     position='3'; text=right(name);
                     end;
                if xvar<0 then do;
                     position='1'; text=left(name);
                     end;
                output;
run;

/* Simbolos */

        symbol v=star i=none c=black;
        axis1 order=(-2 to 2 by 1) length=4 in

  /* tirando os eixos */
        style=0 value=none label=none major=none minor=none;

proc gplot data=auxiliar;
 plot yvar*xvar/ annotate=anno
                vref=0 lvref=1 cvref=black
                href=0 lhref=1 chref=black
                vaxis=axis1 haxis=axis1;
run;
quit;

%mend;

%grafico(2,1);
%grafico(2,2);
%grafico(2,3);
quit;


/*************************************************************************/
/*       Trabalho com as amostras com variâcias diferentes de tal forma  */
/*       de manter a mesma VARIANCIA GENERALIZADA que no caso das amostras */
/*                com variâncias iguais, i.e, det(sigma1)=det(sigma2)     */
/*************************************************************************/

               /* CASO 4. VARIÂNCIAS DIFERENTES */
title;

proc iml;

create tese.sigma var{cD angrad angrau rhoD var1D var2D stdD1 stdD2 LD kx ky};

     vezes=2; rhoD=0.9; vargene1=0.19;

     *Construindo a matriz sigma;

     p=2;
     sigma=j(p,p,0);
     sigma[2,2]=sqrt(vargene1/(vezes*(1-rhoD**2)));
     sigma[1,1]=2*sigma[2,2];
     sigma[1,2]=rhoD*sqrt(vezes)*sigma[2,2];
     sigma[2,1]=sigma[1,2];
     var1D=sigma[1,1]; var2D=sigma[2,2];
     stdD1=sqrt(sigma[1,1]); stdD2=sqrt(sigma[2,2]);
     print sigma;

     * Autovalores e autovetores;

     eval=j(p,1,0); evec=j(p,p,0);
     call eigen(eval,evec,sigma);
     print eval, evec;
```

```
     * calcula o valor da chi-quadrado de interesse ;

     n=3; alfa=0.05;
     gl=n-1;      * graus de libertade da chi-quadrado;
     chi=cinv(1-alfa, gl);
     print chi;
     cD=sqrt(chi);  pi=arcos(-1);

     * calcula o ângulo de inclinação do primeiro autovetor;

     b=evec[2,1]/evec[1,1];
     angrad=atan(b);
     angrau=angrad*180/pi;  * convirtiendo de radianes para graus;

     * calculando as coordenadas das projeções do vértice maior da elipse;
     * calcula os valores do LD para depois poder localizar corretamente as médias;

     LD=cD*sqrt(eval[1,1]); * calculando o cumprimento do semieixo maior da elipse;
     kx=LD*cos(angrad);
     ky=LD*sin(angrad);
     print kx ky;

append;

quit;

proc print data=tese.sigma;
run;

                        /* geração dos dados N(0,1)  */
data tese.geranorD;
  set tese.sigma;
  keep amostras xori yori;
  n=3;                * numero de amostrass a serem geradas = numero de amostrass;
  tamanho=1000;       * tamanho das amostrass;
  c2=sqrt(1-rhoD**2);

    do j=1 to n;

         amostras=j;

       do i = 1 to tamanho;

         /*  gera variaveis N(0,1) com coeficiente de correlaçao rho  */

             xori = rannor(123);
             yori = rhoD*xori+c2*rannor(123);

         /*  transforma estas variaveis para ter a media e a variancia precisada  */
         /* calcula as posições das médias dos amostras (exceto do primeiro),
             segundo o calculo feito para calcular o k anteriormente */

             xori = (2*j-2)*kx + stdD1*xori;
             yori = (2*j-2)*ky + stdD2*yori;

          output;
      end;
  end;
run;

proc corr data=tese.geranorD noprob outp=tese.pearso1D;
     var xori yori;
     by amostras;
run;

proc print data=tese.pearso1D;
run;
```

```
proc gplot data=tese.geranorD;
     plot yori*xori;
run;
quit;

          /*************************************************************/
          /*       DADOS (var. diferentes) APÓS A PERTURBAÇÃO          */
          /*************************************************************/
title;
data tese.perturbD;
     set tese.geranorD;

     vezes=2;       * vezes que uma variância é diferente da outra, supondo sempre var1
maior;
     rhoD=0.9;      * valor da correlação;
     vargene1=0.19;
     var2D=sqrt(vargene1/(vezes*(1-rhoD**2)));
     var1D=2*var2D;
     stdD1=sqrt(var1D); stdD2=sqrt(var2D);
     rmil=sqrt(1000);
     stdmedD1=stdD1/rmil; stdmedD2=stdD2/rmil;

        /* Perturbação aleatória das coordenadas no eixo x e no y */
        /* E feita perturbando gerando um valor que pode ser até duas vezes
           em valor absoluto o desvio padrão da média de cada variável*/

                /* gerar U(-2*stdmedD,+2*stdmedD) */

        if amostras=1 then do;

                x=xori;
                y=yori;
        end;

        if amostras>1 then do;

                uni01X=ranuni(12);
                uni01Y=ranuni(23);

                unixBX=(2*stdmedD1)*uni01X-(stdmedD1);
                unixBY=(2*stdmedD2)*uni01Y-(stdmedD2);

        /* perturbando as coordenadas x e y */

                x=xori+unixBX;
                y=yori+unixBY;
        end;
run;

proc corr data=tese.perturbD noprob outp=tese.pearso2D;
     var x y;
     by amostras;
run;

proc print data=tese.pearso2D;
run;

proc gplot data=tese.perturbD;
        plot y*x;
        title1 "Configuração Positiva-positiva";
        title2 "Variâncias Diferentes";
        title3 "sigma[1,1]=2*sigma[2,2]";
run;
quit;

options ps=30 ls=78 nodate nonumber;

        /*****************************************/
```

```
        /*    Candisc para variâncias diferentes    */
        /*******************************************/
title;
proc candisc data=tese.perturbD out=tese.outD1 outstat=tese.outD2
             wsscp psscp tsscp bsscp wcov pcov tcov bcov;
     var x y;
     class amostras;
run;

title;
        /*******************************************/
        /*    Grafico das Variaveis Canonicas      */
        /*******************************************/

proc gplot data=tese.outD1;
     plot can2*can1=amostras;
     format amostras;
     title2 "Gráfico das Variáveis Canônicas com tres grupos";
     title3 "Posição dos grupos no sentido de correlação positiva entre eles";
     title4 "Coeficiente de correlação dentro dos grupos: rho=+0,9";
     title5 "Variâncias Diferentes";
run;
quit;

        /*****************************************************/
        /* Converte a saida do proc candisc num data set SAS */
        /*         para fazer os graficos                    */
        /*****************************************************/

title;
data testeD;
     set tese.outD2;
     if _type_ in ('STRUCTUR', 'BSTRUCT', 'PSTRUCT', 'SCORE', 'PSCORE', 'RAWSCORE');
run;

proc sort data=testeD;
     by _NAME_;
run;

data teste2;
     set testeD;
     retain x_st y_st x_bs y_bs
            x_ps y_ps  x_sc y_sc
            x_psc y_psc  x_rs y_rs;
     drop x y amostras _type_;

     if _type_='STRUCTUR' then do;
        x_st=x;
        y_st=y;

     end;
     if _type_='BSTRUCT' then do;
        x_bs=x;
        y_bs=y;

     end;
     if _type_='PSTRUCT' then do;
        x_ps=x;
        y_ps=y;

     end;
     if _type_='SCORE' then do;
        x_sc=x;
        y_sc=y;

     end;
     if _type_='PSCORE' then do;
        x_psc=x;
```

```
        y_psc=y;

     end;
     if _type_='RAWSCORE' then do;
        x_rs=x;
        y_rs=y;

        output;
     end;
run;

title;
*proc print data=teste2;
*     id _NAME_;
*run;

proc transpose data=teste2 out=saidaD;
run;

        /*************************/
        /*    Macro - Grafico    */
        /*************************/

%macro grafico(nvar,amostra);

data saida2;
  set saidaD;
  drop n i;
  n=&nvar;
  do i=0 to 5;
    if (i*n+1 le _n_ le (i+1)*n) then classe=i+1;
  end;
run;

data auxiliaD;
       set saida2;
         if classe=&amostra;
         xvar=can1; yvar=can2; name=_name_;

data annoD;
       set auxiliaD;
        retain when 'a';
        length function text style $8;

/* Determina o valor de pi */
        pi=arcos(-1);

/* Determina a longitude do lado mais longe e dos lados adjacentes do triangulo reto */
        if xvar>0 and yvar>0 then do;
                newx=xvar;
                newy=yvar;
           end;
        if xvar<0 and yvar>0 then do;
                newx=abs(xvar - 0);
                newy=yvar-0;
           end;
        if xvar<0 and yvar<0 then do;
                newx=abs(xvar-0);
                newy=abs(yvar-0);
           end;
        if xvar>0 and yvar<0 then do;
                newx=xvar;
                newy=abs(yvar-0);
           end;

/* Determina o angulo adjacente do triangulo reto em radianes  */

        radA=atan(newy/newx);
```

```
/* Determina o angulo adjacente em graus */

        A=(radA*180)/pi;

/* Determina o angulo oposto do triangulo reto em graus */

        B=90-A;

/* Determina a longitude da hipotenusa do triangulo reto   */

        c=sqrt(newx**2 + newy**2);

/* Ajusta as cordenadas Y tal que as cabeças das setas nao sobre-escrevam simbolos */

        adjy=sin (radA)*(c-.1);

/* Ajusta as cordenadas X tal que as cabeças das setas nao sobre-escrevam simbolos */

        adjx=cos (radA)*(c-.1);

/* Especifica condicionalmente o angulo e as coordenadas X/Y  */

        if xvar>0 and yvar>0 then do; adjx=adjx; ang=-B; end;
        if xvar<0 and yvar>0 then do; adjx=-adjx; ang=B; end;
        if xvar<0 and yvar<0 then do; adjx=-adjx; adjy=-adjy; ang=90+A; end;
        if xvar>0 and yvar<0 then do; adjx=adjx; adjy=-adjy; ang=-90-A; end;
        if xvar=0 and yvar>0 then do; adjx=xvar; adjy=yvar-.1; ang=0; end;
        if xvar=0 and yvar<0 then do; adjx=xvar; adjy=yvar+.1; ang=180; end;
        if xvar>0 and yvar=0 then do; adjx=xvar-.1; adjy=yvar; ang=-90; end;
        if xvar<0 and yvar=0 then do; adjx=xvar+.1; adjy=yvar; ang=90; end;

/* Gera ramas de setas começando do (0,0) */

        function='move'; xsys='2'; ysys='2'; x=0; y=0; output;
        function='draw'; xsys='2'; ysys='2'; x=adjx; y=adjy; output;

/* Gera as cabeças das setas*/

        function='label'; xsys='2'; ysys='2'; x=adjx; y=adjy; output;
                style='marker'; text='C';
                angle=ang; position='5'; output;

/* Disenha a circunferencia  */

        function='pie'; xsys='2'; ysys='2'; hsys='2'; x=0; y=0;
                angle=0; rotate=360; size=1;
                style='empty'; when='a'; output;

/* Label da circunferencia */

        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
           style='simplex';
                x=1; y=0; text='1'; position='3'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=-1; y=0; text='1'; position='1'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=0; y=1; text='1'; position='3'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=0; y=-1; text='1'; position='9'; output;

/* Pontos Labels  */

        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=xvar; y=yvar; text=right(name);
                if xvar>0 then do;
                     position='3'; text=right(name);
                     end;
```

```
                    if xvar<0 then do;
                         position='1'; text=left(name);
                         end;
                    output;
run;

/* Simbolos */

        symbol v=star i=none c=black;
        axis1 order=(-2 to 2 by 1) length=4 in

  /* tirando os eixos */
        style=0 value=none label=none major=none minor=none;

proc gplot data=auxiliaD;
 plot yvar*xvar/ annotate=annoD
                 vref=0 lvref=1 cvref=black
                 href=0 lhref=1 chref=black
                 vaxis=axis1 haxis=axis1;
run;
quit;

%mend;

%grafico(2,1);
%grafico(2,2);
%grafico(2,3);
quit;
```

Anexo 3: Exemplo do programa para o caso 1 trivariado[1]

```
libname tese "c:\users\celeste";

                    /*******************************/
                    /*     CASO 1.Trivariada       */
                    /*******************************/
title;
proc iml;

     *Construindo a matriz sigma;

     rho=0.9;
     p=3;
     sigma=j(p,p,0);

     sigma[1,1]=1;
     sigma[2,2]=1;
     sigma[3,3]=1;

     std1=sqrt(sigma[1,1]);
     std2=sqrt(sigma[2,2]);
     std3=sqrt(sigma[3,3]);

     sigma[1,2]=rho*std1*std2; sigma[2,1]=rho*std1*std2;
     sigma[1,3]=rho*std1*std3; sigma[3,1]=rho*std1*std3;
     sigma[2,3]=rho*std2*std3; sigma[3,2]=rho*std2*std3;

     print sigma;

     var1=sigma[1,1];
     var2=sigma[2,2];
     var3=sigma[3,3];
     print var1 var2 var3;

     * Autovalores e autovetores;

     eval=j(p,1,0); evec=j(p,p,0);
     call eigen(eval,evec,sigma);
     print eval evec;

     * calcula o valor da chi-quadrado de interesse ;

     alfa=0.05;
     gl=p;     * graus de libertade da chi-quadrado;
     chi=cinv(1-alfa, gl);
     print chi;
     c=sqrt(chi);  pi=arcos(-1);

     * calcula o ângulo de inclinação do primeiro autovetor no quadrante 12;

     b12=evec[2,1]/evec[1,1];
     angrad12=atan(b12);
     angrau12=angrad12*180/pi;  * convirtiendo de radianes para graus;
```

[1] programa em SAS feito para desenvolver a simulação dos dados, a Análise Discriminante de Fisher e a construção dos Círculos das Correlações.

```
    * calcula o ângulo de inclinação do primeiro autovetor no quadrante 13;

     b13=evec[3,1]/evec[1,1];
     angrad13=atan(b13);
     angrau13=angrad13*180/pi;  * convirtiendo de radianes para graus;

    * calcula o ângulo de inclinação do primeiro autovetor no quadrante 23;

     b23=evec[3,1]/evec[2,1];
     angrad23=atan(b23);
     angrau23=angrad23*180/pi;  * convirtiendo de radianes para graus;

     print b12 b13 b23 angrau12 angrau13 angrau23;

     ************************************************************************;
     * calculando o cumprimento do semieixo maior do elipsoide (no espaço)  ;
     ************************************************************************;

     Lreal= c*sqrt(eval[1,1]);

*****************************************************************************;
* calculando as coordenadas das projeções do vértice maior da elipse         ;
* (no planos) calcula os valores do L e k para depois poder localizar       ;
*   *corretamente as médias                                                  ;

* calcula os valores do L e k para depois poder localizar corretamente as médias;

     *************************;
     ***   QUADRANTE 12    ***;
     *************************;

     L12=Lreal/sqrt(2);; * calculando o cumprimento do semieixo maior da elipse;
     k12=L12*cos(angrad12);
     print L12 k12;

     *************************;
     ***   QUADRANTE 13    ***;
     *************************;

     L13=Lreal/sqrt(2); * calculando o cumprimento do semieixo maior da elipse;
     k13=L13*cos(angrad13);
     print k13;

     *************************;
     ***   QUADRANTE 23    ***;
     *************************;

     L23=Lreal/sqrt(2); * calculando o cumprimento do semieixo maior da elipse;
     k23=L23*cos(angrad23);
     print k23;

     /************************************************/
     /***  CÁLCULO DAS COORDENADAS PARA AS MÉDIAS   ***/
     /************************************************/

     media1=j(p,1,0);
     media2=j(p,1,2*k12);
     media3=j(p,1,4*k12);

  print media1 media2 media3;
  dados = media1||media2||media3||sigma;
  create med from dados;
  append from dados;
```

```
quit;

*proc print data=med;
*run;


/*************************************************************************/
/******      MACRO PARA A GERAÇÃO DOS DADOS TRIVARIADOS     **************/
/*************************************************************************/

%macro mvn(varcov=,        /* dataset for variance-covariance matrix */
           n=,             /* sample size */
           seed=,         /* seed for random number generator */
           sample=);       /* output dataset name */

 /* Generate the multivariate normal data in SAS/IML */

proc iml worksize=100;

   use &varcov;           /* read variance-covariance matrix */
   read all into info;

   p=3;
   cov = info[,4:6];
   mu = info[,1:3];

   v=nrow(cov);            /* calculate number of variables */
   n=&n;
   seed = &seed;
   l=t(root(cov));         /* calculate cholesky root of cov matrix */
   z=normal(j(v,&n,&seed));/* generate nvars*samplesize normals */
   x=l*z;                  /* premultiply by cholesky root */
   x3=x//x//x;
   mu3=mu[,1]//mu[,2]//mu[,3];

   x3=repeat(mu3,1,&n)+x3;    /* add in the means */

   tx=t(x3);
   varnames={'X11' 'X12' 'X13' 'X21' 'X22' 'X23' 'X31' 'X32' 'X33'};

   create &sample from tx [colname=varnames];  /* write out sample data to sas
dataset */
   append from tx;
quit;
%mend mvn;
*_____________________________FIM DA MACRO_____________________________;


%mvn(varcov=med, n=1000, seed=123, sample=sample);

title;
            /*********************************/
            /*       DADOS perturbados       */
            /*********************************/

data tese.perturba;
     set sample;
     RMIL=SQRT(1000);
        /* Perturbação aleatória das coordenadas nos três eixos */

               /* gerar U(-2/sqrt(1000),+2/sqrt(1000)) */

               uni01x21=ranuni(21);
```

```
                uni01x22=ranuni(22);
                uni01x23=ranuni(23);
                uni01x31=ranuni(21);
                uni01x32=ranuni(22);
                uni01x33=ranuni(23);

                uniabx21=2*(2/rmil)*uni01X21-(2/rmil);
                uniabx22=2*(2/rmil)*uni01X22-(2/rmil);
                uniabx23=2*(2/rmil)*uni01X23-(2/rmil);
                uniabx31=2*(2/rmil)*uni01X31-(2/rmil);
                uniabx32=2*(2/rmil)*uni01X32-(2/rmil);
                uniabx33=2*(2/rmil)*uni01X33-(2/rmil);

        /* perturbando as coordenadas x e y */

                x11f=x11;
                x12f=x12;
                x13f=x13;

                x21f=x21+uniabX21;
                x22f=x22+uniabX22;
                x23f=x23+uniabX23;
                x31f=x31+uniabX31;
                x32f=x32+uniabX32;
                x33f=x33+uniabX33;
run;

*PROC PRINT DATA=TESE.PERTURBA;
*     var X11F X12F X13F X21F X22F X23F X31F X32F X33F;
*RUN;

*_____________________FORMANDO OS GRUPOS_____________________________;

data tese.numero;
     do i=1 to 3;
        do j=1 to 1000;
            grupo=i;
           output;
        end;
     end;
run;

data tese.amostra1;
     set tese.perturba;
       drop x21 x22 x23 x31 x32 x33;
       x1=x11;
       x2=x12;
       x3=x13;
       keep x1 x2 x3;
run;

data tese.amostra2;
     set tese.perturba;
       drop x11 x12 x13 x31 x32 x33;
       x1=x21;
       x2=x22;
       x3=x23;
       keep x1 x2 x3;
run;

data tese.amostra3;
     set tese.perturba;
       drop x11 x12 x13 x21 x22 x23;
```

```
       x1=x31;
       x2=x32;
       x3=x33;
       keep x1 x2 x3;
run;

data tese.amostras;
     set tese.amostra1 tese.amostra2 tese.amostra3;
run;

*_____________________DATA SET FINAL COM 3 GRUPOS_____________________;

data tese.grupos;
     merge tese.amostras tese.numero;
run;

proc print data=tese.grupos;
     var grupo x1 x2 x3;
run;

proc corr data=tese.grupos out=resumo;
     var x1 x2 x3;
     by grupo;
run;

proc print data=resumo;
run;

options ps=30 ls=78 nodate nonumber;
  /*************************************************************/
  /*     Metodo da Analise Discriminante Canonica de Fisher      */
  /*************************************************************/

proc candisc data=tese.grupos out=tese.out11 outstat=tese.out12 distance anova
             wsscp wcov psscp pcov tsscp tcov bsscp bcov;
     var x1 x2 x3;
     class grupo;
run;

title;
        /*******************************************/
        /*    Gráfico das Variáveis Canônicas       */
        /*******************************************/

proc gplot data=tese.out11;
     plot can2*can1=grupo;
     format grupo;
run;
quit;

title;

        /*****************************************************/
        /* Converte a saida do proc candisc num data set SAS */
        /*         para fazer os graficos                    */
        /*****************************************************/

title;
data teste;
     set tese.out12;
     if _type_ in ('STRUCTUR', 'BSTRUCT', 'PSTRUCT', 'SCORE', 'PSCORE',
'RAWSCORE');
run;
```

```
proc sort data=teste;
     by _NAME_;
run;

data teste1;
     set teste;
     retain x1_T x2_T x3_T x1_E x2_E x3_E
            x1_D x2_D x3_D x_sc y_sc z_sc
            x_psc y_psc z_psc x_rs y_rs z_rs;
     drop x1 x2  x3 grupo _type_;

     if _type_='STRUCTUR' then do;
        x1_T=x1;
        x2_T=x2;
        x3_T=x3;
     end;

     if _type_='BSTRUCT' then do;
        x1_E=x1;
        x2_E=x2;
        x3_E=x3;
     end;

     if _type_='PSTRUCT' then do;
        x1_D=x1;
        x2_D=x2;
        x3_D=x3;
     end;

     if _type_='SCORE' then do;
        x_sc=x1;
        y_sc=x2;
        z_sc=x3;
     end;

     if _type_='PSCORE' then do;
        x_psc=x1;
        y_psc=x2;
        z_psc=x3;
     end;

     if _type_='RAWSCORE' then do;
        x_rs=x1;
        y_rs=x2;
        z_rs=x3;
        output;
     end;
run;

title;
*proc print data=teste1;
*     id _NAME_;
*run;

proc transpose data=teste1 out=saida;
run;

        /*************************/
        /*    Macro - Grafico    */
        /*************************/

%macro grafico(nvar,grupo);
```

```
data saida1;
  set saida;
  drop n i;
  n=&nvar;
  do i=0 to 5;
    if (i*n+1 le _n_ le (i+1)*n) then classe=i+1;
  end;
run;

data auxiliar;
       set saida1;
         if classe=&grupo;
         xvar=can1; yvar=can2; name=_name_;

data anno;
       set auxiliar;
        retain when 'a';
        length function text style $8;

/* Determina o valor de pi */
        pi=arcos(-1);

/* Determina a longitude do lado mais longe e dos lados adjacentes do triangulo
reto */
        if xvar>0 and yvar>0 then do;
                newx=xvar;
                newy=yvar;
           end;
        if xvar<0 and yvar>0 then do;
                newx=abs(xvar - 0);
                newy=yvar-0;
           end;
        if xvar<0 and yvar<0 then do;
                newx=abs(xvar-0);
                newy=abs(yvar-0);
           end;
        if xvar>0 and yvar<0 then do;
                newx=xvar;
                newy=abs(yvar-0);
           end;

/* Determina o angulo adjacente do triangulo reto em radianes  */

        radA=atan(newy/newx);

/* Determina o angulo adjacente em graus */

        A=(radA*180)/pi;

/* Determina o angulo oposto do triangulo reto em graus */

        B=90-A;

/* Determina a longitude da hipotenusa do triangulo reto   */

        c=sqrt(newx**2 + newy**2);

/* Ajusta as cordenadas Y tal que as cabeças das setas nao sobre-escrevam
simbolos */

        adjy=sin (radA)*(c-.1);
```

```
/* Ajusta as cordenadas X tal que as cabeças das setas nao sobre-escrevam
simbolos */

        adjx=cos (radA)*(c-.1);

/* Especifica condicionalmente o angulo e as coordenadas X/Y  */

        if xvar>0 and yvar>0 then do; adjx=adjx; ang=-B; end;
        if xvar<0 and yvar>0 then do; adjx=-adjx; ang=B; end;
        if xvar<0 and yvar<0 then do; adjx=-adjx; adjy=-adjy; ang=90+A; end;
        if xvar>0 and yvar<0 then do; adjx=adjx; adjy=-adjy; ang=-90-A; end;
        if xvar=0 and yvar>0 then do; adjx=xvar; adjy=yvar-.1; ang=0; end;
        if xvar=0 and yvar<0 then do; adjx=xvar; adjy=yvar+.1; ang=180; end;
        if xvar>0 and yvar=0 then do; adjx=xvar-.1; adjy=yvar; ang=-90; end;
        if xvar<0 and yvar=0 then do; adjx=xvar+.1; adjy=yvar; ang=90; end;

/* Gera ramas de setas começando do (0,0) */

        function='move'; xsys='2'; ysys='2'; x=0; y=0; output;
        function='draw'; xsys='2'; ysys='2'; x=adjx; y=adjy; output;

/* Gera as cabeças das setas*/

        function='label'; xsys='2'; ysys='2'; x=adjx; y=adjy; output;
                style='marker'; text='C';
                angle=ang; position='5'; output;

/* Disenha a circunferencia  */

        function='pie'; xsys='2'; ysys='2'; hsys='2'; x=0; y=0;
                angle=0; rotate=360; size=1;
                style='empty'; when='a'; output;

/* Label da circunferencia */

        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
           style='simplex';
                x=1; y=0; text='1'; position='3'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=-1; y=0; text='1'; position='1'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=0; y=1; text='1'; position='3'; output;
        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=0; y=-1; text='1'; position='9'; output;

/* Pontos Labels  */

        function='label'; xsys='2'; ysys='2'; hsys='4'; when='a';
                x=xvar; y=yvar; text=right(name);
                if xvar>0 then do;
                     position='3'; text=right(name);
                     end;
                if xvar<0 then do;
                     position='1'; text=left(name);
                     end;
                output;
run;

/* Simbolos */

        symbol v=star i=none c=black;
        axis1 order=(-2 to 2 by 1) length=4 in
```

```
  /* tirando os eixos */
       style=0 value=none label=none major=none minor=none;

proc gplot data=auxiliar;
 plot yvar*xvar/ annotate=anno
                vref=0 lvref=1 cvref=black
                href=0 lhref=1 chref=black
                vaxis=axis1 haxis=axis1;
run;
quit;

%mend;

%grafico(3,1);
%grafico(3,2);
%grafico(3,3);
quit;
title;
```

UNICAMP
BIBLIOTECA CENTRAL
SEÇÃO CIRCULANTE