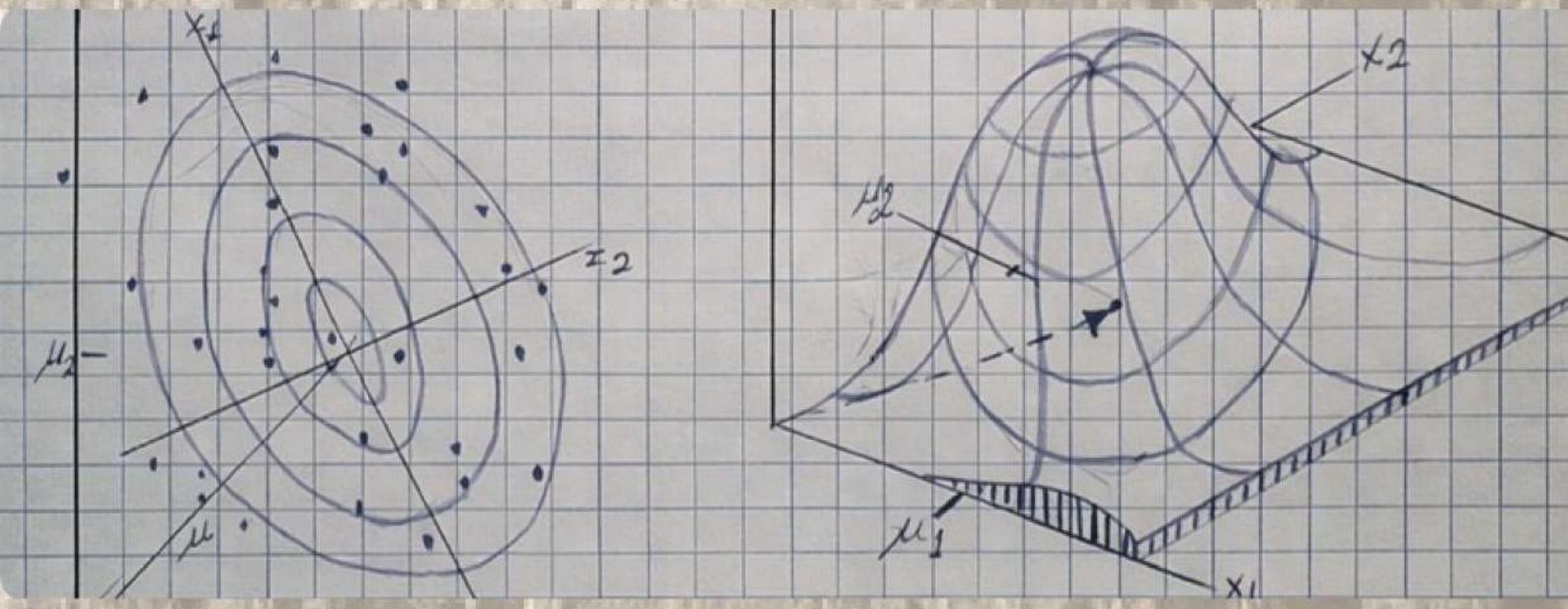
ANÁLISE MULTIVARIADA

PRÍNCIPIOS E PROCEDIMENTOS NA CIÊNCIA FLORESTAL



FLÁVIO FELIPE KIRCHNER
HENRIQUE SOARES KOEHLER
ALEXANDRE BEHLING

2026

Universidade Federal do Paraná
Setor de Ciências Agrárias
Programa de Pós-graduação em Engenharia Florestal

ANÁLISE MULTIVARIADA.
PRINCÍPIOS E PROCEDIMENTOS NA CIÊNCIA FLORESTAL.

Flávio Felipe Kirchner
Henrique Soares Koehler
Alexandre Behling

Curitiba, 2026
Editado pelos autores.

Os dados, funções e códigos computacionais utilizados nos exemplos e aplicações deste livro estão disponibilizados em repositório no GitHub, nas linguagens MATLAB e JULIA. Os materiais podem ser utilizados livremente para reprodução das análises apresentadas, estudo dos procedimentos estatísticos, adaptação dos códigos a diferentes conjuntos de dados e desenvolvimento de novas aplicações. O repositório poderá ser atualizado com correções, aprimoramentos e novos materiais relacionados ao livro.

Todas as figuras e tabelas apresentadas neste livro foram elaboradas pelos autores. Quando necessário, citada a fonte.
Foto Capa: Povoamento da espécie acácia-negra no estado do Rio Grande do Sul, com idade próxima aos 7 anos. Fonte: os autores. A imagem da capa frontal representa o interior do povoamento de acácia-negra, simbolizando a mensuração individual das árvores e a análise de suas características. A imagem da capa externa posterior, por sua vez, apresenta uma visão externa do povoamento, representando a perspectiva global da análise multivariada, que permite estudar simultaneamente múltiplas variáveis e observações, proporcionando uma compreensão integrada do conjunto e com mais clareza nas conclusões.

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ / SISTEMA DE BIBLIOTECAS
BIBLIOTECA DE CIÊNCIAS FLORESTAIS E DA MADEIRA

Kirchner, Flávio Felipe
Análise multivariada: princípios e procedimentos na ciência florestal / Flávio Felipe Kirchner, Henrique Soares Koehler, Alexandre Behling. - Curitiba : Ed. dos Autores, 2026.
1 recurso online : PDF

ISBN 978-65-02-38665-1

1. Análise multivariada. 2. Análise multivariada - Processamento de dados. 3. Engenharia florestal - Métodos estatísticos. 4. Florestas - Métodos estatísticos. 5. MATLAB (programa de computador). 6. Julia (Linguagem de programação de computador). I. Koehler, Henrique Soares. II. Behling, Alexandre. III. Universidade Federal do Paraná. Setor de Ciências Agrárias. Programa de Pós-graduação em Engenharia Florestal. IV. Título.

CDD - 519.535
CDU - 519.237
630:519.237:004.43
CDU/GFDC - 634.0.51

Bibliotecária: Marilene do Rocio Veiga CRB9/424

Qualquer parte desta publicação pode ser reproduzida, desde que citada a fonte.

Este trabalho é dedicado a Celinha, minha mulher,
e as minhas filhas Bruna e Bianca.
(ao Laurêncio também)!
Laurêncio: Wayne L. MYERS a great guy!
FLÁVIO FELIPE KIRCHNER

Aos meus filhos Barbara, Mariana, Pedro, Henrique, Lígia e Maria Vitória. E para minha querida companheira Ana 'Bambina' Eliete pelo apoio e entusiasmo quando da elaboração deste livro.
HENRIQUE SOARES KOEHLER

Benjamin e Juliana. Bóris!
Também à minha querida família de Tucunduva.
Aos meus professores e orientadores!
Ao Rogério Bamberg.
ALEXANDRE BEHLING

CONTEÚDO RESUMIDO

1 VISÃO GERAL DA ESTATÍSTICA MULTIVARIADA
2 REVISÃO DE MATRIZES
3 COMPARAÇÃO DE MÉDIAS MULTIVARIADAS
4 ANÁLISE DE COMPONENTES PRINCIPAIS
5 ANÁLISE DE FATORES
6 ANÁLISE DE VARIÂNCIA MULTIVARIADA
7 ANÁLISE DE DISCRIMINANTES
8 ANÁLISE DE CONGLOMERADOS
9 ANÁLISE DE CORRELAÇÃO CANÔNICA

PARA REFLETIR:

Nós temos que fazer o mais simples possível, não o mais fácil (Albert Einstein).

Cookery has become an art, a noble science; cooks are gentlemen.

TITUS LIVIUS, AB URBE CONDITIA XXXIX.vi (Anatomia da Melancolia 1.2.2.2).

AGRADECIMENTOS ESPECIAIS

Agradecer é para o nosso orientador WAYNE L. MYERS a quem devemos toda a base científica e metodológica desse livro.

Também agradecer ao Prof. SYLVIO PELLICO NETTO e Prof. ROBERTO TUYOSHI HOSOKAWA durante a nossa graduação e pós-graduação.

Também ao Prof. JOSÉ BITTENCOURT DE ANDRADE (*in memoriam*).

Aos professores Dr. ANSELMO CHAVES NETO e Dr. JAIR MENDES MARQUES que são referências na disciplina e em pesquisas de Estatística Multivariada.

PREFÁCIO I

Quando encontrava-me realizando meu doutorado na Universidade Albert Ludwig de Freiburg i.Br. em 1978, nos vários contatos que eu mantive com o Professor Dr. Walter Schöpfer, lembro-me de ele haver mencionado sobre um novo campo da ciência estatística em franca expansão, denominada de Análise Multivariada, e foi desta maneira que eu passei a me interagir com este novo campo de conhecimento, que nos permite tratar estatisticamente muitas variáveis simultaneamente.

Este campo do conhecimento teve início de maneira formal no Programa de Pós-graduação em Engenharia Florestal da UFPR em 2001, quando uma disciplina foi criada e ofertada pelo Professor Dr. Anselmo Chaves Neto, do curso de Matemática da UFPR. Posteriormente, a assumiu o Dr. Maurício Koubay do Amaral, do curso de Matemática da UTFPR e, finalmente, a assumiu os Professores Henrique Soares Koehler e Alexandre Behling, do curso de Engenharia Florestal da UFPR, e atuantes até o momento.

A obra ora apresentada é o resultado de um grande esforço dos Professores Dr. Flávio Felipe Kirchner, Dr. Henrique Soares Koehler e Dr. Alexandre Behling. Ela foi escrita visando disponibilizar as técnicas multivariadas para aplicação na Engenharia Florestal. Os autores se preocuparam em apresentá-la com os mais importantes métodos multivariados, com destaque para a Análise de Componentes Principais, Análise de Fatores, Análise de Variância Multivariada, Análise de Discriminantes, Análise de Conglomerados e Análise de Correlação Canônica, todos eles elucidados com exemplos florestais. Dada a complexidade dos diferentes métodos apresentados, a compreensão da natureza de um conjunto de dados multivariados provenientes da área florestal é fundamental para a escolha de um método e de uma análise apropriados. Os autores salientaram, também, a importância de se interpretar apropriadamente a essência dos resultados obtidos, após o devido processamento dos dados amostrais.

Esta obra está sendo disponibilizada para comunidade florestal brasileira de engenheiros florestais, de professores, de estudantes e de pesquisadores, que tratam variáveis multivariadas de maneira simultânea, com amplo conteúdo teórico e prático, tanto no âmbito do poder público, como das empresas florestais do país.

Os autores, todos professores, ou aposentados ou da ativa, na Universidade Federal do Paraná, compõem uma equipe especializada no campo da estatística teórica e aplicada conseguiram reunir um excelente conteúdo de conhecimentos, elucidados com trabalhos científicos e práticos efetuados em povoamentos florestais do estado do Paraná. Esta obra foi preparada visando atender primordialmente aos estudantes e aos engenheiros florestais recém-formados, que demandam a compreensão mais clara de metodologias estatísticas mais complexas, por meio de exemplos cuidadosamente escolhidos para ilustrá-las. Ainda, mesmo para os engenheiros florestais e pesquisadores mais experientes na aplicação desses métodos estatísticos, algumas inovações estão apresentadas, como a solução de todos os métodos utilizando-se o *software* desenvolvido em linguagem Julia para facilitar a obtenção dos resultados, bem como para abreviar o tempo e assegurar a qualidade do processamento dos dados.

Os autores destacam, com ampla visão, sobre os problemas mais típicos e práticos em que a estatística multivariada pode ser empregada nos diversos campos da ciência florestal.

Na Biometria, é ressaltado a importância de se identificar as inúmeras medições que podem ser efetuadas individualmente em um conjunto de árvores de uma ou mais espécies, para as avalições estatísticas, concernentes ao número apropriado de variáveis

e de medições a serem feitas em cada uma delas, posições mais adequadas para obtenção dessas medições e detecção do seu grau de heterogeneidade.
Na Silvicultura, é ressaltada a possibilidade de se comparar a produtividade de espécies em diferentes áreas de plantio ou em estratos identificados na floresta nativa.

Na Experimentação florestal, principalmente nos arranjos experimentais mais complexos, como no delineamento com arranjo fatorial, com diferentes números de repetições por tratamento, será necessário aplicar técnicas de equalizações para permitir a avaliação de significância dos efeitos principais e para a obtenção de correlações entre as variáveis medidas para cada efeito principal.

Na Economia florestal, as empresas precisam constantemente avaliar a produção anual em suas diferentes modalidades produtivas, bem como sua produtividade, taxas de juros, competitividade, nível qualitativo da mão de obra, entre outras. O desafio será desenvolver um índice que as represente bem e sirva para modelar o seu projeto de desenvolvimento estratégico.

Na Avaliação acadêmica, se for desejável estabelecer uma classificação de alunos por estrita ordem de mérito, uma consideração mais ampla dessa busca cresce em complexidade e carece de objetividade na escolha de cada variável quantitativa usada para representar os itens incluídos.

Na Indústria moveleira, as indústrias de móveis, ao produzi-los, quase sempre avalia a capacidade de atendimento à sua ampla clientela, com produtos de alta qualidade e preços acessíveis. Para tanto, precisa considerar as diversas madeiras disponíveis no mercado, a capacidade de acabamento qualitativo dos seus produtos, ter uma percepção de sua durabilidade, dispor de facilidades para as montagens, entre outras exigências. Esse conjunto de variáveis estruturais visa atingir no processo produtivo a máxima satisfação dos clientes a um mínimo custo.

Na Fitossanidade florestal, madeiras são muitas vezes tratadas para se avaliar o efeito de controle da ação de insetos predadores. Esta ação é efetuada via amostragem e será imprescindível criar um grupo de controle, pois, somente dessa maneira, poder-se-á conseguir uma avaliação de eficácia do tratamento efetuado, bem como do grau e da extensão das partes afetadas.

Nas Propriedades físicas da madeira, corpos de prova são normalmente testados para avaliar suas propriedades físicas, tais como: deformação, dureza, flexão, elasticidade e resistência, visando detectar efeitos sistemáticos e estabelecer relações funcionais que possibilitem estimá-los em função das propriedades físicas estudadas, bem como procedimentos de simulação para testar diferentes usos.

Na Sociologia florestal, populações citadinas e rurais podem ser amostradas, visando conhecer suas opiniões sobre impactos ambientais oriundos da ação antrópica. Por meio de entrevistas, pode-se coletar informações sobre tempo de moradia no local, importância familiar na área, nível de escolaridade atingido pela família e grau de conhecimento sobre diversos aspectos da natureza. O desafio será desenvolver um índice que as represente, bem como avaliar o grau de percepção e de comprometimento dessas populações com a conservação ambiental.

Ao prefaciar esta obra, o que muito me impressionou foi a grande contribuição ora disponibilizada pelos autores à comunidade florestal brasileira. Iniciativas como esta, com expressivo conteúdo científico e técnico, permitirão disseminar conhecimentos úteis à sociedade, e revela a perfeita sintonia da Universidade com a população que a mantém.

Ao concluir a leitura desta obra, sinto-me engrandecido em novos conhecimentos, e estou consciente de que ela muito contribuirá para a formação dos profissionais florestais e, também, para o aperfeiçoamento daqueles que já trabalham na condução de experimentos florestais. Eu apreciei muito a maneira que o texto foi escrito, em bom

português, e muito apropriadamente ilustrado. Tenho convicção de que esta obra será muito apreciada por estudantes, profissionais, professores e empreendedores florestais de todo o país.

Eu parabenizo os autores pela sua dedicação e esforço para concluir esta obra com sucesso. Ressalto, nesta oportunidade, meu grande apreço aos prezados autores, professores e amigos Flávio Felipe Kirchner, Henrique Soares Koehler e Alexandre Behling, com congratulações pela enorme dedicação, perseverança e tenacidade para concluí-la com foco objetivo e prático, e com a qualidade finalmente atingida.

Curitiba, 14 de julho de 2026.

Prof. Dr. Sylvio Péllico Netto

PREFÁCIO II

O Conteúdo do livro "ANÁLISE MULTIVARIADA" dos autores Prof. Dr. Flávio Felipe Kirchner, Prof. Dr. Henrique Soares Koehler e Prof. Dr. Alexandre Behling é um texto para analisar as múltiplas e complexas relações dos eventos de naturezas físicas, químicas, biológicas e sociais. É uma obra referencial de absoluta precisão nas pesquisas globais e no limiar da última fronteira de conhecimentos científicos.

Fazendo um passeio pela história da sociedade humana, pode ser constatada a presença da matemática empírica como a contagem para quantificar o uso de recursos naturais da Biosfera.

Cedo surgiu a necessidade de armazenar alimentos pelas primitivos homens na forma de catação de frutas, nozes, raízes e sementes para a sua segurança alimentar.

Foi manifestação da primeira economia, a poupança seguida de escambo, empréstimos e juros.

A caça de proteínas animais no início era simples apropriação dos restos alimentares de outros predadores. Contudo, com o domínio do uso do fogo, aliado de instrumentos perfurantes construídos com rochas, bronzes e ferros surgiu o trabalho coletivo de grandes caças.

Assim mostram os vários desenhos de registros arqueológicos nas rochas de cavernas. A sociedade humana era itinerante.

Mas a domesticação de animais e vegetais acabou com as perigosas viagens e surge então o capital "terra" como meio de produzir alimentos e também a sua posse significava poder das lideranças construindo então o "ESTADO".

A evolução demográfica começa acelerar no seu crescimento e passa a ser de linear para exponencial com a facilidade induzida pela descoberta de rodas e alavancas para mover e transportar grandes massas.

Então surge a estatística, com previsão catastrófica, feitas por ecólogos e biólogos sobre o colapso da sociedade demonstrada pela assíntota vertical da demografia.

Contudo no século passado, o expoente em matemática e biometria MICHEL PRODAN provou com cálculos usando equações diferenciais e integrais através de álgebra avançada, que tal apocalipse não ocorreria de imediato, mas previu-se que com modelagens de dinâmica físicas, químicas, biológicas e sociais, a disponibilidade de quantitativo dos recursos naturais da Biosfera tenderia evolução à evolução demográfica em forma de SIGMÓIDE.

Assim foi calculado o primeiro ponto de inflexão da função cuja derivada indicou o que os florestais conhecem como máximo incremento corrente.

O ano em que ocorreu este máximo na demografia indicou o início de rendimento decrescente de recursos naturais determinado pelas leis formuladas por KÖNIG e FAUSTMANN.

Ou seja, a população humana continuaria crescendo, mas com taxa cada vez menor devido a progressiva diminuição de proteínas animais e vegetais, seguindo até o máximo incremento médio quando se satura a capacidade da Biosfera em sustentar a massa humana.

Urge, portanto, planejar e racionar o uso de recursos naturais.

A complexa interrelação de vários eventos no crescimento de populações humanas já não se resolve por simples aritmética e cálculos percentuais.

É preciso elaborar e usar algoritmos de racionalização de recursos da Biosfera, que devem ser alimentados com resultados obtidos através de álgebras avançadas, biometria, multivariada, cálculo de probabilidades, erros amostrais etc.

Analisando a grosso modo, quais são nas principais variáveis que limitam o crescimento dos seres vivos, é preciso estudar o que seja um bem econômico.

A característica mais marcante de um bem econômico é a “escassez”.

Neste sentido, vale lembrar as pesquisas em biotério conduzidas pelos antropólogos, sociólogos, etc.

Num experimento foram propiciados todos os elementos necessários para o crescimento de uma população de roedores, mas com limitação de alimentos (calorias).

No limite crucial, iniciou o canibalismo em que recém-nascidos e os mais fracos eram consumidos.

No outro experimento foram fornecidos todos elementos e calorias (alimentos), mas com limitação de espaço.

O apocalipse conduziu os roedores a assassinarem os recém-nascidos, mas não consumiam como no canibalismo.

Isto demostra que os fatores cadeia de calorias e o meio ambiente (espaço) são vitais para a sobrevivência de biodiversidade.

Até antes da culminação do incremento corrente do crescimento da população humana, a água, o oxigênio e a fitomassa não eram considerados como bens econômicos porque não eram escassos.

Mas atualmente se nota o clamor da sociedade em preservar a qualidade da água potável, atmosfera despoluída e conservação da fitomassa.

Feliz e acertada foi a decisão dos autores, destinarem aplicativos avançados com “PRINCÍPIOS E PROCEDIMENTOS NA CIÊNCIA FLORESTAL”. A FITOMASSA FLORESTAL é a principal fonte de energia. A sua produção em grande escala na cadeia de calorias da BIOSFERA, movida por FÓTONS SOLARES, e sendo também a grande massa da Terra (depois de Litosfera, Hidrosfera e Atmosfera), sendo muito importante no equilíbrio dinâmico ambiental.

A sustentabilidade global da BIOSFERA TERRESTRE deve passar por atitude responsável da humanidade, para propiciar os benefícios energéticos e ambientais para gerações vindouras.

Neste sentido, sem dúvidas, o texto elaborado pelos Professores Doutores FLÁVIO FELIPE KIRCHNER, HENRIQUE SOARES KOEHLHER e ALEXANDRE BEHLING será de grande valia para pesquisadores em assuntos Globais da BIOSFERA TERRESTRE.

Curitiba, 22 de abril de 2026.

Prof. Dr. Dr. Honoris Causa Roberto T. Hosokawa

DADOS, FUNÇÕES E CÓDIGOS COMPUTACIONAIS

Os dados, funções e códigos computacionais utilizados nos exemplos e aplicações apresentados neste livro estão disponibilizados em um repositório no GitHub, nas linguagens de programação MATLAB (versão *TRIAL*) e JULIA.

A disponibilização desse material tem como objetivo facilitar a reprodução das análises apresentadas ao longo dos capítulos, bem como permitir que o leitor explore os procedimentos estatísticos, adapte os códigos aos seus próprios conjuntos de dados e aprofunde seus conhecimentos em análise multivariada.

O material pode ser acessado por meio do endereço:

https://github.com/renilsonlisboa/Analise_Multivariada_Principios_E_Procedimentos_Na_Ciencia_Florestal.git

Outra maneira é pesquisar pelo nome do livro no site oficial do GitHub:

https://github.com

A utilização da linguagem JULIA pode ser feita através do *Visual Studio Code* (VS Code), por meio da extensão oficial para JULIA. Nesse ambiente, é possível desenvolver, executar e testar os códigos diretamente no computador, além de organizar os arquivos, funções e conjuntos de dados utilizados nas análises.

O conteúdo do repositório poderá ser atualizado ao longo do tempo para incorporar correções, aprimoramentos e novos materiais relacionados ao livro.

Os autores agradecem especialmente ao Eng. Flo. Renilson Lisboa Junior pela disponibilização e compartilhamento de seu repositório no GitHub, contribuindo para a organização e disponibilização dos materiais computacionais utilizados neste livro.

SUMÁRIO

1 VISÃO GERAL DA ESTATÍSTICA MULTIVARIADA 21
1.1 O CONJUNTO DE DADOS MULTIVARIADOS..... 22
1.2 TIPOS DE DADOS..... 22
1.2.1 Variáveis contínuas e discretas..... 23
1.2.2 Dados agrupados..... 23
1.2.3 Variáveis dicotômicas..... 23
1.2.4 Subdivisões..... 23
1.2.5 Variáveis perdidas..... 24
1.2.6 Métodos não paramétricos e dados ordenados..... 24
1.2.7 Observações simultâneas..... 25
1.3 PROBLEMAS E MÉTODOS..... 25
1.3.1 Análise de Componentes Principais..... 25
1.3.2 Análise de Fatores..... 25
1.3.3 Análise de Correlação Canônica..... 26
1.3.4 Análise de Variância Multivariada (MANOVA)..... 26
1.3.5 Análise de Discriminantes Multivariada..... 26
1.3.6 Análise de Agrupamentos..... 27
1.4 O ESCOPO DA ANÁLISE MULTIVARIADA..... 27
1.4.1 Biometria..... 28
1.4.2 Silvicultura..... 28
1.4.3 Experimentação florestal..... 28
1.4.4 Economia florestal..... 29
1.4.5 Avaliação acadêmica..... 29
1.4.6 Indústria moveleira..... 29
1.4.7 Fitossanidade florestal..... 29
1.4.8 Propriedades físicas da madeira..... 29
1.4.9 Sociologia florestal..... 29
2 REVISÃO DE MATRIZES 30
2.1 VETORES E MATRIZES..... 30
2.1.1 Quantidade escalar..... 30
2.1.2 Rede..... 30
2.1.3 Vetor..... 30
2.1.4 Matriz..... 30
2.1.5 Ordem de uma matriz..... 30
2.1.6 Matriz quadrada..... 30
2.1.7 Matriz nula ou zero..... 31
2.1.8 Diagonal de uma matriz..... 31
2.1.9 Matriz diagonal..... 31
2.1.10 Mariz identidade..... 31
2.1.11 Triângulo superior da matriz..... 31
2.1.12 Triângulo inferior da matriz..... 31
2.1.13 Matriz simétrica..... 31

2.1.14 Matriz escalar..... 31
2.1.15 Igualdade de matrizes..... 32
2.1.16 Transposta de uma matriz..... 32
2.1.17 Adição de matrizes..... 32
2.1.18 Subtração de matriz..... 32
2.1.19 Multiplicação de vetores..... 32
2.1.20 Multiplicação de matrizes..... 32
2.2 EQUAÇÕES LINEARES SIMULTÂNEAS NA FORMA MATRICIAL.... 33
2.2.1 Determinantes..... 33
2.2.2 Menores..... 33
2.2.3 Cofatores..... 34
2.2.4 Expansão por meio dos cofatores..... 34
2.2.5 Singularidade..... 35
2.2.6 Duas propriedades dos determinantes..... 35
2.2.7 Grau de uma matriz..... 35
2.2.8 Matriz cofator..... 35
2.2.9 Matriz adjunta..... 35
2.3 MATRIZ INVERSA..... 35
2.3.1 Inversa por meio de transformações elementares..... 35
2.4 OUTRAS MATRIZES E OPERAÇÕES..... 36
2.4.1 Matriz ortogonal..... 36
2.4.2 Traço da matriz..... 36
2.4.3 Matrizes de transformação..... 36
2.4.4 Escalonamento ou normalização de vetores..... 36
2.4.5 Matriz modal..... 36
2.4.6 Matriz modal normalizada..... 36
2.4.7 Diagonalização de uma matriz..... 36
2.4.8 Matrizes similares..... 36
2.5 VALORES CARACTERÍSTICOS E VETORES CARACTERÍSTICOS (*EIGENVALUES* E *EIGENVECTORS*)..... 37
2.5.1 Propriedades dos valores e dos vetores característicos..... 37
2.5.2 Exemplo..... 37
2.6 EXEMPLO DOS COMANDOS DE ÁLGEBRA MATRICIAL NO MATLAB..... 39
2.7 EXEMPLO DOS COMANDOS DE ÁLGEBRA MATRICIAL NO JULIA..... 44
3 COMPARAÇÃO DE MÉDIAS MULTIVARIADAS **47**
3.1 DENSIDADE NORMAL MULTIVARIADA..... 47
3.1.1 Verificação da normalidade multivariada..... 49
3.2 VERIFICAÇÃO DA HOMOGENEIDADE DA MATRIZ VARIÂNCIA-COVARIÂNCIA..... 52
3.3 TESTE DE T2 DE HOTELLING..... 54
3.3.1 Teste para um vetor de médias μ..... 54
3.3.2 Comparação entre vetores médios de 2 populações..... 55

3.3.2.1 Vetores pareados..... 55
3.3.2.2 Vetores independentes com matrizes variância-covariância homogêneas..... 57
3.3.2.3 Vetores independentes com matrizes variância-covariância heterogêneas..... 58
3.4 ANÁLISE DE VARIÂNCIA MULTIVARIADA – MANOVA..... 60
3.5 SOLUÇÃO DO EXEMPLO 3.1: VERIFICAÇÃO DA NORMALIDADE MULTIVARIADA NO MATLAB E JULIA..... 60
3.6 SOLUÇÃO DO EXEMPLO 3.2: VERIFICAÇÃO DA NORMALIDADE MULTIVARIADA NO MATLAB E JULIA..... 62
3.7 OBTENÇÃO DA FIGURA 3.4 NO MATLAB E JULIA..... 65
3.8 SOLUÇÃO DO EXEMPLO 3.3, VERIFICAÇÃO DA HOMOGENEIDADE DA MATRIZ VARIÂNCIA-COVARIÂNCIA, NO MATLAB E NO JULIA..... 68
3.9 SOLUÇÃO DO EXEMPLO 3.4, TESTE PARA UM VETOR DE MÉDIAS μ, NO MATLAB E JULIA..... 73
3.10 SOLUÇÃO DO EXEMPLO 3.5, VETORES PAREADOS, NO MATLAB E JULIA..... 75
3.11 SOLUÇÃO DO EXEMPLO 3.6, VETORES INDEPENDENTES COM MATRIZES VARIÂNCIA-COVARIÂNCIA HOMOGÊNEAS, NO MATLAB E JULIA..... 78
3.12 SOLUÇÃO DO EXEMPLO 3.7, VETORES INDEPENDENTES COM MATRIZES VARIÂNCIA-COVARIÂNCIA HETEROGÊNEAS, NO MATLAB E JULIA..... 81
4 ANÁLISE DE COMPONENTES PRINCIPAIS **87**
4.1 TRANSFORMAÇÕES LINEARES..... 87
4.2 RELAÇÃO ENTRE MÉDIAS DAS VARIÁVEIS ORIGINAIS E TRANSFORMADAS..... 87
4.3 EFEITO DO DESLOCAMENTO DA MÉDIA NAS VARIÁVEIS ORIGINAIS SOBRE AS VARIÁVEIS TRANSFORMADAS..... 87
4.4 RELAÇÃO ENTRE A VARIÂNCIA DAS VARIÁVEIS ORIGINAIS E AS VARIÁVEIS TRANSFORMADAS..... 88
4.5 EFEITO EM DIVIDIR CADA VARIÁVEL ORIGINAL (X_j)POR UMA CONSTANTE Δ_j... 88
4.6 COVARIÂNCIA DAS VARIÁVEIS ORIGINAIS E TRANSFORMADAS..... 89
4.7 TRANSFORMAÇÕES DE VARIÁVEIS CORRELACIONADAS PARA NÃO CORRELACIONADAS..... 89
4.8 COMPONENTES PRINCIPAIS COMO O MELHOR AJUSTE k-DIMENSIONAL EM m DIMENSÕES..... 94
4.9 USOS E INTERPRETAÇÃO DA ANÁLISE DE COMPONENTES PRINCIPAIS..... 94
4.9.1 Introdução..... 94
4.9.2 Propriedades básicas..... 97
4.9.3 Terminologia e representação..... 98
4.9.4 Sequência operacional..... 99
4.9.5 Obtenção dos valores e vetores característicos de uma matriz..... 101
4.9.6 Teste de esfericidade de Bartlett..... 103

4.9.7 Um exemplo prático de ACP..... 103
4.9.7.1 Introdução..... 104
4.9.7.2 Procedimentos de cálculo..... 105
4.9.7.3 Discussão dos resultados..... 107
4.9.7.4 Interpretação do exemplo..... 107
4.9.7.5 O teste scree..... 111
4.10 EXEMPLO DE ANÁLISE DE COMPONENTES PRINCIPAIS NO MATLAB..... 112
4.11 EXEMPLO DE ANÁLISE DE COMPONENTES PRINCIPAIS NO JULIA..... 120
5 ANÁLISE DE FATORES **124**
5.1 INTRODUÇÃO..... 124
5.2 ANÁLISE DE FATORES..... 124
5.3 EXEMPLO DE ANÁLISE DE FATORES..... 129
5.4 ANÁLISE DE IMAGEM..... 132
5.5 EXEMPLO PRÁTICO DA ANÁLISE DE IMAGEM..... 132
5.6 EXEMPLO DE ANÁLISE DE FATORES NO MATLAB..... 138
5.7 EXEMPLO DE ANÁLISE DE IMAGEM NO MATLAB..... 150
5.8 EXEMPLO DE ANÁLISE DE FATORES NO JULIA..... 154
5.9 EXEMPLO DE ANÁLISE DE IMAGEM NO JULIA..... 159
6 ANÁLISE DE VARIÂNCIA MULTIVARIADA – MANOVA **161**
6.1 INTRODUÇÃO..... 161
6.2 OBTENÇÃO DA MANOVA POR MEIO DO CÁLCULO DOS DESVIOS..... 161
6.3 OBTENÇÃO DA MANOVA POR MEIO DE MODELOS DE REGRESSÃO LINEARES MÚLTIPLOS..... 163
6.4 CRITÉRIOS DE AVALIAÇÃO DA MANOVA - TESTE ESTATÍSTICO DE WILKS..... 165
6.5 EXEMPLO DE MANOVA COM VARIÁVEIS ENERGÉTICAS..... 166
6.6 EXEMPLO DE MANOVA COM VARIÁVEIS DE PRODUÇÃO E DE TECNOLOGIA DA MADEIRA..... 167
6.7 SOLUÇÃO DO EXEMPLO 6.5 DE ANÁLISE DA VARIÂNCIA MULTIVARIDA (MANOVA) NO MATLAB..... 168
6.8 SOLUÇÃO DO EXEMPLO 6.5 DE ANÁLISE DA VARIÂNCIA MULTIVARIDA (MANOVA) NO JULIA..... 176
6.9 SOLUÇÃO DO EXEMPLO 6.6, MANOVA PELO MÉTODO DE REGRESSÃO LINEAR, NO MATLAB E JULIA. USO DO MÉTODO DE MOORE PENROSE PARA OBTENÇÃO DA INVERSA DE X'X..... 181
6.10 SOLUÇÃO DO EXEMPLO 6.6, MANOVA PELO MÉTODO DOS DESVIOS, NO MATLAB E JULIA..... 186
7 ANÁLISE DE DISCRIMINANTES **193**
7.1 INTRODUÇÃO..... 193
7.2 ANÁLISE DE DISCRIMINANTE PARA DOIS GRUPOS..... 194
7.3 ANÁLISE DE DISCRIMINANTE PARA VÁRIOS GRUPOS..... 197
7.4 CLASSIFICAÇÃO DE UMA NOVA OBSERVAÇÃO E AVALIAÇÃO DAS TAXAS DE ACERTO..... 202

7.5 ANÁLISE DISCRIMINANTE PARA DUAS POPULAÇÕES NO MATLAB..... 204
7.6 ANÁLISE DISCRIMINANTE PARA VÁRIAS POPULAÇÕES NO MATLAB..... 207
7.7 ANÁLISE DISCRIMINANTE PARA DUAS POPULAÇÕES NO JULIA. 222
7.8 ANÁLISE DISCRIMINANTE PARA VÁRIAS POPULAÇÕES NO JULIA..... 224
8 ANÁLISE DE CONGLOMERADOS 231
8.1 INTRODUÇÃO..... 231
8.1.1 Escolha das unidades dos dados..... 231
8.1.2 Escolha das variáveis..... 231
8.1.3 Agrupamento de observações e/ou variáveis..... 232
8.1.4 Homogeneização das variáveis..... 232
8.1.5 Medidas de similaridade..... 233
8.1.6 Critério de agrupamento..... 237
8.1.7 Algoritmos e implementação..... 237
8.1.8 Número de conglomerados..... 237
8.1.9 Interpretação dos resultados..... 237
8.2 MÉTODOS DE AGRUPAMENTOS..... 238
8.2.1 Métodos de agrupamentos hierárquicos..... 238
8.2.2 Determinação do número de grupos e avaliação dos agrupamentos..... 240
8.2.3 Agrupamento não-hierárquico..... 240
8.3 EXEMPLO DE ANÁLISE DE CONGLOMERADOS (CLUSTER) NO MATLAB..... 243
8.4 ANÁLISE DE AGRUPAMENTO NÃO HIERÁRQUICO UTILZIANDO COM K-MEANS NO MATLAB..... 247
8.5 EXEMPLO DE ANÁLISE DE CONGLOMERADOS (CLUSTER) NO JULIA..... 249
8.6 ANÁLISE DE AGRUPAMENTO NÃO HIERÁRQUICO - MÉTODO DO K-MEANS NO JULIA..... 252
9 ANÁLISE DE CORRELAÇÃO CANÔNICA 255
9.1 INTRODUÇÃO..... 255
9.2 DESENVOLVIMENTO..... 256
9.3 APLICAÇÃO PRÁTICA..... 262
9.3.1 Introdução..... 262
9.3.2 O conjunto de dados..... 262
9.4 ANÁLISE DE CORRELAÇÃO CANÔNICA NO MATLAB..... 265
9.5 ANÁLISE DE CORRELAÇÃO CANÔNICA NO JULIA..... 277
REFERÊNCIAS 282
BIBLIOGRAFIA COMPLEMENTAR 284

Lista de Tabelas

1.1 Exemplos de dados multivariados em Engenharia Florestal..... 21
3.1 Pares de diâmetros e alturas obtidos em uma amostra de acácia-negra com 10 anos..... 50
3.2 Pares de diâmetros e alturas obtidos em uma amostra de acácia-negra com 10 anos..... 51
3.3 Volume com casca (vcc) e sem casca (vsc) para dois clones (A e B) da espécie acácia-negra..... 53
3.4 Valores das variáveis medidas obtidos para uma amostra em um povoamento de acácia-negra..... 55
3.5 Dados das dimensões comprimento (m), altura (m) e largura (m) de 12 pilhas de madeira, após o baldeio, mensuradas por dois métodos, em uma área de colheita de acácia-negra..... 56
3.6 Teores de materiais voláteis, carbono fixo e cinzas para amostras independentes de acácia-negra, obtida em um povoamento com 10 anos..... 58
3.7 Teores de Cu, Zn, Mn, Na e Fe para duas amostras independentes com matrizes variância-covariância heterogêneas, obtidos em um povoamento de acácia-negra com 7 anos..... 59
4.1 Valores característicos e porcentagem cumulativa da variação associada com os valores característicos da análise de componentes principais de medições em um viveiro de mudas de *Pinus taeda* com 4 anos..... 100
4.2 Conjunto de variáveis e observações usadas como exemplo..... 105
4.3 Variáveis que contribuem com a maior variação em cada componente principal (matriz de carregamento = *loadings*)..... 109
4.4 Fator padrão dos vetores característicos normalizados..... 110
4.5 Relação da variação das variáveis originais e transformadas derivadas da correlação..... 111
5.1 Matriz de carregamento onde se observa a correlação entre os valores originais e os valores transformados..... 134
5.2 Matriz variância-covariância e correlação da matriz H..... 135
5.3 Matriz de correlação da matriz H..... 135
5.4 Matriz transformação (matriz modal) para análise de imagem aplicada a rotação *varimax*..... 136
5.5 Matriz de carregamento (*loading matrix),* correlação entre os valores originais e os valores transformados obtidos pela análise fatorial..... 137
5.6 Matriz de carregamento (*loading matrix),* correlação entre os valores originais e os valores transformados obtidos pela análise de imagem..... 137
5.7 Matriz de transformação (matriz modal) sem o último componente..... 137
5.8 Matriz $A = T\ Q$ obtida pela análise de rotação *varimax*, pela matriz da Tabela 5.7..... 138
5.9 Matriz de transformação rotacionada obtida por *varimax* da análise de fatores componentes principais sem o último componente..... 138
5.10 Matriz $A = T\ Q$ obtida pela análise de rotação *varimax* por meio da análise de imagem sem o último componente..... 138
6.1 Resumo da análise de variância multivariada pelo método de cálculo das somas de quadrados de modo análogo ao caso univariado..... 162
6.2 Resumo da análise de variância multivariada..... 163
6.3 Transformação do valor de Wilks lambda para o valor de F..... 165

6.4 Resultados obtidos para quatro idades (anos) e cinco variáveis energéticas, com 4 repetições cada 166
6.5 Valores das variáveis mensuradas em um experimento com três tipos de adubação..... 167
7.1 Resultados obtidos para a biomassa do fuste e copa e cinco variáveis energéticas, com 21 repetições cada..... 196
7.2 Resultados obtidos para três idades (anos) e quatro variáveis energéticas, com 10 repetições cada..... 201
7.3 Matriz de confusão para dois grupos..... 203
9.1 Dados de 20 árvores de acácia-negra, onde foram medidas as variáveis biomassa do fuste (BF, em kg), biomassa da copa (BC, em kg), biomassa total (BT, em kg), diâmetro (DAP, em cm), a altura (H, em m), comprimento de copa (CC, em m) e massa específica do fuste (MEF, g/cm^3)..... 263
9.2 Resumo da análise de correlação canônica entre um grupo de variáveis de biomassa e outro grupo de variáveis dendrométrica e tecnológicas..... 263

Lista de Figuras

3.1 Distribuição normal univariada padronizada..... 48
3.2 Duas representações de densidade normal..... 49
3.3 Gráfico para avaliação da normalidade p variada..... 51
3.4 Densidades bivariadas normais..... 51
4.1 Representação de vários pontos em um eixo ortogonal bidimensional..... 91
4.2 Representação de cada ponto considerado como um vetor..... 91
4.3 Direções dos cossenos α e β para os novos eixos Y_u e X_u..... 92
4.4 Geometria da relação $Y \cos\alpha + X \cos\beta$..... 92
4.5 Resíduos medidos perpendicularmente para o eixo rotacionado..... 94
4.6 Alguns exemplos de eixos principais para distribuições bivariadas..... 95
4.7 Comparação entre a linha de regressão por mínimos quadrados e o eixo principal componentes principais..... 96
4.8 Fluxograma dos passos para obtenção dos componentes principais..... 108
4.9 Representação gráfica da linha *scree* dos *eingvalues*..... 111
7.1 Análise discriminante para dois grupos e duas variáveis originais. Dois conjuntos de elipses concêntricas representando a sobre posição bivariada para os dois grupos..... 195
7.2 Densidades normais associadas a cada grupo (fuste e copa)..... 197
8.1 A distância entre dois pontos..... 234
8.2 Exemplo de dendrograma..... 238
8.3 Grupos formados pela análise discriminante..... 242

1 VISÃO GERAL DA ESTATÍSTICA MULTIVARIADA

A análise Multivariada consiste em uma coleção de métodos que podem ser usados quando várias variáveis são obtidas de indivíduos ou de objetos em uma ou mais amostras. Na prática, um conjunto de dados multivariados são comuns, apesar de não serem sempre analisados com tal. O uso exclusivo de procedimentos univariados com essa classe de dados não é mais aceitável, dada a disponibilidade de técnicas multivariadas e poder de computação baratos para realizá-los.

Historicamente, a quantidade de aplicações de técnicas multivariadas se deu nas ciências comportamentais e biológicas. Mais recentemente, o interesse em métodos multivariados se espalhou para numerosos outros campos de investigação, entre eles educação, física, geologia, engenharia, literatura, entre outros e, atualmente, no campo da Engenharia Florestal. Na Tabela 1.1 estão listados alguns exemplos de observações multivariadas na Engenharia Florestal.

Tabela 1.1: Exemplos de dados multivariados em Engenharia Florestal.

UNIDADES	VARIÁVEIS / CARACTERÍSTICAS
Árvores	Altura, diâmetro, volume e biomassa
Árvores	Propriedades físicas e mecânicas
Operadores florestais	Variáveis ergonômicas
Solos	Propriedades físicas e químicas
Parcelas	Área basal, volume e biomassa por hectare
Mapas	Bandas, áreas, similaridades
Sementes	Peso, vigor, pureza, tamanho de sementes
Empresas	Custos, preços, demanda, oferta
Áreas de conservação	Índices fitogeográficos e fitossociológicos

O pré-requisito matemático para a leitura e utilização deste livro é álgebra matricial. Cálculo numérico não é utilizado. As ferramentas básicas de álgebra matricial são essenciais e foram apresentadas no Capítulo I, com a intenção de ser suficiente e completa, de modo que o leitor sem experiência prévia possa manipular matrizes ao nível requerido neste livro. Os pré-requisitos estatísticos são familiaridade básica com a distribuição normal, testes de t, intervalos de confiança, análise de regressão e análise de variância.

Normalmente os cálculos multivariados são extensos e trabalhosos, exigindo o uso de programas que permitam a manipulação de matrizes de grande porte. Para exemplificar o uso de programas de computador para a efetuação de análises multivariadas, são apresentados os comandos necessários para obtenção dos resultados nas linguagens MATLAB e JULIA. Os exemplos apresentados têm como objetivo somente mostrar como os cálculos multivariados podem ser efetuados com a utilização desses programas, sem que tenha havido a preocupação de torná-las rotinas programadas sofisticadas e generalizadas.

Este livro foi pensado visando a aplicação de técnicas multivariadas na Engenharia Florestal. Procurou-se, sempre que possível, apresentar exemplos florestais. Com isso em mente, pode-se dizer que três objetivos foram procurados quando do uso da estatística multivariada. O primeiro objetivo é apresentação das várias técnicas multivariadas, seus propósitos, suas condicionantes e suas limitações. O segundo objetivo é selecionar uma ou mais técnicas apropriadas para um determinado conjunto de dados multivariados, com aplicação na área florestal. O reconhecimento da natureza essencial de um conjunto de dados multivariados provenientes da área florestal é o primeiro passo

para uma análise apropriada. O terceiro objetivo é calcular e ser capaz de interpretar os resultados obtidos pelo computador para um conjunto de dados.

Uma sumarização geral dos procedimentos multivariados a serem utilizados dependendo das características do conjunto de dados está apresentada abaixo.

Uma amostra - um conjunto de variáveis

- Teste da hipótese de que as médias possuem um valor especificado.
- Teste da hipótese de que as variáveis não são correlacionadas e têm uma matriz de variância comum.
- Encontrar um conjunto de combinações lineares das variáveis originais que sumarizam a maior parte da variação dos dados – PCA.
- Expressar as variáveis originais como uma função de um conjunto menor de variáveis subjacentes que assimilam as variáveis originais e suas intercorrelações – FA.

Uma amostra - dois conjuntos de variáveis

- Determinar o número, o tamanho e a natureza da relação entre dois conjuntos de variáveis – CC.
- Encontrar um modelo para predizer um conjunto de variáveis a partir de outro conjunto – MMR.

Duas amostras - um conjunto de variáveis

- Comparar as médias das variáveis entre as duas amostras – T^2.
- Encontrar uma combinação linear das variáveis que melhor separam as duas amostras – DA.
- Encontrar uma função das variáveis que acuradamente alocam uma nova observação em um dos dois grupos – CA

Três ou mais amostras - um conjunto de variáveis

- Comparar as médias das variáveis entre os grupos – MANOVA.
- Encontrar uma combinação linear das variáveis que melhor separam mais do que dois grupos – DA.
- Encontrar uma função das variáveis que acuradamente alocam uma nova observação em um dos grupos – CA.

1.1 O CONJUNTO DE DADOS MULTIVARIADOS

Supondo que $\underline{m}$ características ou variáveis foram medidas em $\underline{n}$ unidades observadas, o conjunto de dados resultante toma a forma de uma matriz, como segue:

		Respostas, variáveis, características				
		$\underline{1}$	$\underline{2}$	$\underline{3}$	$\cdots$	$\underline{m}$
Subjetos,	$\underline{1}$	x_{11}	x_{12}	x_{13}	$\cdots$	x_{1m}
indivíduos,	$\underline{2}$	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2m}
unidades	$\underline{3}$	x_{31}	x_{32}	x_{33}	$\cdots$	x_{3m}
observadas	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
etc.	$\underline{n}$	x_{n1}	x_{n2}	x_{n3}	$\cdots$	x_{nm}

As colunas desta matriz são as variáveis ou características medidas ou observadas, e as linhas representam as unidades nas quais foram observadas. A terminologia para as linhas e colunas difere de um campo para outro e, assim, elas são muitas vezes trocadas.

Tal matriz pode ser examinada de duas maneiras. Primeiro, investigar as relações entre as variáveis ou características que constituem as colunas da matriz. Segundo, perguntar sobre a natureza das relações entre as unidades observadas que constituem as linhas da matriz. Contudo, os dois métodos não são mutuamente exclusivos, pois um tipo de análise quase sempre fornece um passo intermediário da outra.

1.2 TIPOS DE DADOS

A primeira tarefa, de crucial importância, é organizar os dados numa forma adequada para uso no computador. Diferentes tipos de variáveis podem estar envolvidas, e podem exigir diferentes tratamentos. De qualquer modo, elas precisam ser codificadas em uma forma numérica adequada para o programa particular a ser utilizado.

1.2.1 Variáveis contínuas e discretas

Uma variável contínua é aquela que pode assumir qualquer valor dentro de um certo domínio (o qual pode ser infinito). Os dados, é claro, nunca são contínuos neste sentido. As medidas são sempre feitas com uma exatidão limitada e sempre existe um certo grau de arredondamento. A diferença prática entre variáveis contínuas e discretas depende da chance que diferentes observações assumam o mesmo valor. Variáveis discretas assumem somente determinados valores dentro do seu domínio. Por exemplo, uma contagem que segue a distribuição de Poisson com média grande pode, para propósitos práticos, ser considerada contínua, pois somente uma pequena proporção das observações terá um único valor.

Variáveis contínuas normalmente não apresentam problemas. Geralmente elas podem ser consideradas com uma distribuição normal conjunta, tanto na sua forma original ou após uma transformação adequada e, quando os dados forem deste tipo (normal multivariada), podem ser analisados por métodos padrões.

1.2.2 Dados agrupados

Os dados podem ser agrupados, porque as medidas são tomadas com exatidão limitada, ou porque elas não são totalmente medidas e são substituídas por avaliações em classes como alto, médio, baixo, ou marcados em alguma escala arbitrária. Para muitos propósitos, este agrupamento não é importante. A avaliação pode ser substituída por marcações adequadas, dado que qualquer peso é considerado apropriado para as diferenças entre os grupos. Em muitos dos métodos multivariados clássicos o teorema do limite central justifica o seu tratamento como se apresentasse uma distribuição normal conjunta. Este ponto será discutido mais adiante.

Existe, contudo, algumas aplicações em que um maior cuidado é necessário. Este é o caso especialmente em análise de conglomerados. Se o objetivo de uma análise é encontrar um agrupamento útil ou significante dos dados, uma variável agrupada em classes pode exercer uma influência desproporcional no resultado.

1.2.3 Variáveis dicotômicas

Um importante caso especial de dados agrupados é o de observações que podem se compor de somente dois valores, usualmente marcados de 0 (zero) e 1 (um), e que, neste caso, são chamadas de variáveis binárias. Variáveis dicotômicas são aquelas em que o método de classificação para cada uma das divisões e subdivisões não contém mais que dois termos, tais como macho e fêmea, presença ou ausência de uma característica, e outras divisões em dois grupos. Geralmente, elas podem ser tratadas da mesma maneira como outras variáveis agrupadas. Assim como em aplicações univariadas, o erro padrão de uma proporção pode ser usado para testes de significância e intervalos de confiança como se a proporção fosse normalmente distribuída. No caso multivariado, o teorema do limite central muitas vezes justifica tratar dados binários como se seguissem uma distribuição normal multivariada.

1.2.4 Subdivisões

Geralmente uma observação pode ser incluída em uma classificação em grupos que não são logicamente ordenados. Por exemplo, suponha que as madeiras possam ser classificadas nas cores vermelha, branca ou amarela. Não existem razões lógicas para marcar a cor por meio de uma única variável. Marcar estes grupos como $0, 1, 2$ implicará que o grupo das madeiras brancas foi intermediário entre as outras duas e, a não ser que exista uma razão, isto pode causar confusão.

Uma maneira de se tratar grupos desordenados deste tipo é substituí-las por variáveis binárias. Estas três cores podem ser representadas por duas variáveis binárias, uma tomando o valor de 1 para as madeiras vermelhas e 0 para as brancas e amarelas, a outra 1 para as brancas e 0 para as vermelhas e amarelas.

Grupos não ordenados podem sempre ser substituídos por variáveis binárias (*dummy* ou categóricas), mas esta ferramenta fica um pouco incômoda se existirem muitos destes grupos, ou muitos grupos são envolvidos. O método também pode ser confuso se for usado em conjunto com uma análise que não leva em consideração as correlações entre as variáveis (alguns tipos de análise de conglomerados, por exemplo) pois um peso desproporcional é dado a uma subdivisão que resulte em muitos grupos. Um método alternativo é basear a análise numa medida de distância ou similaridade e, isto, é preferido na maioria dos casos, quando existem muitas observações deste tipo.

1.2.5 Variáveis perdidas

O problema de observações perdidas é comum na estatística univariada e, geralmente, pode ser facilmente tratada por métodos padrões. Um problema mais difícil, que é particularmente importante na taxonomia numérica, é de variáveis que não podem ser observadas quando uma outra variável toma um valor particular.

Suponha, por exemplo, que dados são disponíveis para um grupo de insetos e alguns dos dados são relacionados com a nervura das asas. Algumas espécies podem ser ápteras, e é conveniente representar esta condição por uma variável binária com o valor de 1 (um) para espécies aladas e 0 (zero) para as ápteras. Sempre que 0 (zero), todas as variáveis relacionadas com a asa são perdidas.

Não existe um caminho satisfatório para tratar desta questão. No geral, os métodos baseados nas medidas de distância ou similaridade são mais flexíveis do que os tipos clássicos de análise para tratar dados deste tipo, e isto é uma razão importante para a sua popularidade. Mesmo usando-se estes métodos, muitas dificuldades permanecem, e a medida de similaridade empregada necessita depender em alguma extensão na contribuição da importância relativa das variáveis. Se a condição de ser áptero é um caráter importante, que provavelmente aparecem uma vez no curso da evolução – será dado um peso considerável na medida de distância e as espécies ápteras improvavelmente serão subdivididas. Do outro lado, pode ser que a condição de ser áptero é uma aberração facilmente produzida pelo meio ambiente, e as espécies ápteras deveriam ser associadas com quaisquer grupos que elas mais assemelham-se. Neste caso, como em outros, métodos numéricos em taxonomia não podem ser completamente numéricos.

1.2.6 Métodos não paramétricos e dados ordenados

Na estatística multivariada, a distinção entre os métodos paramétricos e não-paramétricos é sempre complicada para os usuários. A maioria das técnicas clássicas depende da suposição de haver uma distribuição normal multivariada, mas é razoavelmente robusta para muitos tipos de dados e são geralmente usadas para variáveis agrupadas ou arbitrariamente marcadas. Métodos que dependem das unidades de distância e técnicas de agrupamentos, que produzam um dendrograma, são usualmente não paramétricos. No geral, contudo, o objetivo é produzir alguma descrição das características inerentes dos dados e a estimativa dos parâmetros torna-se um problema secundário.

Um tipo comum destes dados consiste inteiramente de variáveis ordenadas. Isto pode aparecer tanto porque os dados foram coletados desta forma, ou porque é detectado que as medições atuais têm pouco significado e nada é perdido substituindo-as por ordens.

Para muitos propósitos, sendo que as amostras são geralmente grandes, ordens rústicas podem ser tratadas como se fossem normalmente distribuídas. Isto apresenta a desvantagem que ordens dependem no tamanho da amostra e uma função discriminante baseada em ordens terá um conjunto diferente de valores, quando aplicados a uma outra amostra de tamanho diferente. Por esta razão, é sempre mais conveniente transformar as ordens, tanto usando a transformação de ordens de Fisher-Yates (Fisher e Yates, 1938), ou simplesmente dividir as ordens pelo tamanho da amostra.

1.2.7 Observações simultâneas

O uso de observações simultâneas na estatística univariada é um procedimento comum. Um exemplo típico é em experimentos com árvores, quando o diâmetro inicial de uma árvore é sempre usado para ajustar os diâmetros finais no encerramento do experimento e, então, conseguir um aumento na exatidão. A análise de covariância usada neste contexto é uma ferramenta útil e familiar e não apresenta qualquer dificuldade.

Na estatística multivariada, observações simultâneas também podem ser usadas. Por exemplo, em aplicações dendrométricas, observações de idade, espécie, posição fitossociológica, estado fitossanitário e, assim por diante, são sempre disponíveis bem como sinais e sintomas de doença nas árvores. Agora, se o problema é decidir se a normalidade deve ser considerada como um único grupo homogêneo de variáveis, ou se ela é constituída de duas ou mais variáveis separadas (um problema em análise de agrupamento), é claro que estas observações adicionais não devem ser consideradas como do mesmo tipo, como os sintomas, mas devem ser tratadas como observações simultâneas.

1.3 PROBLEMAS E MÉTODOS

O propósito de qualquer análise multivariada é obter uma avaliação do que existe dentro da estrutura da situação e das principais características dos dados. Isto pode tomar várias formas. Pode ser exigido dividir os dados em grupos, ou em decidir que um grupo conhecido seja associado com as medições. Alternativamente, será possível ajustar algum modelo matemático apropriado para a obtenção dos resultados. Pode ser que as medições sejam de dois tipos, exigindo associar uma com a outra. Todas estas situações necessitam diferentes tipos de análises e é importante saber exatamente quais questões serão formuladas, e que tipo de análise pode respondê-las.

1.3.1 Análise de Componentes Principais

O tipo mais simples de tratamento que se deseja realizar nas variáveis é o de transformar um conjunto de m variáveis correlacionadas em um novo conjunto de n variáveis não-correlacionadas, na qual as novas variáveis são combinações lineares das variáveis originais. Isto é chamado de análise de componentes principais.

A transformação do componente principal é derivada de tal maneira que cada nova variável absorve a maior quantidade possível da variação remanescente. Sob esta estratégia é bem possível que relativamente poucas variáveis transformadas explicarão a maior variabilidade existente no conjunto de dados. Neste caso, será possível reduzir a dimensão do conjunto sem uma apreciável perda de informação em eliminando-se as variáveis transformadas, que explicam pouca variação do conjunto de dados.

1.3.2 Análise de Fatores

Em algumas situações pode-se visualizar que $\underline{m}$ variáveis originais possam ser explicadas em termos de poucos fatores comuns básicos, que reproduzem as correlações existentes, enquanto outros fatores únicos podem explicar a completa falta de correlação.

Em tais casos, os fatores comuns básicos serão de central importância para o processo em estudo. A tarefa de separar fatores comuns dos fatores únicos é o principal objetivo da análise de fatores.

A análise de fatores foi desenvolvida principalmente no contexto da procura de fatores básicos de inteligência, que pudessem ser usados para explicar o desempenho de indivíduos em testes. Infelizmente, esta tarefa de separar os componentes comuns dos componentes únicos, para um grupo de variáveis, provou não ser uma tarefa fácil. A dificuldade em criar fatores exige grande conhecimento e experiência por parte do usuário deste tipo de análise.

1.3.3 Análise de Correlação Canônica

Em situações que envolvem relações entre variáveis, as relações existentes podem ser naturalmente separadas em dois conjuntos. Por exemplo, pode-se pensar num conjunto de variáveis "dependentes" como biomassa e volume do fuste da árvore, e um conjunto de variáveis "independentes" como diâmetro e altura. Também pode-se desejar reduzir cada conjunto de variáveis originais para uma só variável, composta de tal modo a conter as variáveis que mais se relacionam, ou obter uma combinação linear das variáveis "independentes" que visa estimar a combinação linear das variáveis "dependentes". Estes problemas caem dentro do escopo da análise de correlação canônica.

1.3.4 Análise de Variância Multivariada (MANOVA)

Quando da realização de análises, nas quais o interesse é centrado nas relações entre as unidades observadas, deve-se primeiro considerar a extensão multivariada da análise de variância.

Frequentemente, o analista mede várias variáveis em cada unidade experimental. O procedimento usual é analisar cada uma destas respostas separadamente. Este procedimento ignora as possíveis correlações existentes entre as variáveis medidas. A abordagem apropriada é usar a análise de variância multivariada (MANOVA), que é uma generalização simples da análise de variância e que visa aumentar o poder dos testes.

Como na situação univariada, estabelece-se um modelo linear aditivo para as respostas, quais são as variáveis categóricas e covariáveis. Na análise multivariada a variável dependente é substituída por um vetor de variável dependentes. Após ajustado o modelo máximo, formula-se e ajusta-se modelos reduzidos apropriados. Como na situação univariada, o objetivo é testar a adequabilidade dos modelos reduzidos em relação ao modelo máximo. Na análise de variância multivariada, a distribuição F não é mais usada como critério de comparação para o teste de hipóteses, e sim uma transformação de seus valores para outros critérios disponíveis para os testes de hipóteses.

1.3.5 Análise de Discriminantes Multivariada

Outro problema relativo às relações entre as unidades amostradas, considera a possibilidade de que as unidades observadas pertencem a "n" grupos distintos. O objetivo é estabelecer uma função com as variáveis que melhor discriminarão entre os grupos conhecidos. Isto é possível por meio da análise de discriminantes multivariada, em que funções discriminantes formadas pelas combinações lineares das variáveis mensuradas são estabelecidas e separarão melhor os grupos.

O termo grupo representa uma subpopulação ou amostra desta população. O primeiro objetivo é a descrição da separação dos grupos por meio de funções lineares (funções discriminantes) a partir das variáveis mensuradas, que são utilizadas para descrever ou explicar as diferenças entre dois ou mais grupos. Outro objetivo é identificar

a contribuição relativa de cada uma das variáveis na separação dos grupos e encontrar um plano ótimo, no qual as observações mensuradas podem ser projetadas para melhor ilustrar a configuração de cada grupo.

A predição ou alocação de observações aos grupos é realizada por meio de funções lineares que são obtidas e usadas para alocar uma unidade de amostra individual a um dos grupos. Os valores medidos no vetor de observações para um indivíduo são avaliados pelas funções de classificação para encontrar o grupo ao qual o indivíduo mais provavelmente pertence.

O resultado da análise é uma regra de alocação. Esta regra pode então ser usada para designar os indivíduos das amostras dentro do grupo mais apropriado. Como em qualquer regra de alocação existe uma probabilidade de classificação errada, e é importante ser capaz de estimar esta probabilidade, e escolher uma regra que gere o menor erro possível.

1.3.6 Análise de Agrupamentos

Na análise de agrupamentos procura-se encontrar padrões existentes no conjunto de dados pelo agrupamento multivariado das observações em grupos. O objetivo é encontrar um agrupamento ótimo, de tal modo que observações similares sejam alocadas em grupos similares, sendo que os grupos são diferentes si. Procura-se encontrar um agrupamento natural nos dados, que fazem sentido ao pesquisador.

A análise de agrupamentos difere fundamentalmente da análise discriminante. Na análise discriminante as observações são alocadas a um número de grupos pré-definidos. Na análise de agrupamentos tanto o número de grupos, quanto os grupos em si não são conhecidos previamente.

Dado um conjunto de N objetos (árvores, pessoas, espécies) cada medida em P diferentes variáveis, pode-se perguntar em que extensão existem grupos naturais entre os N objetos – grupos, que são similares nos seus valores nas P variáveis usadas para descrevê-los?

Para alocar as observações em grupos, várias técnicas começam com a determinação das similaridades entre pares de observações. Em muitos casos, as similaridades são baseadas em uma medida de distância. Outros métodos de agrupamento usam a escolha preliminar do centro do grupo ou a comparação da variabilidade dentro e entre grupos. Também é possível agrupar as variáveis e, neste caso, a similaridade pode ser a correlação entre elas.

Um segundo produto de alguns tipos de análise de agrupamentos é um dendrograma, que é uma representação gráfica dos dados na forma de uma árvore genealógica. Este tipo de gráfico tem uma atração óbvia pelos taxonomistas, que estão acostumados a identificar espécies por meio de árvores de tipo similar, mas sua relação ao problema do agrupamento ótimo das espécies não é sempre clara.

1.4 O ESCOPO DA ANÁLISE MULTIVARIADA

De modo geral a análise multivariada inclui grande parte da teoria da estatística. Mesmo nos denominados problemas univariados, tal como o teste de comparação de médias de Student, é necessário a ideia de independência (da média e variância em amostras normais) e a ideia complementar de dependência. Assim, qualquer amostra é um caso especial de uma variável múltipla, em que os indivíduos, como regra, são independentes e identicamente distribuídos.

A estatística univariada é concernente à distribuição de uma única variável, ou variável aleatória, especificada por meio da distribuição de frequência, dada as

probabilidades com a qual ela toma cada um de seus possíveis valores. A estatística multivariada é concernente à distribuição conjunta de variáveis.

Como exemplo, regressão múltipla é um problema de estatística univariada. Somente uma das variáveis envolvidas pode ser a variável dependente. As outras podem ser tanto variáveis controladas com os valores definidos pelo pesquisador, ou podem ter valores aleatórios. A única distribuição envolvida é a da variável dependente em função das variáveis independentes.

Por outro lado, encontrar limites de confiança para coeficientes de correlação envolve a hipótese que as variáveis sigam uma distribuição normal bivariada, ou outra distribuição conjunta especificada. Então, isso é um problema que envolve a estatística multivariada.

A maior parte da estatística multivariada não é recente e as distribuições fundamentais relacionadas com a distribuição normal multivariada foram derivadas nos anos 30, e os métodos desenvolvidos são a base para a maioria dos métodos multivariados que são empregados hoje em dia (Marriott, 1974).

No início do desenvolvimento da estatística multivariada, certos cálculos não eram possíveis de ser efetuados. Com a invenção do computador foi então possível, como por exemplo, a inversão de matrizes de ordem n, cálculo das raízes de polinômios de grau n, e assim sucessivamente. Hoje em dia, com a melhoria dos processadores matemáticos, aumento da capacidade de armazenamento em discos rígidos, placas gráficas de alta definição e monitores, sem falar dos *softwares* disponíveis, a estatística multivariada se tornou realidade.

Existe um senão em tudo isso, que é o entendimento da teoria, o arranjo dos dados, a interpretação dos resultados e o próprio manuseio do *software* e do *hardware*. Sem a integração desses elementos o pesquisador está fadado ao fracasso e a desilusão.

Em geral, o termo "análise multivariada" é usado em um sentido mais restrito. Assim, duas características principais são reconhecidas:

1.- O conjunto de n indivíduos é caracterizado pelo valor p de variáveis diferentes. O caráter multivariado reside da multiplicidade de p variáveis, não no tamanho do conjunto de n observações.

2.- As variáveis são dependentes entre si, sendo que não é possível separar uma ou mais variáveis das demais e considerá-la por si só. As variáveis precisam ser consideradas conjuntamente (Kendall, 1972).

Para um entendimento concreto, alguns problemas típicos práticos em que a estatística multivariada é empregada são a seguir abordados.

1.4.1 Biometria

Em um determinado número de árvores, que podem ser de uma espécie ou de diferentes espécies, muitas medições podem ser feitas em uma única árvore. Assim, quais são as medidas mais apropriadas a serem obtidas? Qual é o número mínimo necessário e como usá-las para testar a homogeneidade ou heterogeneidade na amostra?

1.4.2 Silvicultura

A partir de n áreas diferentes, em que cada uma tem k diferentes espécies, deseja-se saber se é possível fazer comparações da produtividade geral entre áreas e se for o caso, como? Quais são as espécies mais produtivas por área?

1.4.3 Experimentação florestal

Um delineamento com arranjo fatorial dos tratamentos, porém com diferente número de repetições por tratamento, como seria avaliada a significância dos efeitos

principais? Qual a correlação entre as variáveis medidas para cada efeito principal avaliado?

1.4.4 Economia florestal

Para uma determinada empresa, cada ano existe uma produção que de alguma forma é atrelada com a sua atividade comercial geral, tal como a produção de determinado item, taxa de juros, competitividade, emprego de mão de obra qualificada. É possível calcular um número índice da atividade econômica com essa complexidade de informações, e se for possível o seu objetivo tem algum significado?

1.4.5 Avaliação acadêmica

Um certo número n de alunos é examinado em p disciplinas e uma pontuação é dada por disciplina. Qual é o melhor sistema para classificar os alunos em ordem de mérito, tal que ela tenha um significado justificado?

1.4.6 Indústria moveleira

Uma indústria de móveis que produz, dentre os vários modelos, móveis prontos para usar, deseja produzir o suficiente para atender as exigências de sua ampla clientela com o mínimo de defeitos e peças não vendidas. As variáveis envolvidas são o tipo de madeira, o acabamento, a durabilidade, a facilidade de montagem, entre outras. Em qual medida deve-se trabalhar e como proceder para produzir o máximo de satisfação a um custo mínimo?

1.4.7 Fitossanidade florestal

Um determinado princípio ativo é aplicado em brocas marítimas da madeira, as quais são subsequentemente amostradas e examinadas. Amostras afetadas são encontradas quando comparadas com o grupo de controle. Quantos desses efeitos são significantes? Quais são as partes mais afetadas?

1.4.8 Propriedades físicas da madeira

Um conjunto de corpos de prova de madeira são testados para as várias propriedades físicas tais como deformação, dureza, flexão, elasticidade e resistência. É possível detectar nos resultados alguns efeitos sistemáticos que permitam estimá-los a partir das propriedades físicas conhecidas da madeira? É então possível usar os resultados para desenvolver simulação para diferentes usos?

1.4.9 Sociologia florestal

As respostas dadas pelos membros de uma comunidade, em um questionário, que faz parte de um estudo de impacto ambiental são esperadas variar de acordo com o nível de educação. Informações são coletadas para certas características objetivas da amostra, como por exemplo, tempo de moradia na área em questão, importância familiar da área, nível de educação e formação na área de ambiental. Pode um determinado índice ser construído a partir desse material?

2 REVISÃO DE MATRIZES

Neste capítulo estão apresentados elementos básicos de álgebra matricial utilizados ao longo do livro. É essencialmente uma revisão das ferramentas utilizadas em matrizes, não pretendendo ser uma abordagem completa de álgebra matricial. Entretanto, vai auxiliar aqueles que não são familiarizados com álgebra de matrizes.

Com exceção de algumas poucas derivações que são instrutivas, a maior parte dos resultados é dada sem prova. Os exemplos apresentados são desenvolvidos com o uso dos programas MATLAB e JULIA. A programação apresentada visa somente mostrar como se pode utilizar estes programas, sem haver uma preocupação com modos mais eficientes de se obter os mesmos resultados.

2.1 VETORES E MATRIZES

2.1.1 Quantidade escalar

Uma quantidade escalar é aquela que pode ser descrita por um único número.

2.1.2. Rede

Uma rede é um conjunto de quantidades. As quantidades que a integram são elementos da rede. A ordem pode incluir tantas dimensões quantas forem necessárias.

2.1.3 Vetor

Vetor é uma rede linear de quantidades. O ordenamento linear pode existir tanto em linhas como em colunas, sendo a distinção entre eles feitas por se chamarem vetores linhas e vetores colunas. Vetores são usualmente representados entre colchetes, e as vírgulas são usadas para separar elementos, visando melhor compreensão e entendimento dessa estrutura.

$$[a_1, a_2, a_3] \rightarrow \text{vetor linha}$$

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \rightarrow \text{vetor coluna}$$

Observação: atentar para a notação do subscrito para os elementos do vetor.

2.1.4 Matriz

Matriz é uma rede retangular de quantidades.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Observação: atentar dentro dos colchetes para a notação dos subscritos, sendo que o primeiro representa o número da linha e o segundo o número da coluna.

2.1.5 Ordem de uma matriz

Uma matriz de M linhas e C colunas é denominada ser de ordem $\underline{M \cdot C}$ e escrito na forma reduzida como $M \cdot C$.

2.1.6 Matriz quadrada

Uma matriz que possua um número igual de linhas e colunas é denominada de matriz quadrada de ordem m.

2.1.7 Matriz nula ou zero

Uma matriz composta inteiramente de zeros é chamada de matriz nula. A ordem precisa ser especificada. $\begin{bmatrix}0 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$ →matriz nula de ordem 3.

2.1.8 Diagonal de uma matriz

Em uma matriz quadrada, os elementos que possuem os mesmos índices para as linhas e colunas compreendem a diagonal principal.

$$\begin{bmatrix}a_{11} & a_{12} & a_{13}\\a_{21} & a_{22} & a_{23}\\a_{31} & a_{32} & a_{33}\end{bmatrix}$$

2.1.9 Matriz diagonal

Uma matriz que contém zeros fora das posições da diagonal é chamada de matriz diagonal.

$$\begin{bmatrix}2 & 0 & 0\\0 & 3 & 0\\0 & 0 & 4\end{bmatrix}$$

2.1.10 Mariz identidade

Uma matriz diagonal em que todos os elementos da diagonal são iguais a 1 é chamada de matriz identidade (I).

$$\begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$$

2.1.11 Triângulo superior da matriz

Uma matriz quadrada A cujos elementos $a_{ij} = 0$ para i maior que j é chamada de triângulo superior da matriz.

$$\begin{bmatrix}a_{11} & a_{12} & a_{13}\\0 & a_{22} & a_{23}\\0 & 0 & a_{33}\end{bmatrix}$$

2.1.12 Triângulo inferior da matriz

Uma matriz quadrada B cujos elementos $b_{ij} = 0$ para i menor que j é chamada de triângulo inferior da matriz.

$$\begin{bmatrix}b_{11} & 0 & 0\\b_{21} & b_{22} & 0\\b_{31} & b_{32} & b_{33}\end{bmatrix}$$

2.1.13 Matriz simétrica

Uma matriz em que os elementos $a_{ij} = a_{ji}$ é denominada simétrica.

$\begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix}$ sendo $a_{12} = a_{21}$

2.1.14 Matriz escalar

Uma matriz em que todos os elementos possuem um fator comum é denominada matriz escalar.

$$B = \begin{bmatrix}ka_{11} & ka_{12}\\ka_{21} & ka_{22}\end{bmatrix} \text{ ou } B = kA = k\begin{bmatrix}a_{11} & a_{12}\\a_{21} & a_{22}\end{bmatrix}$$

2.1.15 Igualdade de matrizes

Duas matrizes são iguais somente se elas forem idênticas. Isto é, de mesma ordem e com os mesmos elementos.

$$A = \begin{bmatrix} 4 & 2 \\ 6 & 1 \end{bmatrix} \quad B = \begin{bmatrix} 4 & 2 \\ 6 & 1 \end{bmatrix}$$

2.1.16 Transposta de uma matriz

A matriz obtida de uma outra matriz A mudando-se as linhas e colunas de A é chamada de transposta de A e é representada por A' ou A^T.

$$\begin{bmatrix} 2 & 1 \\ 4 & 3 \\ 3 & 5 \end{bmatrix} \text{ ordem } (3 \cdot 2); \; A' = \begin{bmatrix} 2 & 4 & 3 \\ 1 & 3 & 5 \end{bmatrix} \text{ ordem } (2 \cdot 3)$$

2.1.17 Adição de matrizes

A soma de duas matrizes que possuem a mesma ordem é uma matriz obtida em somando-se os elementos correspondentes das duas matrizes. Adição de matrizes é comutativa e associativa e obedece a lei do cancelamento. Se duas matrizes não são de mesma ordem, elas não são similares para adição.

$$C = A + B = \begin{bmatrix} 4 & 3 & 2 \\ 8 & 1 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 2 & -2 \\ -2 & 7 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 5 & 0 \\ 6 & 8 & -1 \end{bmatrix}$$

2.1.18 Subtração de matrizes

Se duas matrizes A e B possuem a mesma ordem, a diferença $C = A - B$ é obtida pela subtração de cada elemento de A do correspondente elemento de B.

2.1.19 Multiplicação de vetores

O produto (interno) de uma linha vetor A e uma coluna vetor B que possuem o mesmo número de elementos é um único valor C obtido pela soma dos produtos dos elementos correspondentes. Então,

$$C^{1\cdot 1} = A^{1\cdot n} \, B^{n\cdot 1}$$

$$\underset{(1 \cdot 3)}{\begin{bmatrix} 2 & 5 & -2 \end{bmatrix}} \underset{(3 \cdot 1)}{\begin{bmatrix} 3 \\ -3 \\ 4 \end{bmatrix}} = [(2)(3) + (5)(-3) + (-2)(4)] = -17$$

Multiplicação de vetores não é comutativa, a não ser que ambos os vetores sejam transpostos.

2.1.20 Multiplicação de matrizes

O produto de uma matriz $A^{m\cdot n}$ e uma matriz $B^{n\cdot c}$ é uma matriz $P^{m\cdot c}$, cujo elemento ij é obtido como o vetor produto da linha i de A e da coluna j de B. Isto é,

$$P^{m\cdot n} = A^{m\cdot n} \, B^{n\cdot c}$$

$$P^{2\cdot 2} = \underset{(2 \cdot 3)}{A \begin{bmatrix} 4 & -1 & 0 \\ 2 & 3 & 8 \end{bmatrix}} \underset{(3 \cdot 2)}{B \begin{bmatrix} 3 & 4 \\ 2 & 1 \\ 0 & 2 \end{bmatrix}}$$

$$= \begin{bmatrix} 4(3) - 1(2) + 0(0) & 4(4) - 1(1) + 0(2) \\ 2(3) + 3(2) + 8(0) & 2(4) + 3(1) + 8(2) \end{bmatrix} = \begin{bmatrix} 10 & 15 \\ 12 & 27 \end{bmatrix}$$

B é dita ser pré-multiplicada por A, e A é dita ser pós-multiplicada por B. A multiplicação de matrizes é associativa, mas não comutativa. Importante notar que o número de colunas

da primeira matriz precisa ser o mesmo do que o número de linhas da segunda matriz para que as matrizes sejam conformes para multiplicação.

2.2 EQUAÇÕES LINEARES SIMULTÂNEAS NA FORMA MATRICIAL

Considerando um conjunto simultâneo de equações lineares,

$$5\,X_1 + 8\,X_2 = 6$$
$$9\,X_1 + 4\,X_2 = 4$$

disponha os coeficientes da esquerda como sendo a matriz A, as incógnitas como um vetor coluna X, e os vetores da direita como um vetor coluna B.

Examine a equação matricial

$$AX = B$$

Note que ela gera as equações originais. Este método de formular equações lineares simultâneas é facilmente estendido para várias equações com várias incógnitas.

2.2.1 Determinantes

Considere a matriz A. Faça $|A|$ ser o determinante da matriz A. O determinante da matriz de ordem $1 \cdot 1$ é simplesmente o número que compreende a matriz. O determinante de uma matriz de ordem $2 \cdot 2$ é assim determinando:

$$|A| = \overset{-}{a_{11}\quad a_{12}}\;\underset{+}{a_{21}\quad a_{22}} = (a_{11}\,a_{22}) - (a_{12}a_{21})$$

$$\begin{bmatrix}3 & 2\\ 6 & 5\end{bmatrix} = (3)(5) - (2)(6) = 15 - 12 = 3$$

em que cada linha tracejada indica o produto dos elementos cortados pela flecha.

O determinante de uma matriz $3 \cdot 3$ é calculado da seguinte maneira:

$$|A| = \begin{matrix} a_{11} & a_{12} & a_{13} & a_{11} & a_{12}\\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22}\\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32}\end{matrix}$$

$$\begin{matrix}- & - & - & + & + & +\end{matrix}$$

$$\begin{bmatrix}6 & 2 & 1\\ 3 & 0 & 1\\ 8 & 2 & 3\end{bmatrix} = \begin{bmatrix}6 & 2 & 1 & 6 & 2\\ 3 & 0 & 1 & 3 & 0\\ 8 & 2 & 3 & 8 & 2\end{bmatrix}$$

$$= (6)(0)(3) + (2)(1)(8) + (1)(3)(2) - (1)(0)(8) - (6)(1)(2) - (-2)(3)(3) = -8$$

Determinantes de ordens maiores podem ser avaliados pela expansão usando os menores e os cofatores.

2.2.2 Menores

O menor de primeira ordem é o determinante da matriz formada em eliminando-se uma linha e uma coluna da matriz original. Se a linha i e a coluna j são eliminadas, o menor que resulta é simbolizado como $|M_{ij}|$.

O menor de um determinante é um outro determinante formado quando se elimina um número igual de linhas e colunas do determinante original. A ordem do menor indica o número de linhas (ou colunas) no menor.

Para um determinante de ordem 3, tem-se nove menores de segunda ordem obtidos eliminando-se uma linha em uma coluna.

$$\det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Os três menores obtidos em eliminando-se a primeira linha são:

$$|M_{11}| = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}; |M_{12}| = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}; |M_{13}| = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Em que: $|M_{ij}|$ refere-se ao menor de $|A|$ obtido em eliminando-se a linha i e a coluna j.

Como exemplo, os três menores obtidos da primeira linha do determinante.

$$\det A = \begin{vmatrix} 2 & 4 & 1 \\ 3 & 5 & 2 \\ 2 & 1 & 0 \end{vmatrix}$$

O primeiro menor $|M_{11}|$ que é o próprio determinante, é obtido em removendo-se a primeira linha e a primeira coluna de $|A|$, sendo:

$$|M_{11}| = \begin{vmatrix} 5 & 2 \\ 1 & 0 \end{vmatrix} = 5(0) - 1(2) = -2$$

Igualmente,

$$|M_{12}| = \begin{vmatrix} 3 & 2 \\ 2 & 0 \end{vmatrix} = 3(0) - 2(2) = 4$$

que é obtido em removendo-se a primeira linha e a segunda coluna de $|A|$. E, para $|M_{13}|$ tem-se:

$$|M_{13}| = \begin{vmatrix} 3 & 5 \\ 2 & 1 \end{vmatrix} = 3(1) - 5(2) = -7$$

E assim sucessivamente até que todos os menores tenham sido calculados, no exemplo, nove menores.

2.2.3 Cofatores

Um cofator é formado para estabelecer um sinal para o menor de primeira ordem, dependendo da linha e coluna eliminada na formação do menor. O sinal é determinado por

$$(-1)^{i+j}$$

onde i é o número da linha eliminada e j é o número da coluna eliminada. O cofator é representado por α_{ij}. Então,

$$\alpha_{ij} = (-1)^{i+j}|M_{ij}|$$

Então, considerando-se a matriz

$$A = \begin{vmatrix} 2 & 4 & 1 \\ 3 & 5 & 2 \\ 2 & 1 & 0 \end{vmatrix}$$

O cofator $\alpha_{11} = (-1)^{1+1}\begin{vmatrix} 5 & 2 \\ 1 & 0 \end{vmatrix} = (-1)^2|-2|$

$$\alpha_{11} = 2$$

2.2.4 Expansão por meio dos cofatores

O determinante de uma matriz pode ser obtido pela expansão ao longo da linha i, a seguir:

$$|A| = a_{i1}\alpha_{i1} + a_{2j}\alpha_{2j} + \cdots + a_{nj}\alpha_{nj}$$

$$\det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\alpha_{11} + a_{12}\alpha_{12} + a_{13}\alpha_{13}$$
$$= a_{11}|M_{11}| + a_{12}(-1)|M_{12}| + a_{13}|M_{13}|$$

2.2.5 Singularidade

Uma matriz que possui determinante igual a zero é dita ser singular. A consequência é que a matriz não possui sua inversa.

2.2.6 Duas propriedades dos determinantes

$$|A| = |A'|$$
$$|AB| = |A||B|$$

2.2.7 Grau de uma matriz

O grau de uma matriz quadrada e não singular é a ordem da matriz.

Menores de ordens maiores são formados em eliminando-se várias linhas e colunas para formar subconjuntos quadrados. O grau de uma matriz singular ou uma matriz que não seja quadrada é a ordem do maior menor que não seja zero.

2.2.8 Matriz cofator

A matriz cofator, A^C, é obtida em substituindo-se cada elemento da matriz quadrada A por seu correspondente cofator α_{ij}.

2.2.9 Matriz adjunta

A adjunta de uma matriz, $adj\ A$, é a transposta da matriz cofator. Duas propriedades da matriz adjunta são:

$$A\ (adj\ A) = (adj\ A)A = |A|I$$

e

$$A\ |adj\ A| = |adj\ A||A| = \ |A|^n$$

onde n é a ordem da matriz A não singular.

2.3 MATRIZ INVERSA

Dada duas matrizes A e B, se $A\ B = B\ A = I$, então A^{-1} é dita ser a inversa de A e vice e versa. A notação para a inversa é A^{-1}. Então,

$$A\ A^{-1} = A^{-1}A = I$$

Sendo $(adj\ A)A = |A|I$ é possível fazer uma multiplicação escalar por $\frac{1}{|A|}$ em ambos os lados para obter-se

$$\frac{1}{|A|}(adj\ A)A = \frac{|A|}{|A|}I = I = A^{-1}A$$

Então,

$$A^{-1} = \frac{1}{|A|}(adj\ A)$$

Igualmente,

$$|A|A^{-1} = adj\ A$$

2.3.1 Inversa por meio de transformações elementares

As seguintes operações, denominadas de transformações elementares, não modificam tanto a ordem como o grau de uma matriz.

(1) mudar duas linhas ou duas colunas;

(2) multiplicar toda uma linha ou toda uma coluna por um escalar diferente de zero.
(3) adicionar um escalar múltiplo de uma linha (coluna) para outra linha (coluna).

Para obter a inversa por de transformações elementares, procede-se:

$$[A][I]$$

Então reduza A para uma identidade por meio das transformações elementares, e proceda as mesmas operações em I. Quando A foi reduzida para uma matriz identidade, a matriz identidade anterior irá conter A^{-1}.

2.3.2 Inversa de um produto

A inversa de um produto é o produto das inversas em ordem reversa. Isto é,

$$(AB)^{-1} = B^{-1}A^{-1}$$

2.4 OUTRAS MATRIZES E OPERAÇÕES

2.4.1 Matriz ortogonal

Uma matriz ortogonal é dita ortogonal se $A^{-1} = A$.

2.4.2 Traço da matriz

O traço de uma matriz quadrada é a soma dos elementos da diagonal principal.

2.4.3 Matrizes de transformação

Um vetor X em um sistema de coordenadas pode ser transformada ou expressa em termos de um outro sistema de coordenadas por

$$Y = T\,X$$

onde T é a matriz de transformação e X é um vetor coluna. Se T não é singular, a transformação reversa é:

$$T^{-1}Y = T^{-1}TX = IX = X$$

2.4.4 Escalonamento ou normalização de vetores

Um vetor X pode ser escalonado sendo que $X'X = C$ pela multiplicação de cada componente por $\sqrt{\frac{C}{X'X}}$, e se $X'X = 1$ o vetor é dito ser normalizado.

2.4.5 Matriz modal

A matriz espectral S da matriz A é a matriz diagonal constituída dos valores característicos de A.

2.4.6 Matriz modal normalizada

Se A é uma matriz simétrica com elementos reais, então a matriz modal normalizada associada com A é uma matriz ortogonal, isto é,

$$N' = N^{-1}$$

2.4.7 Diagonalização de uma matriz

Se a matriz modal não é singular $M^{-1}AM = S$ onde A é uma matriz quadrada, M é a sua matriz modal, e S a sua matriz espectral.

2.4.8 Matrizes similares

Duas matrizes S e A são ditas similares se existir uma matriz M não singular tal que:

$$S = M^{-1}\,A\,M$$

2.5 VALORES CARACTERÍSTICOS E VETORES CARACTERÍSTICOS (*EIGENVALUES* E *EIGENVECTORS*)

Dada uma matriz A, o chamado problema valor característico é encontrar escalares λ e seus associados vetores X tal que,

$$A\,X = \lambda\,X$$

Para resolver $A\,X = \lambda\,X$, deve-se subtrair $\lambda\,X$ de ambos os lados para obter

$$A\,X - \lambda\,X = 0$$

Em seguida fatorar X para se obter

$$(A - \lambda I)X = 0$$

Esta é a matriz de formulação do conjunto de equações lineares homogêneas com os $X's$ como incógnitas e a matriz de coeficientes tendo elementos envolvendo $\lambda's$ na diagonal.

O conjunto de equações homogêneas possui soluções não triviais somente se o determinante da matriz de coeficientes for zero. Isto é,

$$|A - \lambda I| = 0$$

A expansão de $|A - \lambda I|$ resulta em um polinômio em λ, e quando igualado a zero resulta na equação característica da matriz A. As raízes desta equação (valores de λ) são chamados de vetores característicos, valores *eigen*, raízes características ou raízes latentes de A.

Para cada valor característico existe um vetor – X determinado pelo conjunto de equações

$$(A - \lambda I)X = 0$$

Estes vetores – X são chamados de vetores característicos, vetores *eign* ou vetores latentes da matriz A.

Uma vez que os vetores característicos são soluções não triviais para o conjunto de equações lineares, eles possuem pelo menos um componente arbitrário, e qualquer escalar múltiplo de um vetor característico irá também ser um vetor característico.

Então, é comum escalonar cada vetor característico sendo que

$$X'X = \lambda$$

onde λ é o valor característico correspondente ao vetor característico.

2.5.1 Propriedades dos valores e dos vetores característicos

1- Se uma matriz é multiplicada por um escalar, os valores característicos serão multiplicados por este escalar, enquanto os vetores característicos não mudam.
2- Os valores característicos da inversa de uma matriz são os recíprocos dos valores característicos da matriz original, mas os vetores característicos permanecem os mesmos.
3- Qualquer valor característico de uma matriz ortogonal é tanto $+1$ ou -1.
4- Os valores característicos de uma matriz diagonal ou triangular são os elementos da diagonal da matriz.
5- O somatório dos valores característicos é o traço da matriz.
6- Os valores característicos de uma matriz simétrica real são todos reais.

2.5.2 Exemplo

Para determinar o polinômio característico para a matriz

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

tem-se:

$$|A - \lambda I| = \begin{vmatrix} (1-\lambda) & 2 \\ 2 & (1-\lambda) \end{vmatrix} = (1-\lambda)^2 - 4 = \lambda^2 - 2\lambda - 3$$

Para encontrar as raízes do polinômio característico, iguala-se a zero obtendo-se:

$$|A - \lambda I| = 0$$

a qual é chamada equação característica. É importante observar que a matriz A de ordem n gera a equação característica de ordem n que possui n raízes. A equação característica obtida para:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

é

$$|A - \lambda I| = 0$$

ou

$$\lambda^2 - 2\lambda - 3$$

As raízes, ou fatores, da equação característica são chamados de raízes características, autovalores ou valores *eign*. Neste caso as raízes características são:

$$\lambda_1 = -1, \lambda_2 = 3$$

pois

$$(\lambda + 1)(\lambda - 3) = \lambda^2 - 2\lambda - 3 = 0$$

Para cada raiz característica, λ_i, da equação característica, existe um vetor característico x_i. Este vetor característico é a solução das equações homogêneas que são assim obtidas:

$$(A - \lambda_i I)\boldsymbol{x_i} = 0$$

O vetor em negrito $\boldsymbol{x_i}$ é o vetor correspondente a raiz i^{th}.

$$\boldsymbol{x_i} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Considerando a matriz

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

com as raízes características $\lambda_1 = -1$ e $\lambda_2 = 3$, determinadas acima, tem-se associado com a primeira raiz:

$$(A - \lambda_1 I)\boldsymbol{x_1} = 0$$

ou

$$\left(\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} - (-1)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

o que implica em

$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$\boldsymbol{x_1}$ é o vetor característico associado com a raiz $\lambda_1 = -1$.

Os elementos de $\boldsymbol{x_1}$ são x_1 e x_2. Assim, $x_1 = -x_2$ sendo que $2\,x_1 + 2\,x_2 = 0$ que é obtido de qualquer linha da equação $(A - \lambda_1 I)\boldsymbol{x_1} = 0$.

Fazendo $\boldsymbol{x_1} = \beta$, uma constante arbitrária. Então,

$$\boldsymbol{x_1} = \beta\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

O subscrito 1 em x_1 implica que se tem um vetor característico associado com a primeira raiz $\lambda_1 = -1$. Importante notar que a matriz $(A - \lambda_1 I)$ possui ordem 1 e tem um valor arbitrário e um valor único em $\boldsymbol{x}$.

A segunda raiz característica associado a $\lambda_2 = 3$ é, então, $(A - \lambda_2 I)\boldsymbol{x_2} = 0$

e,

$$\left(\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} - (3)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

ou,

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Lembrando aqui que

$$\boldsymbol{x_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Novamente para λ_2

$$(A - \lambda_2 I) = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}$$

é de ordem 1 e tem um valor arbitrário em $\boldsymbol{x_2}$ e um valor único determinado.

Os valores e os vetores característicos devem ser bem entendidos, pois serão amplamente utilizados em todas as análises da estatística multivariada.

Maiores detalhes serão abordados principalmente no capítulo de Componentes Principais.

Para fins de facilitação do entendimento das operações matriciais ao longo deste livro será utilizada a seguinte notação para autovalores e autovetores:

Autovalores serão representados pela notação λ ou l nas linguagens de programação e os autovetores serão representados pela notação e ou c.

2.6 EXEMPLO DOS COMANDOS DE ÁLGEBRA MATRICIAL NO MATLAB

QUANTIDADE ESCALAR

```
A=2
A =
     2
```

VETOR LINHA

```
A=[2 3 4]
A =
     2     3     4
```

VETOR COLUNA

```
B=[2 ; 3; 4]
B =
     2
     3
     4
```

MATRIZ

```
A=[3 5 7 ; 8 9 10]
A =
     3     5     7
     8     9    10
```

ORDEM DE UMA MATRIZ

```
A=[5 6 7 ; 9 8 7 ;5 4 3]
A =
     5     6     7
     9     8     7
     5     4     3
[linhas , colunas]=size(A)
linhas =
     3
colunas =
     3
```

MATRIZ QUADRADA

```
A=[2 4 ; 5 6]
A =
```

```
     2     4
     5     6
```

MATRIZ NULIDADE OU ZERO

```
A=zeros(3,3)
A =
     0     0     0
     0     0     0
     0     0     0
```

DIAGONAL DE UMA MATRIZ

```
A=[5 6 7 ; 9 8 7 ; 5 4 3]
A =
     5     6     7
     9     8     7
     5     4     3
DIAGONAL=diag(A)
DIAGONAL =
     5
     8
     3
```

MATRIZ DIAGONAL

```
V=[3 ; 5 ; 1]
V =
     3
     5
     1
A=diag(V)
A =
     3     0     0
     0     5     0
     0     0     1
```

MATRIZ IDENTIDADE

```
A=eye(4)
A =
     1     0     0     0
     0     1     0     0
     0     0     1     0
     0     0     0     1
```

TRIÂNGULO SUPERIOR DA MATRIZ

```
A=[5 6 7 1 ; 9 8 7 3 ; 5 4 3 2 ; 7 9 5 1]
A =
     5     6     7     1
     9     8     7     3
     5     4     3     2
     7     9     5     1
B=tril(A)
B =
     5     6     7     1
     0     8     7     3
     0     0     3     2
     0     0     0     1
```

TRIÂNGULO INFERIOR DA MATRIZ

```
C=tril(A)
```

```
C =
     5     0     0     0
     9     8     0     0
     5     4     3     0
     7     9     5     1
```

MATRIZ SIMÉTRICA

```
A=[ 4 3 ; 3 5]
A =
     4     3
     3     5
```

MATRIZ ESCALAR

```
C=10*[1 2 ; 3 4]
C =
    10    20
    30    40
```

IGUALDADE DE MATRIZES

```
A=[4 2 ; 6 1]
A =
     4     2
     6     1
B=[4 2 ; 3 1]
B =
     4     2
     3     1
B=A
B =
     4     2
     6     1
```

TRANSPOSTA DE UMA MATRIZ

```
A=[4 5 6 ; 7 8 9]
A =
     4     5     6
     7     8     9
B=A'
B =
     4     7
     5     8
     6     9
```

ADIÇÃO DE MATRIZES

```
A=[2 2 ; 3 3]
A =
     2     2
     3     3
B=[ 5 6 ; 6 5]
B =
     5     6
     6     5
C=A+B
C =
     7     8
     9     8
```

SUBTRAÇÃO DE MATRIZES

```
D=A-B
D =
    -3    -4
    -3    -2
```

MULTIPLICAÇÃO DE VETORES

```
A=[2 5 -2 ]
A =
     2     5    -2
B=[3 ; -1 ; 4]
B =
     3
    -1
     4
C=A*B
C =
    -7
```

MULTIPLICAÇÃO DE MATRIZES

```
A=[2 2 ; 3 3]
A =
     2     2
     3     3
B=[5 6 ; 6 5]
B =
     5     6
     6     5
C = A * B
C =
     22     22
     33     33
```

EQUAÇÕES LINEARES SIMULTÂNEAS NA FORMA MATRICIAL

```
A =[5 8; 9 4]
A =
     5     8
     9     4
B = [6;4]
B =
     6
     4
X=inv(A)*B
X =
    0.1538
    0.6538
B=A*X
B =
    6.0000
    4.0000
```

DETERMINANTES

```
A=[ 3 2 ; 6 5]
A =
     3     2
     6     5
B=det(A)
```

```
B =
     3
```

SINGULARIDADE

```
A=[2 2 ; 2 2]
A =
     2     2
     2     2
B=det(A)
B =
     0
C=inv(A)
[Warning: Matrix is singular to working precision.]
C =
   Inf   Inf
   Inf   Inf
```

MATRIZ INVERSA

```
A=[1 3 4 ; 5 6 7 ; 9 8 7]
A =
     1     3     4
     5     6     7
     9     8     7
B=inv(A)
B =
   -1.0000    0.7857   -0.2143
    2.0000   -2.0714    0.9286
   -1.0000    1.3571   -0.6429
```

TRAÇO DA MATRIZ

```
A=[1 3 4 ; 5 6 7 ; 9 8 7 ]
A =
     1     3     4
     5     6     7
     9     8     7
B=trace(A)
B =
    14
```

VALORES CARACTERÍSTICOS E VETORES CARACTERÍSITICOS (*EIGENVALUES* E *EIGENVECTORS*)

```
A=[ 1 2 ; 2 1]
A =
     1     2
     2     1

[e , l] = eig(A)
e =
   -0.7071    0.7071
    0.7071    0.7071
l =
    -1     0
     0     3
```

2.7 EXEMPLO DOS COMANDOS DE ÁLGEBRA MATRICIAL NO JULIA

```
#Pacotes
using LinearAlgebra

#Quantidade escalar
A=2

#Vetor linha
A=[2 3 4]

#Vetor coluna
B=[2;3;4]
B=A'

#Matriz
A=[3 5 7; 8 9 10]

#Ordem de uma matriz
A=[5 6 7; 9 8 7; 5 4 3]
size(A)
A_linhas, A_colunas = size (A)
A_linhas
A_colunas

#Matriz quadrada
A=[2 4; 5 6]

#Matriz nulidade ou zero
A = zeros (3,3)

#Diagonal de uma matriz
A=[5 6 7; 9 8 7; 5 4 3]
Diagonal=diag(A)
Diagonal=diagonal(A)

#Matriz diagonal
V = [3, 5, 1]
A=diagm(V)
A=Diagonal(V)

#Mariz identidade
A=Matrix(I, 4, 4)

#Triângulo superior da matriz
A=[5 6 7 1; 9 8 7 3; 5 4 3 2; 7 9 5 1]
B=UpperTriangular(A)

#Triângulo inferior da matriz
C=LowerTriangular(A)
```

```
#Matriz simétrica
A=[4 3; 1 5]
B=Symmetric(A)

#Matriz escalar
C=10*[1 2; 3 4]

#Igualdade de matrizes
A=[4 2; 6 1]
B=[4 2; 6 1]
B=A

#Transposta de uma matriz
A=[4 5 6; 7 8 9]
B=A'

#Adição de matrizes
A=[2 2; 3 3]
B=[5 6; 6 5]
C=A+B

#Subtração de matrizes
D=A-B

#Multiplicação de vetores
A=[2 5 -2]
B=[3;-3;4]
C=A*B

#Multiplicação de matrizes
A=[2 2; 3 3]
B=[5 6; 6 5]
C=A*B

#Equações lineares simultâneas na forma matricial
A=[5 8; 9 4]
B=[6;4]
X=inv(A)*B
B=A*X

#Determinantes
A=[3 2; 6 5]
B=det(A)

#Singularidade
A=[2 2; 2 2]
B=det(A)
C=inv(A)
```

```
#Matriz adjunta
A=[4 5; 6 7]
B=Adjoint(A)

#Matriz inversa
A=[1 3 4; 5 6 7; 9 8 7]
B=inv(A)

#Traço da matriz
A=[1 3 4; 5 6 7; 9 8 7]
B=tr(A)

#Valores característicos e vetores característicos (eigenvalues e eigenvectors)
A=[1 2; 2 1]
B=eigen(A)
```

3 COMPARAÇÃO DE MÉDIAS MULTIVARIADAS

Testes de hipóteses no contexto multivariado são mais complexos quando comparados com os testes univariados. No caso univariado, para duas amostras, são utilizados o teste de t para amostras independentes e pareadas. Quando o número de amostras é maior do que duas a razão F de Fischer é utilizada. No caso multivariado o número de parâmetros pode ser muito grande, uma vez que envolve p médias, p variâncias e a combinação de todas as variáveis duas a duas representando a covariância. O número total de parâmetros é dado por:

$$0{,}5 * p * (p + 3) \tag{3.1}$$

Para um $p = 5$, por exemplo, o número de parâmetros é 20, e para cada um deles poderia ser formulada uma hipótese.

A motivação para testar p variáveis de modo multivariado, ao invés de univariado, pode ser explicada por quatro argumentos:

- O uso de p testes univariados inflam o erro tipo I, α, enquanto o teste multivariado preserva o nível α.
- Os testes univariados ignoram por completo a correlação entre as variáveis, enquanto os testes multivariados fazem o uso direto das correlações.
- O teste multivariado é mais poderoso em muitos casos.
- Muitos testes multivariados envolvendo médias tem como subproduto a construção de combinação lineares das variáveis que revelam mais sobre como as variáveis se unem para rejeitar a hipótese.

Vários testes univariados e intervalos de confiança são baseados na distribuição normal. Similarmente a maioria dos procedimentos multivariados tem a distribuição normal multivariada como sua base. Como a distribuição normal multivariada é uma extensão da distribuição normal univariada e compartilha muito de suas características, ambas as funções devem ser descritas.

3.1 DENSIDADE NORMAL MULTIVARIADA

A densidade normal univariada é representada pela fórmula:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]} \tag{3.2}$$

sendo que

$$E[x] = \int_{-\infty}^{+\infty} xp\,(x)dx = \mu \tag{3.3}$$

e

$$E[(x-\mu)^2] = \int_{-\infty}^{+\infty}(x-\mu)^2 p(x)dx = \sigma^2 \tag{3.4}$$

A densidade normal univariada é caracterizada por dois parâmetros, a média μ e a variância σ^2, e é representada por $p(x) \sim N(\mu, \sigma^2)$, isto é, x é normalmente distribuído com média μ e variância σ^2. Amostras distribuídas normalmente possuem a tendência em agruparem-se ao redor da média, com a dispersão proporcional ao desvio padrão σ. Para amostras obtidas de uma população normal, 95% dos valores estarão dentro do intervalo $|x - \mu| \leq 2\sigma$.

A Figura 3.1 representa a distribuição normal univariada padrão, com média $\mu = 0$ e variância $\sigma^2 = 1$.

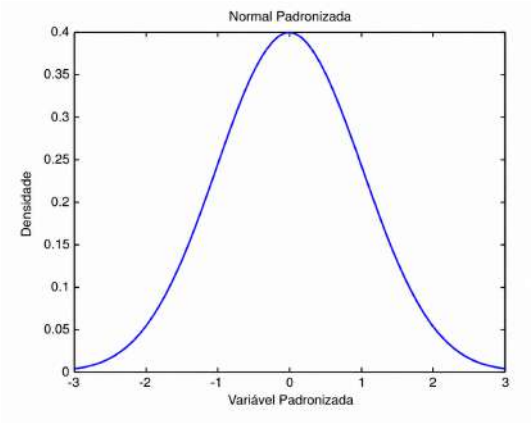


Figura 3.1: Distribuição normal univariada padronizada.

Se x tem uma distribuição normal multivariada com vetor de médias μ e matriz variância-covariância $\sum$ a A densidade normal multivariada é representada pela seguinte fórmula:

$$p(x) = \frac{1}{(2\pi)^{p/2}\,|\sum|^{1/2}}\; e^{\left[-\frac{1}{2}(x-\mu)'\sum^{-1}(x-\mu)\right]} \tag{3.5}$$

Em que: x é um vetor coluna, p é o número de variáveis, μ é o vetor de médias e $\sum$ é a matriz variância-covariância.

A distribuição normal multivariada é representada por:

$$p(x) \sim N(\mu, \sum)$$

ou seja,

$$\mu = E[X] \tag{3.6}$$

$$\sum = E[(x-\mu)(x-\mu)'] \tag{3.7}$$

$$\mu_i = \mathrm{E}[x_i] \tag{3.8}$$

$$\sigma_{ij} = E\left[(x_i-\mu_i)\left(x_j-\mu_j\right)'\right] \tag{3.9}$$

A matriz variância-covariância $\sum$ é sempre simétrica e positiva semi-definida, que quando assumida dessa forma resulta no determinante de $\sum$ estritamente positivo. A diagonal da matriz representada por σ_{ij} é a variância de x_i e fora da diagonal são as covariâncias de x_i e x_j representadas por σ_{ij}. Caso x_i e x_j sejam estatisticamente independentes, então $\sigma_{ij} = 0$. Se todos os valores fora da diagonal da matriz forem zero, a $p(x)$ é reduzida ao produto da densidade normal univariada para os componentes de x.

Amostras sorteadas de uma população normal tendem a formar uma única nuvem de pontos ou conglomerado. Duas representações da densidade normal estão representadas na Figura 3.2.

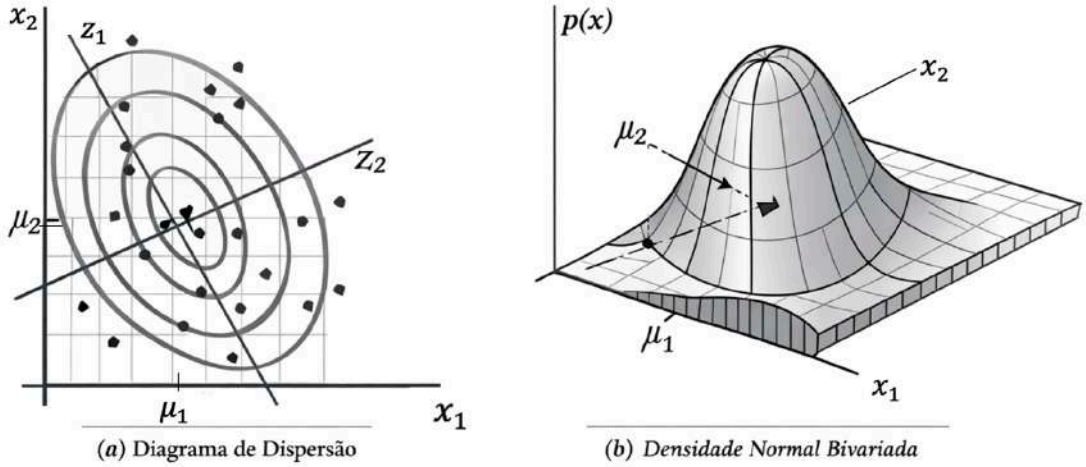


Figura 3.2: Duas representações de densidade normal (adaptado de Morisson, 1976 e Finn, 1974). Diagrama de dispersão e Densidade normal bivariada, respectivamente.

O centro do conglomerado é determinado pelo vetor médio e a forma de agrupamento (conglomerado) é determinada pela matriz variância-covariância. Pela equação 3.5 segue que o conjunto de pontos de uma densidade constante são hiperelipsóides para a qual a forma quadrática $(x-\mu)'\sum^{-1}(x-\mu)$ é constante. Os eixos principais, z_1 e z_2, desses hiperelipsóides são representados pelos vetores característicos de v, e os valores característicos (raízes características) determinam o comprimento desses eixos. A quantidade

$$m^2 = (x-\mu)'\sum^{-1}(x-\mu) \qquad (3.10)$$

é denominada o quadrado da distância de Mahalanobis de X para μ. As curvas de densidade constante são hiperelipsóides da distância constante Mahalanobis para a μ. O volume dos hiperelipsóides medem a dispersão das amostras em relação a média.

3.1.1 Verificação da normalidade multivariada

A verificação da normalidade multivariada não é conceitualmente tão direta como para a normalidade univariada. No caso univariado os valores de uma amostra são divididos em intervalos (classes) e são contados quantos valores estão em cada intervalo. Essas frequências observadas (contagens) são comparadas com as frequências esperadas sob a condição que a amostra tem origem numa distribuição normal com a mesma média e variância da amostra. Se as observações são multivariadas tem-se uma região p −dimensional que deveria ser dividida em mais sub-regiões do que no caso univariado. As frequências esperadas para essas sub-regiões seriam menos facilmente obtidas. Muitas sub-regiões provavelmente teriam poucas observações, pois os dados multivariados são mais espaçados do que univariados, e o teste torna-se impraticável.

Mesmo diante disso, alguma verificação da normalidade multivariada é desejável, e existem vários procedimentos propostos. Um procedimento comum na área florestal é baseado na distância padronizada entre cada valor em relação a média.

A normalidade p variada é verificada para dimensões $p \geq 2$. Para a normal bivariada ($p = 2$), tem-se:

$$\mathrm{P}\left[\underline{X}: \left(\underline{\mathrm{X}} - \underline{\mu}\right)' \sum^{-1} \left(\underline{\mathrm{X}} - \underline{\mu}\right) \leq \chi_2^2(0{,}5)\right] = 1 - \alpha = 1 - 0{,}5 = 0{,}5 \qquad (3.11)$$

Em que: $\left(\underline{\mathrm{X}} - \underline{\mu}\right)' \sum^{-1} \left(\underline{\mathrm{X}} - \underline{\mu}\right)$ é a distância d_j^2.

A distribuição das distâncias d_j^2 pode ser adequadamente aproximada de uma distribuição de Qui-quadrado (χ_p^2). Nesse caso, deve-se esperar a mesma porcentagem (50%) das observações amostrais, tais que $\left(\underline{X} - \overline{\underline{X}}\right)' S^{-1}\left(\underline{X} - \overline{\underline{X}}\right) \leq \chi_2^2(0{,}5)$, caso contrário, existe evidência que a hipótese da normalidade bivariada deve ser rejeitada.

Exemplo 3.1

Considere 10 valores de diâmetro à altura do peito (cm) e altura (m) de árvores de acácia negra (Tabela 3.1).

Tabela 3.1: Pares de diâmetros e alturas obtidos em uma amostra de acácia-negra com 10 anos.

N	dap (cm)	h (m)	d_j^2	$\chi_2^2(0{,}5)$
1	19,58	19,70	1,1963	1,3863
2	7,48	14,90	1,2525	1,3863
3	16,07	18,20	0,2942	1,3863
4	9,23	15,70	0,7804	1,3863
5	9,39	11,80	2,8040	1,3863
6	5,41	12,00	1,5590	1,3863
7	8,44	12,00	1,9392	1,3863
8	10,19	18,50	4,0631	1,3863
9	18,94	18,80	1,0950	1,3863
10	23,55	21,70	3,0164	1,3863

As distâncias $d_j^2 \leq \chi_2^2$ perfazem 50% das observações (5 valores), indicando a presença de normalidade multivariada.

Um segundo procedimento, quando $p > 2$, consiste em julgar a normalidade conjunta de dados multivariados por meio do quadrado da distância generalizada:

$$d_j^2 = \left(\underline{X}_j - \overline{\underline{X}}\right)' S^{-1}\left(\underline{X}_j - \overline{\underline{X}}\right),\ j = 1{,}2 \ldots, n \qquad (3.12)$$

Em que: $\underline{X}_1$, $\underline{X}_2$, …, $\underline{X}_n$ são observações amostrais.

Quando as amostras têm origem em uma população normal multivariada e ambos n e $n - p$ são maiores que 25, cada uma das distâncias $d_1^2, d_2^2, \ldots, d_n^2$ comporta-se como uma variável aleatória com distribuição de Qui-quadrado. O método consiste em 2 passos:
i) Ordenar as distâncias em ordem crescente.

ii) Representar graficamente os pares $\left[d_j^2,\ \chi_p^2\left(\frac{j-\frac{1}{2}}{n}\right)\right]$, obtendo-se χ_p^2 por meio da distribuição de Qui-quadrado para p graus de liberdade. O gráfico obtido deve se aproximar de uma reta.

Exemplo 3.2

Para o exemplo anterior, os resultados estão apresentados na Tabela 3.2 e Figura 3.3.

As distâncias d_j^2 se aproximam de uma reta, indicando a presença de normalidade multivariada.

Para o caso bivariado, a comparação da distribuição com uma variância pequena e uma distribuição com variância grande é apresentado na Figura 3.4. A Figura 3.4A apresenta alta correlação entre as variáveis. Em geral, para qualquer número de variáveis

p uma diminuição nas intercorrelações entre as variáveis ou um aumento nas variâncias ocasionará uma maior variância generalizada $|\sum|$ (Figura 3.4B).

Tabela 3.2: Pares de diâmetros e alturas obtidos em uma amostra de acácia-negra com 10 anos.

n	dap (cm)	h (m)	d_j^2	$\chi_p^2\left(\frac{j-\frac{1}{2}}{n}\right)$
1	19,58	19,70	0,2942	0,1026
2	7,48	14,90	0,7804	0,3250
3	16,07	18,20	1,0950	0,5754
4	9,23	15,70	1,1963	0,8616
5	9,39	11,80	1,2525	1,1957
6	5,41	12,00	1,5590	1,5070
7	8,44	12,00	1,9392	2,0996
8	10,19	18,50	2,8040	2,7726
9	18,94	18,80	3,0164	3,7942
10	23,55	21,70	4,0631	5,9915

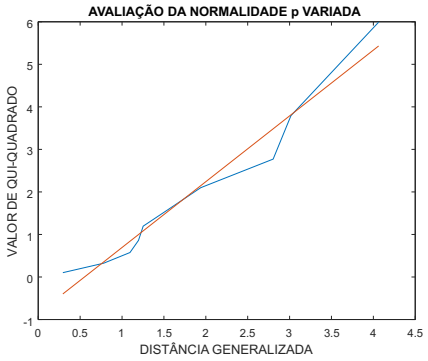


Figura 3.3: Gráfico para avaliação da normalidade p variada.

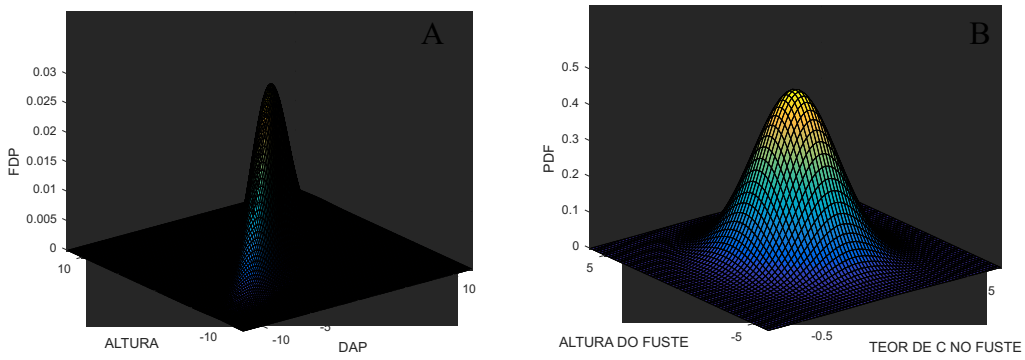


Figura 3.4: Densidades bivariadas normais. Figura A: Gráfico da distribuição normal bivariada para as variáveis diâmetro à altura do peito e altura com correlação alta (r = 0,88). Figura B: Gráfico da distribuição normal bivariada para as variáveis teor de carbono do fuste e de sua altura com correlação muito baixa (r = - 0,056).

3.2 VERIFICAÇÃO DA HOMOGENEIDADE DA MATRIZ VARIÂNCIA-COVARIÂNCIA

Uma condicionante para o teste T^2 de Hotelling (3.3) ou para MANOVA (3.4), comparando dois ou mais vetores de média, é que as correspondentes matrizes variância-covariância da população não difiram entre si, ou seja $\sum_1 = \sum_2 = \cdots = \sum_k$ (matrizes de variância-covariância homogêneas). Considerando essa hipótese as matrizes variância-covariância das amostras $S_1 = S_2 = \cdots = S_k$ refletem a variância comum da população $\sum$ e, desta forma, são combinadas para obter uma estimativa de $\sum$. Se $\sum_1 = \sum_2 = \cdots = \sum_k$ não for verdadeira, grandes diferenças entre $S_1 = S_2 = \cdots = S_k$ podem levar a rejeição de H_0. Entretanto, os testes de T^2 e da MANOVA são robustos, desde que os tamanhos das amostras sejam grandes e iguais.

Para k populações multivariadas, a hipótese de igualdade das matrizes de variânica-covariância é:

$$H_0: \sum_1 = \sum_2 = \cdots = \sum_k$$

Esta hipótese é testada por meio da estatística de razão de verossimilhança generalizada:

$$M = \sum v_i \ln|S| - \sum_{i=1}^{k} v_i \ln|S_i| \tag{3.13}$$

Em que: v_i: graus de liberdade, $v_i = n_i - 1$ para uma amostra aleatória de tamanho n_i proveniente da i-ésima população.

$$S = \frac{1}{\sum v_i} \sum_{i=1}^{k} v_i S_i \tag{3.13}$$

O fator de escala é dado por:

$$C^{-1} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum_{i=1}^{k} \frac{1}{v_i} - \frac{1}{\sum v_i} \right) \tag{3.14}$$

Se esse fator de escala for introduzido, a quantidade MC^{-1} tem distribuição que se aproxima da distribuição de Qui-quadrado com graus de liberdade $\frac{1}{2}(k-1)p(p+1)$ para grandes valores de v_i. Se todos os v_i forem iguais a v, então:

$$C^{-1} = 1 - \frac{(2p^2 + 3p - 1)(k+1)}{6(p+1)\, k\, n} \tag{3.15}$$

A aproximação para a distribuição de Qui-quadrado fornece bons resultados se k (número de grupos) e p forem no máximo 5 e cada v_i for pelo menos 20.

Para grandes valores de v_i pode-se tomar a aproximação para a distribuição F.

$$a_1 = 1 - c \tag{3.16}$$

$$a_2 = \frac{(p-1)(p+2)}{6(k-1)} \left(\sum_{j=1}^{k} \frac{1}{v_j^2} - \frac{1}{(n-k)^2} \right) \tag{3.17}$$

$$v = \frac{p(p+1)(k-1)}{2} \tag{3.18}$$

$$v_0 = \frac{v+2}{a_2 - a_1^2} \tag{3.19}$$

$$M \frac{\left(1 - a_1 - \frac{v}{v_0}\right)}{v} \sim F_{(v, v_0)} \tag{3.20}$$

Exemplo 3.3

Os volumes de madeira com e sem casca (em m^3) de dois clones de Acácia-negra (A e B) foram avaliados com amostras de 30 árvores de cada clone. Os dados estão apresentados na Tabela 3.3. Testar, ao nível de 5% de significância, se as matrizes covariâncias são as mesmas para os dois clones.

Tabela 3.3: Volume com casca (vcc) e sem casca (vsc) para dois clones (A e B) da espécie acácia-negra.

Clone	n	vcc	vsc	Clone	n	vcc	vsc
A	1	0,05627	0,04566	B	1	0,04058	0,03349
A	2	0,18012	0,15376	B	2	0,09321	0,07654
A	3	0,29275	0,24932	B	3	0,17324	0,14701
A	4	0,05525	0,04569	B	4	0,15231	0,13220
A	5	0,12055	0,10179	B	5	0,20407	0,17811
A	6	0,21626	0,18614	B	6	0,07278	0,06258
A	7	0,05468	0,04580	B	7	0,06487	0,05284
A	8	0,11396	0,09670	B	8	0,19848	0,16709
A	9	0,25274	0,20748	B	9	0,15004	0,12780
A	10	0,04135	0,03316	B	10	0,03816	0,03125
A	11	0,11727	0,09503	B	11	0,11024	0,09459
A	12	0,14707	0,11882	B	12	0,18531	0,15747
A	13	0,03518	0,03071	B	13	0,04311	0,03596
A	14	0,13309	0,11207	B	14	0,15980	0,12702
A	15	0,22043	0,18568	B	15	0,33257	0,28749
A	16	0,04550	0,03724	B	16	0,18213	0,15175
A	17	0,09090	0,07592	B	17	0,38640	0,33046
A	18	0,15289	0,12450	B	18	0,16669	0,13862
A	19	0,04923	0,03961	B	19	0,01863	0,01478
A	20	0,13854	0,11848	B	20	0,06900	0,05616
A	21	0,22338	0,19068	B	21	0,12995	0,10654
A	22	0,04144	0,03249	B	22	0,05183	0,04243
A	23	0,21322	0,18033	B	23	0,13582	0,11089
A	24	0,18346	0,15311	B	24	0,23464	0,19872
A	25	0,04857	0,04109	B	25	0,03010	0,02526
A	26	0,13621	0,11484	B	26	0,11272	0,09614
A	27	0,31096	0,26852	B	27	0,21245	0,17713
A	28	0,05460	0,04813	B	28	0,04858	0,03935
A	29	0,12250	0,09829	B	29	0,11949	0,09769
A	30	0,19152	0,16087	B	30	0,23579	0,20158

As matrizes de variância-covariância amostrais são:

$$S_A = \begin{bmatrix} 0{,}00636 & 0{,}00543 \\ 0{,}00543 & 0{,}00464 \end{bmatrix} \quad S_B = \begin{bmatrix} 0{,}00783 & 0{,}00674 \\ 0{,}00674 & 0{,}00580 \end{bmatrix}$$

O valor da estatística M de Box ($MC^{-1} = 0{,}92$) foi menor do que o valor crítico de Qui-quadrado ($\chi^2 = 7{,}81$), revelando que as matrizes variância-covariância são homogêneas, ao nível de 5% de significância.

3.3 TESTE DE T^2 DE HOTELLING

No caso univariado, o teste de *t* para uma amostra com somente uma variável medida em cada unidade de amostra é familiar. Assume-se que uma amostra aleatória de uma variável $x_1, x_2, \ldots, x_n$, com n medidas é disponível e oriunda de uma população $N(\mu, \sigma^2)$. Os parâmetros μ e σ^2 são estimados por $\bar{y}$ e s^2. Para testar a hipótese $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ é utilizado a estatística *t*.

Para o caso multivariado, no qual p variáveis são medidas em cada unidade, assume-se que uma amostra aleatória para cada uma das variáveis $x_1, x_2, \ldots, x_p$ é disponível de uma população $N(\mu, \sum)$, onde cada x_p contém n medidas. Os valores de $\underline{\mu}$ são estimados por $\underline{\bar{x}}$ e $\sum$ por S.

3.3.1 Teste para um vetor de médias $\underline{\mu}$

O procedimento utilizado para testar um vetor de médias contra um valor fixo é realizado por meio do teste T^2 de Hotelling. Este teste é análogo ao teste *t-student* univariado. As hipóteses testadas são:

$$\begin{cases} H_o: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu_{1,0} \\ \mu_{2,0} \\ \vdots \\ \mu_{p,0} \end{bmatrix} \\ H_1: \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \neq \begin{bmatrix} \mu_{1,0} \\ \mu_{2,0} \\ \vdots \\ \mu_{p,0} \end{bmatrix} \end{cases} \quad \begin{cases} H_o: \underline{\mu} = \underline{\mu}_0 \\ H_1: \underline{\mu} \neq \underline{\mu}_0 \end{cases}$$

A estatística T^2 de Hotelling é dada por:

$$T^2 = n\left(\underline{\overline{X}} - \underline{\mu}_0\right)' S^{-1}\left(\underline{\overline{X}} - \underline{\mu}_0\right) \tag{3.21}$$

Em que:

$$\underline{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.22}$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(\underline{X}_i - \underline{\overline{X}}\right)\left(\underline{X}_i - \underline{\overline{X}}\right)' \tag{3.23}$$

$$\underline{\mu}_0' = [\mu_{1,0} \quad \mu_{2,0} \quad \cdots \quad \mu_{p,0}] \tag{3.24}$$

A conclusão do teste é dada por:

$$T^2 > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

para rejeitar H_0.

Quando $n - p$ for maior do que 30, o valor de T^2 assume uma distribuição de χ^2, e a conclusão do teste torna-se $n\left(\underline{\overline{X}} - \underline{\mu}_0\right)' S^{-1}\left(\underline{\overline{X}} - \underline{\mu}_0\right) > \chi_p^2(\alpha)$.

Exemplo 3.4

Para uma amostra de tamanho $n = 10$ foram medidas as variáveis volume de madeira sem casca de árvores vivas (Vsc m^3) e volume de madeira com casca de árvores vivas (Vcc m^3), em parcelas de 450 m^2. Os valores estão apresentados na Tabela 3.4.

Calcule a estatística T^2 de Hotelling, considerando o vetor de médias: $\underline{\mu}_0' = [8 \quad 10]$, que representa a produção padrão desejada para as variáveis, respectivamente.

O resultado do teste revelou que o vetor de médias da amostra não difere do valor padrão, uma vez que T^2 (7,25) foi menor do que o valor tabelado de F (10,04).

A estatística T^2 pode ser interpretada como a distância padronizada entre o vetor de médias da amostra observada e o vetor de médias padrão. Se o vetor de médias é bem distante do vetor padrão, existe evidências que os dois vetores sejam diferentes estatisticamente.

Tabela 3.4: Valores das variáveis medidas obtidos para uma amostra em um povoamento de acácia-negra.

n	Vsc	Vcc
1	7,63	9,50
2	7,90	9,80
3	7,98	9,88
4	8,07	9,98
5	8,17	10,13
6	8,32	9,45
7	8,33	10,30
8	8,35	10,31
9	8,41	10,26
10	8,53	10,13

3.3.2 Comparação entre vetores médios de 2 populações

3.3.2.1 *Vetores pareados*

Assume-se um pareamento natural entre as unidades amostrais, nas quais foram medidas p variáveis em cada unidade amostral. Desta forma, $x_{11j}, x_{12j}, \ldots, x_{1pj}$ para o primeiro grupo está pareado com $x_{21j}, x_{22j}, \ldots, x_{2pj}$ do segundo grupo. Em termos de dois tratamentos aplicados para cada unidade experimental j, a situação experimental é como segue.

Amostra	Tratamento 1		Tratamento 2	
	X_{11j}	X_{12j}	X_{21j}	X_{22j}
1				
2				
3				
.				
.				
.				
n				

As distâncias pareadas das variáveis aleatórias tornam-se:

$$d_{1j} = X_{11j} - X_{21j}$$
$$d_{2j} = X_{12j} - X_{22j}$$
$$\vdots$$
$$d_{pj} = X_{1pj} - X_{2pj}$$

Uma vez calculadas as distâncias, para cada variável, a matriz de distâncias pode ser obtida:

$$D_j' = [\underline{d}_1 \quad \underline{d}_2 \quad \cdots \quad \underline{d}_p] \tag{3.25}$$

O vetor médio de distâncias é dado por:

$$\overline{\underline{D}} = \frac{1}{n}\sum_{j=1}^{n} \underline{D}_j \tag{3.26}$$

A matriz variância-covariância das distâncias é dada por:

$$S_d = \frac{1}{n-1}\sum_{j=1}^{n} \left(\underline{D}_j - \overline{\underline{D}}\right)\left(\underline{D}_j - \overline{\underline{D}}\right)' \tag{3.27}$$

Uma vez obtido esses valores, a estatística de Hotelling é dada por:

$$T^2 = n\left(\overline{\underline{D}}\right)' S_d^{-1}\left(\overline{\underline{D}}\right) \tag{3.28}$$

As hipóteses testadas são:

$$H_0: \underline{\delta} = \underline{0}$$
$$H_1: \underline{\delta} \neq \underline{0}$$

A interpretação do teste é feita comparando-se o valor de T^2 calculado com o valor de F, sendo que, se o valor for menor do que F aceita-se a hipótese da nulidade (H_0).

$$T^2 \geq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

Quando $n - p$ for ≥ 30 o valor de $F_{p,n-p}$ é equivalente a χ_p^2.

Exemplo 3.5

Pilhas de madeira foram medidas em suas três dimensões (comprimento, altura e largura) por dois métodos distintos, com trena (método A) e imagens obtidas com drones (método B), para cada uma das 12 pilhas amostradas. Testar a hipótese de que não há diferenças entre os resultados obtidos pelos dois métodos, ao nível de 1% de significância, para os dados da Tabela 3.5.

A análise revelou que não existe diferença entre as medições com trena e a partir de imagens de drone, ao nível de 1% de significância, uma vez que T^2 (15,15) foi menor do que o valor tabelado de F (25,64).

Tabela 3.5: Dados das dimensões comprimento (m), altura (m) e largura (m) de 12 pilhas de madeira, após o baldeio, mensuradas por dois métodos, em uma área de colheita de acácia-negra.

Pilha	Método A			Método B		
	Comprimento	Altura	Largura	Comprimento	Altura	Largura
1	50,2	2,9	2,1	49,9	2,8	1,9
2	54,1	3,1	2,2	52,3	3,2	2,3
3	48,9	2,4	2,4	46,7	2,2	2,5
4	50,6	2,7	2,3	49,0	2,9	2,0
5	57,4	3,0	1,9	54,1	3,0	1,7
6	49,6	3,7	1,8	48,2	3,2	2,0
7	45,7	2,8	2,0	46,4	3,0	1,8
8	44,0	2,9	2,0	44,0	2,7	1,6
9	52,1	3,2	2,2	49,6	2,9	2,0
10	46,7	3,1	1,7	44,7	3,3	1,5
11	49,3	3,3	1,9	50,1	3,4	1,9
12	51,8	2,8	2,0	49,9	2,9	2,4

3.3.2.2 *Vetores independentes com matrizes variância-covariância homogêneas*

Para o caso em que p variáveis são mensuradas em cada unidade amostral em dois grupos independentes com matrizes variância-covarância homogêneas, $\sum_1 = \sum_2 = \sum$, e $\sum$ desconhecido. Estas condições são necessárias para que a estatística T^2 tenha uma distribuição T^2. As hipóteses testadas são:

$$\begin{cases} H_o: \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix} \\ H_1: \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} \neq \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix} \end{cases} \begin{cases} H_o: \underline{\mu}_1 = \underline{\mu}_2 \\ H_1: \underline{\mu}_1 \neq \underline{\mu}_2 \end{cases}$$

Quando $\sum_1 = \sum_2 = \sum$, aplicado em pequenas amostras independentes, a estatística utilizada é:

$$T^2 = \left[\left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right)\right]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_p\right]^{-1} \left[\left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right)\right] \quad (3.29)$$

O vetor de médias $\underline{\overline{X}}_1$ estima $\underline{\mu}_1$ e o vetor de médias $\underline{\overline{X}}_2$ estima $\underline{\mu}_2$.

As matrizes variância-covariância S_1 e S_2 são utilizadas para calcular a matriz variância-covariância ponderada S_p que estima $\sum$, sendo dada por:

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \quad (3.30)$$

Em que: n_1 e n_2 são os tamanhos das amostras de cada grupo.

O resultado de T^2 obtido é comparado com o valor de F para testar H_0. Se

$$T^2 \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} > F_{p,n_1+n_2-p-1}(\alpha)$$

então H_0 é rejeitada.

Exemplo 3.6

Os teores de materiais voláteis, carbono fixo e cinzas, em dois componentes, copa e fuste, de uma amostra de acácia-negra foram medidos e obtidos em porcentagem. Deseja-se comparar se existe diferença dos teores entre os dois componentes, a partir dos dados da Tabela 3.6.

O valor da estatística M de Box ($MC^{-1} = 3{,}42$) foi de menor do que o valor crítico de Qui-quadrado ($\chi^2 = 12{,}59$), revelando que as matrizes variância-covariância são homogêneas, ao nível de 5% de significância.

O teste T^2 revelou existir diferença significante entre os teores dos componentes testados, ao nível de 5% de significância, uma vez que o valor de T^2 transformado para F calculado (35,77) foi maior do que o valor tabelado de F (2,92).

Tabela 3.6: Teores de materiais voláteis, carbono fixo e cinzas para amostras independentes de acácia-negra, obtida em um povoamento com 10 anos.

n	Componente Copa			Componente Fuste		
	TEOR DE VOLATEIS (%)	TEOR DE CARBONO FIXO (%)	TEOR DE CINZAS (%)	TEOR DE VOLATEIS (%)	TEOR DE CARBONO FIXO (%)	TEOR DE CINZAS (%)
1	72,25	25,02	2,73	77,71	21,49	0,79
2	71,54	25,67	2,78	77,04	22,02	0,94
3	71,79	25,56	2,65	79,34	19,95	0,72
4	73,21	25,06	1,73	79,02	19,26	1,72
5	70,64	26,74	2,62	77,05	22,35	0,60
6	70,70	27,12	2,19	75,09	23,82	1,09
7	69,65	27,39	2,96	77,86	21,19	0,95
8	69,86	28,04	2,10	76,30	22,64	1,06
9	75,71	23,50	0,79	79,82	18,19	1,99
10	72,84	25,09	2,07	74,61	23,53	1,86
11	71,20	26,13	2,67	79,59	19,70	0,70
12	70,48	27,44	2,08	78,31	20,79	0,89
13	71,76	25,36	2,88	75,24	23,39	1,36
14	69,26	28,09	2,65	75,26	23,19	1,55
15	72,51	25,89	1,61	78,08	21,14	0,78
16	69,42	28,84	1,74	76,54	21,65	1,81
17	71,80	25,89	2,30	79,34	20,03	0,63

3.3.2.3 *Vetores independentes com matrizes variância-covariância heterogêneas*

Quando as matrizes variância-covariância são heterogêneas, $\sum_1 \neq \sum_2$ e desconhecidas, o procedimento adotado é similar ao utilizado quando as variâncias são homogêneas, com pequenas variações nos cálculos. Esse procedimento é utilizado para amostras de tamanhos grandes, em que $n_1 - p$ e $n_2 - p$ são maiores do que 30. As hipóteses testadas são:

$$\begin{cases} H_0: \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix} \\ H_1: \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} \neq \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix} \end{cases} \begin{cases} H_o: \underline{\mu}_1 = \underline{\mu}_2 \\ H_1: \underline{\mu}_1 \neq \underline{\mu}_2 \end{cases}$$

A estatística do teste de T^2 de Hotelling é obtida por:

$$T^2 = \left[\left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right)\right]' \left[\left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right)\right]^{-1} \left[\left(\underline{\overline{X}}_1 - \underline{\overline{X}}_2\right)\right] \quad (3.31)$$

A rejeição da hipótese de nulidade H_0 se dá quando $T^2 > \chi_p^2(\alpha)$.

Exemplo 3.7

Dois tipos de preparo do solo (A e B) foram aplicados para o plantio da espécie acácia-negra. Após sete anos foram feitas análises de solo e deseja-se saber se os teores dos elementos são diferentes para os dois tipos de preparo. Os dados dos teores dos elementos estão apresentados na Tabela 3.7.

Tabela 3.7: Teores de Cu, Zn, Mn, Na e Fe para duas amostras independentes com matrizes variância-covariância heterogêneas, obtidos em um povoamento de acácia-negra com 7 anos.

n	Grupo	Cu	Zn	Mn	Na	Fe	n	Grupo	Cu	Zn	Mn	Na	Fe
1	A	0,6	0,3	33	11	0,09	1	B	0,6	0,3	127	10	0,08
2	A	0,6	0,3	64	7	0,04	2	B	0,6	0,3	87	7	0,06
3	A	0,2	0,3	71	6	0,06	3	B	0,6	0,3	69	20	0,07
4	A	0,1	0,7	50	9	0,05	4	B	0,2	0,4	58	8	0,05
5	A	0,6	0,4	98	6	0,07	5	B	0,6	0,3	42	8	0,04
6	A	0,2	0,6	24	9	0,08	6	B	0,6	0,4	32	7	0,04
7	A	0,2	0,4	66	5	0,06	7	B	0,2	0,3	37	8	0,08
8	A	0,4	0,4	105	8	0,08	8	B	0,8	0,5	52	20	0,07
9	A	0,8	0,6	113,8	7	0,1	9	B	0,9	0,4	87	5	0,07
10	A	0,6	0,4	73	6	0,05	10	B	0,6	0,5	44	8	0,06
11	A	0,9	0,4	78	5	0,07	11	B	0,4	0,4	49	8	0,06
12	A	0,1	0,5	31	10	0,05	12	B	0,4	0,7	61	15	0,12
13	A	0,6	0,5	171	8	0,08	13	B	0,2	0,3	78	7	0,07
14	A	0,2	0,6	78	9	0,06	14	B	0,2	0,4	53	11	0,07
15	A	0,2	0,4	56	6	0,06	15	B	1,1	0,5	75	8	0,09
16	A	0,4	0,4	78	8	0,07	16	B	0,9	0,4	73	5	0,07
17	A	0,1	0,3	64	6	0,04	17	B	0,6	0,5	99	6	0,07
18	A	0,2	0,4	50	10	0,07	18	B	0,4	0,4	78	9	0,06
19	A	0,6	0,4	24	8	0,07	19	B	0,4	0,3	40	6	0,1
20	A	0,1	0,4	32	5	0,07	20	B	0,9	0,3	80	5	0,06
21	A	0,2	0,5	71	6	0,08	21	B	0,8	0,3	76	13	0,06
22	A	0,4	0,4	76	6	0,06	22	B	0,9	0,3	87	10	0,08
23	A	0,2	0,4	18	7	0,08	23	B	0,6	0,3	44	10	0,07
24	A	0,2	0,6	29	6	0,07	24	B	0,8	0,3	171	10	0,06
25	A	0,4	0,4	29	8	0,1	25	B	0,9	0,5	40	9	0,08
26	A	0,4	0,3	73	6	0,07	26	B	0,6	0,4	57	9	0,09
27	A	0,4	0,4	71	13	0,09	27	B	0,9	0,4	104	7	0,07
28	A	0,7	0,5	146	7	0,08	28	B	0,8	0,4	71	7	0,05
29	A	0,2	0,5	20	7	0,12	29	B	0,8	0,3	66	6	0,05
30	A	0,2	0,6	91	11	0,05	30	B	1,2	0,4	118	6	0,08
31	A	0,1	0,3	93	8	0,1	31	B	0,4	0,4	41	8	0,13
32	A	0,2	0,6	21	10	0,12	32	B	0,9	0,5	107	12	0,6
33	A	0,6	0,5	91	6	0,08	33	B	0,1	0,4	41	7	0,1
34	A	0,2	0,3	53	7	0,07	34	B	0,4	0,6	95	20	0,12
35	A	0,4	0,4	54	7	0,06	35	B	0,2	0,8	79	13	0,07
36	A	0,4	0,4	49	4	0,1	36	B	0,4	0,5	59	8	0,09
37	A	0,2	0,4	71	8	0,07	37	B	0,4	0,4	82	6	0,05

O valor da estatística M de Box ($MC^{-1} = 85{,}32$) foi maior do que o valor crítico de Qui-quadrado ($\chi^2 = 24{,}99$), revelando que as matrizes variância-covariância são

heterogêneas, ao nível de 5% de significância. O teste T^2 revelou existir diferença significante entre os dois tipos de preparo do solo ao nível de 5% de significância, uma vez que o valor de T^2 (30,94) foi maior do que o valor tabelado de χ^2 (11,07).

3.4 ANÁLISE DE VARIÂNCIA MULTIVARIADA – MANOVA

Na análise de variância univariada, quando se tem dois ou mais grupos (tratamentos), a soma total dos quadrados é decomposta em soma de quadrados dentro dos grupos e soma de quadrados entre grupos, com o intuito de se testar a hipótese de não diferença entre as médias dos grupos. A variância do grupo dentro é assumida ser homogênea. A hipótese é testada pelo teste F.

A extensão multivariada deste procedimento ocorre quando existem várias variáveis resposta sendo analisadas simultaneamente. A hipótese multivariada correspondente é feita assumindo que as matrizes de variância-covariância dentro dos grupos são homogêneas. O procedimento para a comparação de vetores de médias de mais de dois grupos está apresentado na seção 6, análise de variância multivariada – MANOVA.

3.5 SOLUÇÃO DO EXEMPLO 3.1: VERIFICAÇÃO DA NORMALIDADE MULTIVARIADA NO MATLAB E JULIA

```
% SOLUÇÃO DO EXEMPLO 3.1 NO MATLAB

% VARIÁVEIS DAP E ALTURA
X =
   19.5800   19.7000
    7.4800   14.9000
   16.0700   18.2000
    9.2300   15.7000
    9.3900   11.8000
    5.4100   12.0000
    8.4400   12.0000
   10.1900   18.5000
   18.9400   18.8000
   23.5500   21.7000

% MÉDIA
MEDIA=mean(X)
MEDIA =
   12.8280   16.3300

% MATRIZ VARIÂNCIA-COVARIÂNCIA
S=cov(X)
S =
   38.1139   19.1527
   19.1527   12.7734

% INVERSA DA MARIZ VARIÂNCIA-COVARIÂNCIA
INVS=inv(S)
INVS =
    0.1064   -0.1596
   -0.1596    0.3176

% CÁLCULOS DAS DISTÃNCIAS GENERALIZADA
D2J=sum((((X-MEDIA)*INVS).*(X-MEDIA)),2)
```

```
D2J =
    1.1963
    1.2525
    0.2942
    0.7804
    2.8040
    1.5590
    1.9392
    4.0631
    1.0950
    3.0164

% TAMANHO DA AMOSTRA (n) E NÚMERO DE VARIÁVEIS (p)
[n,p]=size(X)
n =
    10

p =
     2

% OBTENÇÃO DO VALOR DE QUI-QUADRADO TABELADO
VALORTABELADO=chi2inv(0.5,p)
VALORTABELADO =
    1.3863

% NÚMERO DE VOLORES MENORES OU IGUAIS AO VALOR DE QUI-QUADRADO
TABELADO
PROPORCAOMENOR=nnz((D2J(:,:)<=VALORTABELADO))
PROPORCAOMENOR =
     5

#SOLUÇÃO DO EXEMPLO 3.1 NO JULIA
using DataFrames, CSV, Statistics, LinearAlgebra, Distributions, PyPlot,
LinearRegression
#Variáveis DAP e Altura
X=[19.58    19.7
7.48    14.9
16.07   18.2
9.23    15.7
9.39    11.8
5.41    12
8.44    12
10.19   18.5
18.94   18.8
23.55   21.7]

#Média
MEDIA=mean(X, dims=1)

#Matriz variância-covariância
S=cov(X)

#Inversa da matriz variância-covariância
INVS=inv(S)
```

```
#Cálculo da distância generalizada
D2j=sum(((X.-MEDIA)*INVS).*(X.-MEDIA),dims=2)

#Tamanho da amostra (n) e número de variáveis (p)
(n,p)=size(X)

#Obtenção do valor de Qui-quadrado tabelado
QUIQUADTAB = quantile(Chisq(p), 0.5)

#Filtro de valores menores ou igual ao valor de Qui-Quadrado tabelado
Contagem=(filtered = D2j[(D2j[:,:] .<= VALORDEQUI)])

#Número de valores menores ou iguais ao valor de Qui-Quadrado tabelado
ProporçãoMenor=length(Contagem)
```

3.6 SOLUÇÃO DO EXEMPLO 3.2: VERIFICAÇÃO DA NORMALIDADE MULTIVARIADA NO MATLAB E JULIA

```
% SOLUÇÃO DO EXEMPLO 3.2 NO MATLAB

% ORDENAÇÃO DAS DISTÂNCIAS GENERALIZADAS CALCULADAS NO EXEMPLO 3.1
D2JORD=sort(D2J)
D2JORD =
    0.2942
    0.7804
    1.0950
    1.1963
    1.2525
    1.5590
    1.9392
    2.8040
    3.0164
    4.0631

% OBTENÇÃO DO VETOR DE POSIÇOES (j)
POS=[1:1:n]'
POS =
     1
     2
     3
     4
     5
     6
     7
     8
     9
    10

% VALORES DE ALFA PARA CADA POSIÇÃO (j)
ALFA=(POS-0.5)/n
ALFA =
    0.0500
    0.1500
    0.2500
    0.3500
    0.4500
```

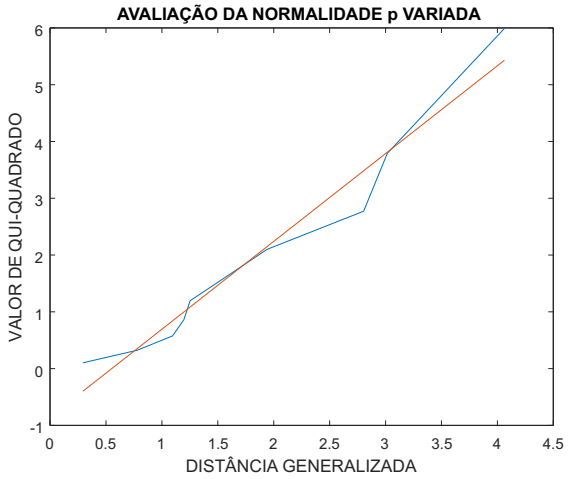
```
    0.5500
    0.6500
    0.7500
    0.8500
    0.9500

% VALORES DE QUI-QUDRADO PARA CADA POSIÇÃO (j)
y=chi2inv(ALFA,p)
y =
    0.1026
    0.3250
    0.5754
    0.8616
    1.1957
    1.5970
    2.0996
    2.7726
    3.7942
    5.9915

% AJUSTE DA EQUAÇÃO DA RETA PARA O GRÁFICO
coef=polyfit(D2JORD,y,1)
coef =
    1.5466   -0.8524
b0=coef(2)
b0 =
   -0.8524
b1=coef(1)
b1 =
    1.5466

% ESTIMATIVAS DOS VALORES DA RETA PARA O GRÁFICO
ybest=b0+b1*D2JORD
ybest =
   -0.3974
    0.3545
    0.8411
    0.9978
    1.0848
    1.5587
    2.1467
    3.4844
    3.8128
    5.4317

% GRÁFICO PARA AVALIAÇÃO DA NORMALIDADE p VARIADA
plot(D2JORD,y,D2JORD,ybest)
xlabel('DISTÂNCIA GENERALIZADA')
ylabel('VALOR DE QUI QUADRADO')
title('AVALIAÇÃO DA NORMALIDADE p VARIADA')
ylabel('VALOR DE QUI-QUADRADO')
```




```
#SOLUÇÃO DO EXEMPLO 3.2 NO JULIA

#Ordenação das distâncias generalizadas cáculadas no exemplo 3.1
D2JORD=sort(D2j,dims=1)

#Obtenção do vetor de posições (j)
POS=collect(1:1:n)

#Valores de alfa para cada posição j
ALFA=(POS.-0.5)./n

#Valores de Qui-quadrado para cada posição j
y = quantile(Chisq(p), ALFA)

#Gráfico para avaliação da normalidade p variada
plot(D2JORD,y)
gcf()

#Ajuste da equação da reta
lr = linregress(D2JORD, y)

#Estimativa dos valores da reta para o gráfico
ybest = lr(D2JORD)

#Gráfico para avaliação da normalidade p variada
plot(D2JORD,ybest)
xlabel("DISTÂNCIA GENERALIZADA")
ylabel("VALOR DE QUI-QUADRADO")
title("AVALIAÇÃO DA NORMALIDADE p VARIADA")
gcf()
```

3.7 OBTENÇÃO DA FIGURA 3.4 NO MATLAB E JULIA

```
% OBTENÇÃO DA FIGURA 3.4 NO MATLAB

% GRÁFICO DA DISTRIBUIÇÃO NORMAL BIVARIADA PARA AS VARIÁVEIS
  DIÂMETRO À ALTURA DO PEITO E ALTURA COM CORRELAÇÃO ALTA (r=0,88)

% VETOR DE MÉDIAS CENTRALIZADO
mu=[0 0]
mu =
     0     0

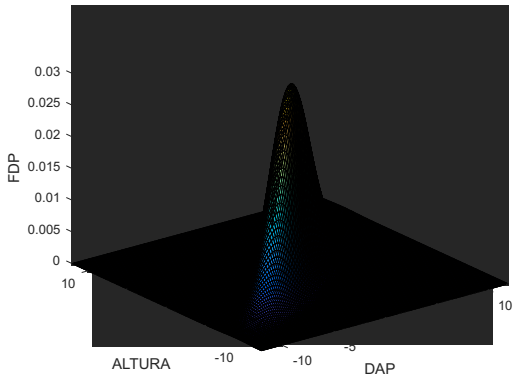
% MATRIZ VARIÂNCIA-COVARIÂNCIA DAS VARIÁVEIS ORIGINAIS
S=[13.38 9.80 ; 9.80 9.31]
S =
   13.3800    9.8000
    9.8000    9.3100

% ORDENADAS PARA AS VARIÁVEIS
x1=-12 : 0.2 : 12;
x2=-12 : 0.2 : 12;

% PLOTAGEM DO GRID
[X1,X2]=meshgrid(x1,x2);
X=[X1(:) X2(:)];

% CÁLCULO DOS VALORES DE DENSIDADE PROBABILÍSTICA
y=mvnpdf(X,mu,S);
y=reshape(y,length(x2),length(x1));

% OBTENÇÃO DO GRÁFICO
surf(x1,x2,y)
axis([-12 12 -12 12 0 0.03])
xlabel('DAP')
ylabel('ALTURA')
zlabel('FDP')
```




```
% GRÁFICO DA DISTRIBUIÇÃO NORMAL BIVARIADA PARA AS VARIÁVEIS TEOR
  DE CARBONO NO FUSTE E ALTURA DO FUSTE COM CORRELAÇÃO BAIXA (r=-
  0,056)

% VETOR DE MÉDIAS CENTRALIZADO
mu=[0 0]
mu =
     0     0
```

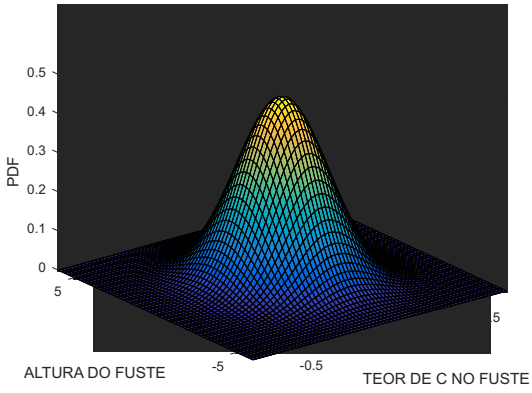
```
% MATRIZ VARIÂNCIA-COVARIÂNCIA DAS VARIÁVEIS ORIGINAIS
S=[0.035 -0.032 ; -0.032 3.312]
S =
    0.0350   -0.0320
   -0.0320    3.3120

% ORDENADAS PARA AS VARIÁVEIS
x1=-0.7 : 0.02 : 0.7;
x2=-6 : 0.2 : 6;

% PLOTAGEM DO GRID
[X1,X2]=meshgrid(x1,x2);
X=[X1(:) X2(:)];

% CÁLCULO DOS VALORES DE DENSIDADE PROBABILÍSTICA
y=mvnpdf(X,mu,S);
y=reshape(y,length(x2),length(x1));

% OBTENÇÃO DO GRÁFICO
surf(x1,x2,y)
axis([-0.7 0.7 -6 6 0 0.5])
xlabel('TEOR DE C NO FUSTE')
ylabel('ALTURA DO FUSTE')
zlabel('PDF')
```




```
# OBTENÇÃO DA FIGURA 3.4 NO JULIA

# GRÁFICO DA DISTRIBUIÇÃO NORMAL BIVARIADA PARA AS VARIÁVEIS
# DIÂMETRO À ALTURA DO PEITO E ALTURA COM CORRELAÇÃO ALTA (r = 0.88)

using LinearAlgebra, Distributions, PyPlot

# VETOR DE MÉDIAS CENTRALIZADO
mu = [0.0, 0.0]

# MATRIZ VARIÂNCIA-COVARIÂNCIA DAS VARIÁVEIS ORIGINAIS
S = [13.38 9.80;
     9.80 9.31]

# ORDENADAS PARA AS VARIÁVEIS
x1 = -12:0.2:12
x2 = -12:0.2:12
```

```
# PLOTAGEM DO GRID (equivalente ao meshgrid do MATLAB)
X1 = repeat(collect(x1)', length(x2), 1)
X2 = repeat(collect(x2), 1, length(x1))

# ORGANIZAÇÃO DOS PONTOS EM MATRIZ (equivalente a X=[X1(:) X2(:)])
X = hcat(vec(X1), vec(X2))

# CÁLCULO DOS VALORES DE DENSIDADE PROBABILÍSTICA
# (equivalente ao mvnpdf do MATLAB)
dist = MvNormal(mu, S)
y = pdf.(dist, eachrow(X))
# REORGANIZAÇÃO PARA MATRIZ (reshape como no MATLAB)
y = reshape(y, length(x2), length(x1))

# OBTENÇÃO DO GRÁFICO
figure()
surf(x1, x2, y)
axis([-12, 12, -12, 12, 0, 0.03])
xlabel("DAP")
ylabel("ALTURA")
zlabel("FDP")

# GRÁFICO DA DISTRIBUIÇÃO NORMAL BIVARIADA PARA AS VARIÁVEIS
# TEOR DE CARBONO NO FUSTE E ALTURA DO FUSTE COM CORRELAÇÃO BAIXA (r = -0,056)

using LinearAlgebra, Distributions, PyPlot

# VETOR DE MÉDIAS CENTRALIZADO
mu = [0.0, 0.0]

# MATRIZ VARIÂNCIA-COVARIÂNCIA DAS VARIÁVEIS ORIGINAIS
S = [0.035  -0.032;
     -0.032  3.312]

# ORDENADAS PARA AS VARIÁVEIS
x1 = -0.7:0.02:0.7
x2 = -6:0.2:6

# PLOTAGEM DO GRID (equivalente ao meshgrid do MATLAB)
X1 = repeat(collect(x1)', length(x2), 1)
X2 = repeat(collect(x2), 1, length(x1))

# ORGANIZAÇÃO DOS PONTOS EM MATRIZ (equivalente a X=[X1(:) X2(:)])
X = hcat(vec(X1), vec(X2))

# CÁLCULO DOS VALORES DE DENSIDADE PROBABILÍSTICA
# (equivalente ao mvnpdf do MATLAB)
dist = MvNormal(mu, S)
y = pdf.(dist, eachrow(X))
```

```
# REORGANIZAÇÃO PARA MATRIZ (reshape como no MATLAB)
y = reshape(y, length(x2), length(x1))

# OBTENÇÃO DO GRÁFICO
figure()
surf(x1, x2, y)
axis([-0.7, 0.7, -6, 6, 0, 0.5])
xlabel("TEOR DE C NO FUSTE")
ylabel("ALTURA DO FUSTE")
zlabel("PDF")
```

3.8 SOLUÇÃO DO EXEMPLO 3.3, VERIFICAÇÃO DA HOMOGENEIDADE DA MATRIZ VARIÂNCIA-COVARIÂNCIA, NO MATLAB E NO JULIA

```
% ANÁLISE DE HOMOGENEIDADE DE MATRIZES VARIÂNCIA-COVARIÂNCIA NO
MATLAB
% SOLUÇÃO DO EXEMPLO 3.3

% MATRIZ DE DADOS DO GRUPO 1 (CLONE A)
XA =
    0.0563    0.0457
    0.1801    0.1538
    0.2928    0.2493
    0.0553    0.0457
    0.1206    0.1018
    0.2163    0.1861
    0.0547    0.0458
    0.1140    0.0967
    0.2527    0.2075
    0.0413    0.0332
    0.1173    0.0950
    0.1471    0.1188
    0.0352    0.0307
    0.1331    0.1121
    0.2204    0.1857
    0.0455    0.0372
    0.0909    0.0759
    0.1529    0.1245
    0.0492    0.0396
    0.1385    0.1185
    0.2234    0.1907
    0.0414    0.0325
    0.2132    0.1803
    0.1835    0.1531
    0.0486    0.0411
    0.1362    0.1148
    0.3110    0.2685
    0.0546    0.0481
    0.1225    0.0983
    0.1915    0.1609
```

```
% MATRIZ DE DADOS DO GRUPO 2 (CLONE B)
XB =
    0.0406    0.0335
    0.0932    0.0765
    0.1732    0.1470
    0.1523    0.1322
    0.2041    0.1781
    0.0728    0.0626
    0.0649    0.0528
    0.1985    0.1671
    0.1500    0.1278
    0.0382    0.0313
    0.1102    0.0946
    0.1853    0.1575
    0.0431    0.0360
    0.1598    0.1270
    0.3326    0.2875
    0.1821    0.1518
    0.3864    0.3305
    0.1667    0.1386
    0.0186    0.0148
    0.0690    0.0562
    0.1300    0.1065
    0.0518    0.0424
    0.1358    0.1109
    0.2346    0.1987
    0.0301    0.0253
    0.1127    0.0961
    0.2125    0.1771
    0.0486    0.0394
    0.1195    0.0977
    0.2358    0.2016

% MATRIZ VARIÂNCIA-COVARIÂNCIA DO GRUPO 1
S1=cov(XA)
S1 =

    0.0064    0.0054
    0.0054    0.0046

% MATRIZ VARIÂNCIA-COVARIÂNCIA DO GRUPO 2
S2=cov(XB)
S2 =

    0.0078    0.0067
    0.0067    0.0058

% NÚMERO DE GRUPOS
k=2
k =
     2

% TAMANHO DA AMOSTRA DO GRUPO 1 (n1) E NÚMERO DE VARIÁVEIS (p)
[n1,p]=size(XA)
n1 =
    30
p =
     2
```

```
% TAMANHO DA AMOSTRA DO GRUPO 2 (n2) E NÚMERO DE VARIÁVEIS (p)
[n2,p]=size(XB)
n2 =
    30
p =
     2

% TAMANHO TOTAL DAS AMOSTRAS
n=n1+n2
n =
    60

% GRAUS DE LIBERDADE PARA O GRUPO 1
glX1=n1-1
glX1 =
    29

% GRAUS DE LIBERDADE PARA O GRUPO 2
glX2=n2-1
glX2 =
    29

% MATRIZ VARIÂNCIA-COVARIÂNCIA COMBINADA
S=(1/(glX1+glX2))*((glX1*S1)+(glX2*S2))
S =
    0.0071    0.0061
    0.0061    0.0052

% DETERMINANTE DA MATRIZ VARIÂNCIA-COVARIÂNCIA COMBINADA
DETS=det(S)
DETS =
   4.7584e-08

% DETERMINANTE DA MATRIZ VARIÂNCIA-COVARIÂNCIA GRUPO 1
DETS1=det(S1)
DETS1 =
   4.1607e-08

% DETERMINANTE DA MATRIZ VARIÂNCIA-COVARIÂNCIA DO GRUPO 2
DETS2=det(S2)
DETS2 =
   5.2702e-08

% CÁLCULO DA ESTATÍSTICA M DE BOX
M=(glX1*log(DETS)+glX2*log(DETS))-(glX1*log(DETS1)+glX2*log(DETS2))
M =
    0.9295

% FATOR DE ESCALA PARA AMOSTRAS DE MESMO TAMANHO
C1=1-(2*p^2+3*p-1*k+1)/(6*(p+1)*k*n)
C1 =
    0.9940

% CÁLCULO DE M DE BOX CORRIGIDO
MC1=M*C1
MC1 =
    0.9239
```

```
% GRAUS DE LIBERDADE PARA OBTENÇÃO DO VALOR DE QUI-QUADRADO
Gl=0.5*(k-1)*p*(p+1)
Gl =
     3

% VALOR CRÍTICO DE QUI-QUAQDRADO
Qui2=chi2inv(0.95,Gl)
Qui2 =
    7.8147
```

```
#Análise da homogeneidade de matrizes variância-covariância
#SOLUÇÃO DO EXEMPLO 3.3 NO JULIA

using DataFrames, Statistics, LinearAlgebra

#Matriz de dados do grupo 1 (clone A)
XA=[0.05627 0.04566
0.18012 0.15376
0.29275 0.24932
0.05525 0.04569
0.12055 0.10179
0.21626 0.18614
0.05468 0.04580
0.11396 0.09670
0.25274 0.20748
0.04135 0.03316
0.11727 0.09503
0.14707 0.11882
0.03518 0.03071
0.13309 0.11207
0.22043 0.18568
0.04550 0.03724
0.09090 0.07592
0.15289 0.12450
0.04923 0.03961
0.13854 0.11848
0.22338 0.19068
0.04144 0.03249
0.21322 0.18033
0.18346 0.15311
0.04857 0.04109
0.13621 0.11484
0.31096 0.26852
0.05460 0.04813
0.12250 0.09829
0.19152 0.16087]

#Matriz de dados do grupo 2 (clone B)
XB=[0.04058 0.03349
0.09321 0.07654
0.17324 0.14701
0.15231 0.13220
```

```
0.20407 0.17811
0.07278 0.06258
0.06487 0.05284
0.19848 0.16709
0.15004 0.12780
0.03816 0.03125
0.11024 0.09459
0.18531 0.15747
0.04311 0.03596
0.15980 0.12702
0.33257 0.28749
0.18213 0.15175
0.38640 0.33046
0.16669 0.13862
0.01863 0.01478
0.06900 0.05616
0.12995 0.10654
0.05183 0.04243
0.13582 0.11089
0.23464 0.19872
0.03010 0.02526
0.11272 0.09614
0.21245 0.17713
0.04858 0.03935
0.11949 0.09769
0.23579 0.20158]

#Matriz variância-covariância do grupo1
S1=cov(XA)

#Matriz variância-covariância do grupo 2
S2=cov(XB)

#Número de grupos
k=2

#Tamanho da amostra do grupo 1 (n1) e número de variáveis (p)
(n1,p)=size(XA)

#Tamanho da amostra do grupo 2 (n2) e número de variáveis (p)
(n2,p)=size(XB)

#Tamanho total das amostras
n=n1+n2

#Graus de liberdade para o grupo 1
glX1=n1-1

#Graus de liberdade para o grupo 2
glX2=n2-1
```

```
#Matriz variância-covariância combinada
S=(1/(glX1+glX2))*((glX1*S1)+(glX2*S2))

#Determinante da matriz variância-covariância combinada
DETS=det(S)

#Determinante da matriz variância-covariância para o grupo 1
DETS1=det(S1)

#Determinante da matriz variância-covariância combinada para o grupo 2
DETS2=det(S2)

#Cálculo da estatística M de Box
M=(glX1*log(DETS)+glX2*log(DETS))-(glX1*log(DETS1)+glX2*log(DETS2))

#Fator de escala para amostras de mesmo tamanho
C1=1-(2*p^2+3*p-1*k+1)/(6*(p+1)*k*n)

#Cálculo de M de Box corrigido
MC1=M*C1

#Graus de liberdade para obtenção do valor de Qui-Quadrado
Gl=0.5*(k-1)*p*(p+1)

#Valor crítico de Qui-Quadrado
Qui2 = quantile(Chisq(Gl), 0.95)
```

3.9 SOLUÇÃO DO EXEMPLO 3.4, TESTE PARA UM VETOR DE MÉDIAS $\underline{\mu}$, NO MATLAB E JULIA

```
% TESTE PARA UM VETOR DE MÉDIAS
% SOLUÇÃO DO EXEMPLO 3.4 NO MATLAB
% MATRIZ DE DADOS PARA VOLUME SEM CASCA E VOLUME COM CASCA
X =
    7.6300    9.5000
    7.9000    9.8000
    7.9800    9.8800
    8.0700    9.9800
    8.1700   10.1300
    8.3200    9.4500
    8.3300   10.3000
    8.3500   10.3100
    8.4100   10.2600
    8.5300   10.1300

% VETOR DE MÉDIAS A SER COMPARADO(VALORES ESPERADOS)
mizero=[8 10]
mizero =
     8    10

% VETOR DE MÉDIAS
media=mean(X)
media =
    8.1690    9.9740
```

```
% MATRIZ VARIÂNCIA-COVARIÂNCIA
S=cov(X)
S =
    0.0754    0.0532
    0.0532    0.0987

% INVERSA DA MATRIZ VARIÂNCIA-COVARIÂNCIA
invS=inv(S)
invS =
   21.4347  -11.5627
  -11.5627   16.3720

% TAMANHO DA AMOSTRA (n) E NÚMERO DE VARIÁVEIS (p)
[n,p]=size(X)
n =
    10
p =
     2

% CÁLCULO DA ESTATÍSTICA T² DE HOTELLING
T2=n*(media-mizero)*invS*(media-mizero)'
T2 =
    7.2488

% CÁCULO DO VALOR CRÍTICO DE F
F=(((n-1)*p)/(n-p))*finv(0.95,p,(n-p))
F =
   10.0327
```

```
#Teste para um vetor de médias
#SOLUÇÃO DO EXEMPLO 3.4 NO JULIA

using DataFrames, Statistics, LinearAlgebra

#Matriz de dados para o volume sem casca e volume com casca
X=[7.63 9.50
7.90    9.80
7.98    9.88
8.07    9.98
8.17    10.13
8.32    9.45
8.33    10.30
8.35    10.31
8.41    10.26
8.53    10.13]

#Vetor de médias a ser comparado (valores esperados)
mizero=[8 10]

#Vetor de médias
Media=mean(X, dims=1)

#Matriz variância-covariância
S=cov(X)
```

```
#Inversa da matriz variância-covariância
invS=inv(S)

#Tamanho da amostra (n) e número de variáveis (p)
(n,p)=size(X)

#Cálculo da estatística T2 de Hotelling
T2=n*(Media-mizero)*invS*(Media-mizero)'

#Cálculo do valor crítico de F
F = (((n - 1) * p) / (n - p)) * quantile(FDist(p, n - p), 0.95)
```

3.10 SOLUÇÃO DO EXEMPLO 3.5, VETORES PAREADOS, NO MATLAB E JULIA

```
% TESTE PARA VETORES PAREADOS
% SOLUÇÃO DO EXEMPLO 3.5 NO MATLAB

% MATRIZ DE DADOS PARA O GRUPO 1 (MÉTODO A)
X1 =
   50.2000    2.9000    2.1000
   54.1000    3.1000    2.2000
   48.9000    2.4000    2.4000
   50.6000    2.7000    2.3000
   57.4000    3.0000    1.9000
   49.6000    3.7000    1.8000
   45.7000    2.8000    2.0000
   44.0000    2.9000    2.0000
   52.1000    3.2000    2.2000
   46.7000    3.1000    1.7000
   49.3000    3.3000    1.9000
   51.8000    2.8000    2.0000

% MATRIZ DE DADOS PARA O GRUPO 2 (MÉTODO B)
X2 =
   49.9000    2.8000    1.9000
   52.3000    3.2000    2.3000
   46.7000    2.2000    2.5000
   49.0000    2.9000    2.0000
   54.1000    3.0000    1.7000
   48.2000    3.2000    2.0000
   46.4000    3.0000    1.8000
   44.0000    2.7000    1.6000
   49.6000    2.9000    2.0000
   44.7000    3.3000    1.5000
   50.1000    3.4000    1.9000
   49.9000    2.9000    2.4000

% MATRIZ DAS DIFERENÇAS ENTRE GRUPOS
DIF=X1-X2
DIF =
    0.3000    0.1000    0.2000
    1.8000   -0.1000   -0.1000
    2.2000    0.2000   -0.1000
    1.6000   -0.2000    0.3000
```

```
    3.3000         0    0.2000
    1.4000    0.5000   -0.2000
   -0.7000   -0.2000    0.2000
         0    0.2000    0.4000
    2.5000    0.3000    0.2000
    2.0000   -0.2000    0.2000
   -0.8000   -0.1000         0
    1.9000   -0.1000   -0.4000

% MATRIZ VARIÂNCIA-COVARIÂNCIA DAS DIFERENÇAS
SDIF=cov(DIF)
SDIF =
    1.6863    0.0485   -0.0584
    0.0485    0.0515   -0.0073
   -0.0584   -0.0073    0.0548

% INVERSA DA MATRIZ VARIÂNCIA-COVARIÂNCIA DAS DIFERENÇAS
INVSDIF=inv(SDIF)
INVSDIF =
    0.6285   -0.5064    0.6030
   -0.5064   20.1906    2.1409
    0.6030    2.1409   19.1845

% MEDIAS DAS DIFERENÇAS
DIFMEDIA=mean(DIF)
DIFMEDIA =
    1.2917    0.0333    0.0750

% TAMANHO DA AMOSTRA (n) E NÚMERO DE VARIÁVEIS (p)
[n,p]=size(DIF)
n =
    12
p =
     3

% CÁLCULO DA ESTATÍSTICA T² DE HOTELLING
T2=n*DIFMEDIA*INVSDIF*DIFMEDIA'
T2 =
   15.1536

% VALOR DE F CRÍTICO CORRIGIDO
F=((n-1)*p)/(n-p)*finv(0.99,p,(n-p))
F =
   25.6370

#Teste para vetores pareados
#SOLUÇÃO DO EXEMPLO 3.5 NO JULIA
using DataFrames, Statistics, LinearAlgebra

#Matriz de dados para o grupo 1 (Método A)
X1=[50.2    2.9 2.1
54.1    3.1 2.2
48.9    2.4 2.4
50.6    2.7 2.3
57.4    3.0 1.9
49.6    3.7 1.8
45.7    2.8 2.0
```

```
44.0    2.9 2.0
52.1    3.2 2.2
46.7    3.1 1.7
49.3    3.3 1.9
51.8    2.8 2.0]

#Matriz de dados para o grupo 2 (Método B)
X2=[49.9    2.8 1.9
52.3    3.2 2.3
46.7    2.2 2.5
49.0    2.9 2.0
54.1    3.0 1.7
48.2    3.2 2.0
46.4    3.0 1.8
44.0    2.7 1.6
49.6    2.9 2.0
44.7    3.3 1.5
50.1    3.4 1.9
49.9    2.9 2.4]

#Matriz das diferenças entre grupos
DIF=X1-X2

#Matriz variância-covariância das diferenças
SDIF=cov(DIF)

#Inversa da matriz variância-covariância das diferenças
INVSDIF=inv(SDIF)

#Médias das diferenças
DIFMEDIA=mean(DIF,dims=1)

#Tamanho da amostra (n) e número de variáveis (p)
(n,p)=size(DIF)

#Cálculo da estatística T2 de Hotelling
T2=n*DIFMEDIA*INVSDIF*DIFMEDIA'

#Cálculo do valor de F crítico corrigido
F = ((n - 1) * p) / (n - p) * quantile(FDist(p, n - p), 0.99)
```

3.11 SOLUÇÃO DO EXEMPLO 3.6, VETORES INDEPENDENTES COM MATRIZES VARIÂNCIA-COVARIÂNCIA HOMOGÊNEAS, NO MATLAB E JULIA

```
% TESTE PARA VETORES INDEPENDENTES COM MATRIZES VARIÂNCIA-
COVÂRIANCIA HOMOGENEAS
% SOLUÇÃO DO EXEMPLO 3.6 NO MATLAB

% MATRIZ DE DADOS PARA O GRUPO 1 (COMPONENTE COPA)
X1 =
   72.2500   25.0200    2.7300
   71.5400   25.6700    2.7800
   71.7900   25.5600    2.6500
   73.2100   25.0600    1.7300
   70.6400   26.7400    2.6200
   70.7000   27.1200    2.1900
   69.6500   27.3900    2.9600
   69.8600   28.0400    2.1000
   75.7100   23.5000    0.7900
   72.8400   25.0900    2.0700
   71.2000   26.1300    2.6700
   70.4800   27.4400    2.0800
   71.7600   25.3600    2.8800
   69.2600   28.0900    2.6500
   72.5100   25.8900    1.6100
   69.4200   28.8400    1.7400
   71.8000   25.8900    2.3000

% MATRIZ DE DADOS PARA O GRUPO 1 (COMPONENTE FUSTE)
X2 =
   77.7100   21.4900    0.7900
   77.0400   22.0200    0.9400
   79.3400   19.9500    0.7200
   79.0200   19.2600    1.7200
   77.0500   22.3500    0.6000
   75.0900   23.8200    1.0900
   77.8600   21.1900    0.9500
   76.3000   22.6400    1.0600
   79.8200   18.1900    1.9900
   74.6100   23.5300    1.8600
   79.5900   19.7000    0.7000
   78.3100   20.7900    0.8900
   75.2400   23.3900    1.3600
   75.2600   23.1900    1.5500
   78.0800   21.1400    0.7800
   76.5400   21.6500    1.8100
   79.3400   20.0300    0.6300

% TAMANHO DA AMOSTRA DO GRUPO 1 (n1) E NÚMERO DE VARIÁVEIS (p)
[n1,p]=size(X1)
n1 =
    17
p =
     3

% TAMANHO DA AMOSTRA DO GRUPO 2 (n2) E NÚMERO DE VARIÁVEIS (p)
[n2,p]=size(X2)
```

```
n2 =
    17
p =
     3

% VETOR DE MÉDIAS PARA O GRUPO 1
mediaX1=mean(X1)
mediaX1 =
   71.4482   26.2841    2.2676

% VETOR DE MÉDIAS PARA O GRUPO 2
mediaX2=mean(X2)
mediaX2 =

   77.4235   21.4312    1.1435

% MATRIZ VARIÂNCIA-COVARIÂNCIA PARA O GRUPO 1
S1=cov(X1)
S1 =
    2.6369   -2.1146   -0.5223
   -2.1146    1.9191    0.1964
   -0.5223    0.1964    0.3251

% MATRIZ VARIÂNCIA-COVARIÂNCIA PARA O GRUPO 2
S2=cov(X2)
S2 =
    2.9350   -2.6899   -0.2446
   -2.6899    2.6671    0.0221
   -0.2446    0.0221    0.2227

% MATRIZ VARIÂNCIA-COVARIÂNCIA COMBINADA
Sp=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
Sp =
    2.7859   -2.4023   -0.3835
   -2.4023    2.2931    0.1093
   -0.3835    0.1093    0.2739

% CÁLCULO DA ESTATÍSTICA T² DE HOTELLING
T2=(mediaX1-mediaX2)*(inv((1/n1 + 1/n2)*(Sp)))*(mediaX1-mediaX2)'
T2 =
  114.4944

% TRANFORMAÇÃO DE T² DE HOTELLING PARA O VALOR DE F CALCULADO
Fc=(T2*(n1+n2-p-1))/((n1+n2-2)*p)
Fc =
   35.7795

% CÁLCULO DO VALOR DO F CRÍTICO
FT=finv(0.95,p,(n1+n2-p-1))
FT =
    2.9223
```

```
#Teste para vetores independentes com matrizes variância-covariância
homogêneas
#SOLUÇÃO DO EXEMPLO 3.6 NO JULIA

using DataFrames, Statistics, LinearAlgebra

#Matriz de dados para o grupo 1 (Componente copa)
X1=[72.25   25.02   2.73
71.54   25.67   2.78
71.79   25.56   2.65
73.21   25.06   1.73
70.64   26.74   2.62
70.70   27.12   2.19
69.65   27.39   2.96
69.86   28.04   2.10
75.71   23.50   0.79
72.84   25.09   2.07
71.20   26.13   2.67
70.48   27.44   2.08
71.76   25.36   2.88
69.26   28.09   2.65
72.51   25.89   1.61
69.42   28.84   1.74
71.80   25.89   2.30]

#Matriz de dados para o grupo 2 (Componente Fuste)
X2=[77.71   21.49   0.79
77.04   22.02   0.94
79.34   19.95   0.72
79.02   19.26   1.72
77.05   22.35   0.60
75.09   23.82   1.09
77.86   21.19   0.95
76.30   22.64   1.06
79.82   18.19   1.99
74.61   23.53   1.86
79.59   19.70   0.70
78.31   20.79   0.89
75.24   23.39   1.36
75.26   23.19   1.55
78.08   21.14   0.78
76.54   21.65   1.81
79.34   20.03   0.63]

#Tamanho da amostra do grupo 1 (n1) e número de variáveis (p)
(n1,p)=size(X1)

#Tamanho da amostra do grupo 2 (n2) e número de variáveis (p)
(n2,p)=size(X2)
```

```
#Vetor de médias para o grupo 1
MEDIAX1=mean(X1,dims=1)

#Vetor de médias para o grupo 2
MEDIAX2=mean(X2,dims=1)

#Matriz variância-covariância para o grupo 1
S1=cov(X1)

#Matriz variância-covariância para o grupo 2
S2=cov(X2)

#Matriz variância-covariância combinada
Sp=((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)

#Cálculo da estatística T2 de Hotelling
T2=(MEDIAX1-MEDIAX2)*(inv((1/n1 + 1/n2)*(Sp)))*(MEDIAX1-MEDIAX2)'

#Transformação de T2 de Hotelling para o valor de F calculado
Fc=(T2*(n1+n2-p-1))/((n1+n2-2)*p)

#Cálculo do valor de F crítico
FT = quantile(FDist(p, n1 + n2 - p - 1), 0.95)
```

3.12 SOLUÇÃO DO EXEMPLO 3.7, VETORES INDEPENDENTES COM MATRIZES VARIÂNCIA-COVARIÂNCIA HETEROGÊNEAS, NO MATLAB E JULIA

```
% TESTE PARA VETORES INDEPENDENTES COM COM MATRIZES VARIÂNCIA-
COVÂRIANCIA HETEROGÊNEA
% SOLUÇÃO DO EXEMPLO 3.7 NO MATLAB

% MATRIZ DE DADOS PARA O GRUPO 1 (PREPARO DO SOLO A)
X1 =
    0.6000    0.3000   33.0000   11.0000    0.0900
    0.6000    0.3000   64.0000    7.0000    0.0400
    0.2000    0.3000   71.0000    6.0000    0.0600
    0.1000    0.7000   50.0000    9.0000    0.0500
    0.6000    0.4000   98.0000    6.0000    0.0700
    0.2000    0.6000   24.0000    9.0000    0.0800
    0.2000    0.4000   66.0000    5.0000    0.0600
    0.4000    0.4000  105.0000    8.0000    0.0800
    0.8000    0.6000  113.8000    7.0000    0.1000
    0.6000    0.4000   73.0000    6.0000    0.0500
    0.9000    0.4000   78.0000    5.0000    0.0700
    0.1000    0.5000   31.0000   10.0000    0.0500
    0.6000    0.5000  171.0000    8.0000    0.0800
    0.2000    0.6000   78.0000    9.0000    0.0600
    0.2000    0.4000   56.0000    6.0000    0.0600
    0.4000    0.4000   78.0000    8.0000    0.0700
    0.1000    0.3000   64.0000    6.0000    0.0400
    0.2000    0.4000   50.0000   10.0000    0.0700
    0.6000    0.4000   24.0000    8.0000    0.0700
```

```
    0.1000    0.4000   32.0000    5.0000    0.0700
    0.2000    0.5000   71.0000    6.0000    0.0800
    0.4000    0.4000   76.0000    6.0000    0.0600
    0.2000    0.4000   18.0000    7.0000    0.0800
    0.2000    0.6000   29.0000    6.0000    0.0700
    0.4000    0.4000   29.0000    8.0000    0.1000
    0.4000    0.3000   73.0000    6.0000    0.0700
    0.4000    0.4000   71.0000   13.0000    0.0900
    0.7000    0.5000  146.0000    7.0000    0.0800
    0.2000    0.5000   20.0000    7.0000    0.1200
    0.2000    0.6000   91.0000   11.0000    0.0500
    0.1000    0.3000   93.0000    8.0000    0.1000
    0.2000    0.6000   21.0000   10.0000    0.1200
    0.6000    0.5000   91.0000    6.0000    0.0800
    0.2000    0.3000   53.0000    7.0000    0.0700
    0.4000    0.4000   54.0000    7.0000    0.0600
    0.4000    0.4000   49.0000    4.0000    0.1000
    0.2000    0.4000   71.0000    8.0000    0.0700

% MATRIZ DE DADOS PARA O GRUPO 1 (PREPARO DO SOLO B)
X2 =
    0.6000    0.3000  127.0000   10.0000    0.0800
    0.6000    0.3000   87.0000    7.0000    0.0600
    0.6000    0.3000   69.0000   20.0000    0.0700
    0.2000    0.4000   58.0000    8.0000    0.0500
    0.6000    0.3000   42.0000    8.0000    0.0400
    0.6000    0.4000   32.0000    7.0000    0.0400
    0.2000    0.3000   37.0000    8.0000    0.0800
    0.8000    0.5000   52.0000   20.0000    0.0700
    0.9000    0.4000   87.0000    5.0000    0.0700
    0.6000    0.5000   44.0000    8.0000    0.0600
    0.4000    0.4000   49.0000    8.0000    0.0600
    0.4000    0.7000   61.0000   15.0000    0.1200
    0.2000    0.3000   78.0000    7.0000    0.0700
    0.2000    0.4000   53.0000   11.0000    0.0700
    1.1000    0.5000   75.0000    8.0000    0.0900
    0.9000    0.4000   73.0000    5.0000    0.0700
    0.6000    0.5000   99.0000    6.0000    0.0700
    0.4000    0.4000   78.0000    9.0000    0.0600
    0.4000    0.3000   40.0000    6.0000    0.1000
    0.9000    0.3000   80.0000    5.0000    0.0600
    0.8000    0.3000   76.0000   13.0000    0.0600
    0.9000    0.3000   87.0000   10.0000    0.0800
    0.6000    0.3000   44.0000   10.0000    0.0700
    0.8000    0.3000  171.0000   10.0000    0.0600
    0.9000    0.5000   40.0000    9.0000    0.0800
    0.6000    0.4000   57.0000    9.0000    0.0900
    0.9000    0.4000  104.0000    7.0000    0.0700
    0.8000    0.4000   71.0000    7.0000    0.0500
    0.8000    0.3000   66.0000    6.0000    0.0500
    1.2000    0.4000  118.0000    6.0000    0.0800
    0.4000    0.4000   41.0000    8.0000    0.1300
    0.9000    0.5000  107.0000   12.0000    0.6000
    0.1000    0.4000   41.0000    7.0000    0.1000
    0.4000    0.6000   95.0000   20.0000    0.1200
    0.2000    0.8000   79.0000   13.0000    0.0700
    0.4000    0.5000   59.0000    8.0000    0.0900
    0.4000    0.4000   82.0000    6.0000    0.0500
```

```
% TAMANHO DA AMOSTRA DO GRUPO 1 (n1) E NÚMERO DE VARIÁVEIS (p)
 [n1,p]=size(X1)
n1 =
    37
p =
     5

% TAMANHO DA AMOSTRA DO GRUPO 2 (n2) E NÚMERO DE VARIÁVEIS (p)
[n2,p]=size(X2)
n2 =
    37
p =
     5

% VETOR DE MÉDIAS PARA O GRUPO 1
mediaX1=mean(X1)
mediaX1 =
    0.3541    0.4378   65.2919    7.4595    0.0735

% VETOR DE MÉDIAS PARA O GRUPO 2
mediaX2=mean(X2)
mediaX2 =
    0.6027    0.4081   71.8649    9.2432    0.0876

% MATRIZ VARIÂNCIA-COVARIÂNCIA PARA O GRUPO 1
S1=cov(X1)
S1 =
    0.0476   -0.0024    3.4810   -0.0644    0.0005
   -0.0024    0.0113    0.0792    0.0544    0.0003
    3.4810    0.0792 1176.3413   -5.7767   -0.0640
   -0.0644    0.0544   -5.7767    3.8108    0.0050
    0.0005    0.0003   -0.0640    0.0050    0.0004

% MATRIZ VARIÂNCIA-COVARIÂNCIA PARA O GRUPO 2
S2=cov(X2)
S2 =
    0.0780   -0.0047    3.2837   -0.1257    0.0033
   -0.0047    0.0135   -0.1517    0.1674    0.0022
    3.2837   -0.1517  854.0090    4.5893    0.4719
   -0.1257    0.1674    4.5893   15.8559    0.0642
    0.0033    0.0022    0.4719    0.0642    0.0079

% MATRIZ VARIÂNCIA-COVARIÂNCIA PONDERADA
SP=((1/n1)*S1)+((1/n2)*S2)
SP =
    0.0034   -0.0002    0.1828   -0.0051    0.0001
   -0.0002    0.0007   -0.0020    0.0060    0.0001
    0.1828   -0.0020   54.8743   -0.0321    0.0110
   -0.0051    0.0060   -0.0321    0.5315    0.0019
    0.0001    0.0001    0.0110    0.0019    0.0002

% CÁLCULO DA ESTATÍSTICA T² DE HOTELLING
T2=(mediaX1-mediaX2)*inv(SP)*(mediaX1-mediaX2)'
T2 =
   30.9432

% CÁLCULO DO VALOR DE QUI-QUADRADO CRÍTICO
QUI2=chi2inv(0.95,p)
```

```
QUI2 =
   11.0705

#Teste para vetores independentes com matrizes variância-covariância
heterogêneas
#SOLUÇÃO DO EXEMPLO 3.7 NO JULIA

using DataFrames, Statistics, LinearAlgebra

#Matriz de dados para o grupo 1 (Preparo do solo A)
X1=[0.6 0.3 33  11  0.09
0.6 0.3 64  7   0.04
0.2 0.3 71  6   0.06
0.1 0.7 50  9   0.05
0.6 0.4 98  6   0.07
0.2 0.6 24  9   0.08
0.2 0.4 66  5   0.06
0.4 0.4 105 8   0.08
0.8 0.6 113.8   7   0.1
0.6 0.4 73  6   0.05
0.9 0.4 78  5   0.07
0.1 0.5 31  10  0.05
0.6 0.5 171 8   0.08
0.2 0.6 78  9   0.06
0.2 0.4 56  6   0.06
0.4 0.4 78  8   0.07
0.1 0.3 64  6   0.04
0.2 0.4 50  10  0.07
0.6 0.4 24  8   0.07
0.1 0.4 32  5   0.07
0.2 0.5 71  6   0.08
0.4 0.4 76  6   0.06
0.2 0.4 18  7   0.08
0.2 0.6 29  6   0.07
0.4 0.4 29  8   0.1
0.4 0.3 73  6   0.07
0.4 0.4 71  13  0.09
0.7 0.5 146 7   0.08
0.2 0.5 20  7   0.12
0.2 0.6 91  11  0.05
0.1 0.3 93  8   0.1
0.2 0.6 21  10  0.12
0.6 0.5 91  6   0.08
0.2 0.3 53  7   0.07
0.4 0.4 54  7   0.06
0.4 0.4 49  4   0.1
0.2 0.4 71  8   0.07]

#Matriz de dados para o grupo 2 (Preparo do solo B)
X2=[0.6 0.3 127 10  0.08
0.6 0.3 87  7   0.06
0.6 0.3 69  20  0.07
```

```
0.2 0.4 58  8   0.05
0.6 0.3 42  8   0.04
0.6 0.4 32  7   0.04
0.2 0.3 37  8   0.08
0.8 0.5 52  20  0.07
0.9 0.4 87  5   0.07
0.6 0.5 44  8   0.06
0.4 0.4 49  8   0.06
0.4 0.7 61  15  0.12
0.2 0.3 78  7   0.07
0.2 0.4 53  11  0.07
1.1 0.5 75  8   0.09
0.9 0.4 73  5   0.07
0.6 0.5 99  6   0.07
0.4 0.4 78  9   0.06
0.4 0.3 40  6   0.1
0.9 0.3 80  5   0.06
0.8 0.3 76  13  0.06
0.9 0.3 87  10  0.08
0.6 0.3 44  10  0.07
0.8 0.3 171 10  0.06
0.9 0.5 40  9   0.08
0.6 0.4 57  9   0.09
0.9 0.4 104 7   0.07
0.8 0.4 71  7   0.05
0.8 0.3 66  6   0.05
1.2 0.4 118 6   0.08
0.4 0.4 41  8   0.13
0.9 0.5 107 12  0.6
0.1 0.4 41  7   0.1
0.4 0.6 95  20  0.12
0.2 0.8 79  13  0.07
0.4 0.5 59  8   0.09
0.4 0.4 82  6   0.05]

#Tamanho da amostra do grupo 1 (n1) e número de variáveis (p)
(n1,p)=size(X1)


#Tamanho da amostra do grupo 2 (n2) e número de variáveis (p)
(n2,p)=size(X2)

#Vetor de médias para o grupo 1
MEDIAX1=mean(X1,dims=1)

#Vetor de médias para o grupo 2
MEDIAX2=mean(X2,dims=1)

#Matriz variância-covariância para o grupo 1
S1=cov(X1)
```

```
#Matriz variância-covariância para o grupo 2
S2=cov(X2)

#Matriz variância-covariância ponderada
SP=((1/n1)*S1)+((1/n2)*S2)

#Cálculo da estatística T2 de Hotelling
T2=(MEDIAX1-MEDIAX2)*inv(SP)*(MEDIAX1-MEDIAX2)'

#Cálculo do valor de Qui-quadrado crítico
QUI2 = quantile(Chisq(p), 0.95)
```

4 ANÁLISE DE COMPONENTES PRINCIPAIS

Considere o problema de transformar um conjunto de variáveis correlacionadas em um conjunto de variáveis não correlacionadas, cujas variáveis transformadas são combinações lineares das variáveis originais.

4.1 TRANSFORMAÇÕES LINEARES

A partir de uma matriz X de dados $n\ x\ m$, multiplique-a por um vetor coluna T contendo m elementos. Assim, tem-se a seguinte forma:

$$Y_{(n,1)} = X_{(n,m)} \cdot T_{(m,1)} \tag{4.1}$$

O resultado é uma nova variável, ou seja, uma combinação das variáveis originais. Dessa forma, o vetor coluna T define a transformação das variáveis originais. T pode também ser uma matriz composta de várias colunas, sendo que, neste caso, cada coluna da matriz T definirá uma nova variável, que é uma combinação linear das variáveis originais. Essas novas variáveis podem ou não ser correlacionadas entre si. As novas variáveis formadas são chamadas de componentes principais.

Uma transformação linear que preserve o número original de variáveis pode também ser considerada como uma rotação de eixos. Os novos eixos serão mutuamente perpendiculares, desde que a matriz de transformação seja ortogonal, por exemplo, $T'T = I$.

4.2 RELAÇÃO ENTRE MÉDIAS DAS VARIÁVEIS ORIGINAIS E TRANSFORMADAS

O vetor médio para a matriz original X é:

$$\left[\frac{\sum_{i=1}^{n} X_{i1}}{n}, \frac{\sum_{i=1}^{n} X_{i2}}{n}, \dots, \frac{\sum_{i=1}^{n} X_{im}}{n}\right] \tag{4.2}$$

Sendo que a *inésima* linha da matriz transformada é:

$$(T_{11} \cdot X_{i1} + T_{21} \cdot X_{i2} + \cdots + T_{m1} \cdot X_{im}) \dots (T_{1m} \cdot X_{i1} + T_{2m} \cdot X_{i2} + \cdots + T_{mm} \cdot X_{im}) \tag{4.3}$$

As médias das variáveis transformadas são:

$$\frac{\sum_{i=1}^{n}(T_{11} \cdot X_{i1} + T_{21} \cdot X_{i2} + \cdots + T_{m1} \cdot X_{im})}{n} \dots \frac{\sum_{i=1}^{n}(T_{1m} \cdot X_{i1} + T_{2m} \cdot X_{i2} + \cdots + T_{mm} \cdot X_{im})}{n} = \tag{4.4}$$

$$\frac{(T_{11}\sum_{i=1}^{n} X_{i1} + T_{21}\sum_{i=1}^{n} X_{i2} + \cdots + T_{m1}\sum_{i=1}^{n} X_{im})}{n} \dots \frac{(T_{1m}\sum_{i=1}^{n} X_{i1} + T_{2m}\sum_{i=1}^{n} X_{i2} + \cdots + T_{mm}\sum_{i=1}^{n} X_{im})}{n}$$

$$= (T_{11}\bar{X}_1 + T_{21}\bar{X}_2 + \cdots + T_{m1}\bar{X}_m), \dots, (T_{1m}\bar{X}_1 + T_{2m}\bar{X}_2 + \cdots + T_{mm}\bar{X}_m)$$

Em outras palavras, uma transformação linear das variáveis produz novas variáveis, cujas médias são as mesmas que as originais transformadas. Uma consequência disso é que as variáveis transformadas terão média zero, desde que as variáveis originais tenham também médias zero.

4.3 EFEITO DO DESLOCAMENTO DA MÉDIA NAS VARIÁVEIS ORIGINAIS SOBRE AS VARIÁVEIS TRANSFORMADAS

O deslocamento da média nas variáveis originais X resulta da soma da matriz

$$X + \Delta \tag{4.5}$$

Em que cada linha da matriz Δ consiste no vetor

$$[\Delta_1, \Delta_2, \dots, \Delta_i\ ] \tag{4.6}$$

com Δ_j sendo a quantidade de deslocamento da média para a j-ésima variável X.

Então, a *inésima* linha da matriz de dados, após a média ser deslocada é

$$[(X_{i1} + \Delta_1), (X_{i2} + \Delta_2), \dots, (X_{im} + \Delta_m)\ ] \tag{4.7}$$

e o vetor de médias será:

$$\left[\frac{\sum_{i=1}^{n} X_{i1}+\Delta_1}{n}, \frac{\sum_{i=1}^{n} X_{i2}+\Delta_2}{n}, \ldots, \frac{\sum_{i=1}^{n} X_{im}+\Delta_m}{n}\right] = \quad (4.8)$$
$$[(\bar{X}_1+\Delta_1), (\bar{X}_2+\Delta_2), \ldots, (\bar{X}_m+\Delta_m)]$$

Aplicando a transformação T para a matriz de dados, após o deslocamento da média, tem-se a matriz transformada, onde a *inésima* linha é:

$$[\{T_{11}(X_{i1}+\Delta_1)+T_{21}(X_{i2}+\Delta_2)+\cdots + T_{m1}(X_{im}+\Delta_m)\}, \ldots, \{T_{1m}(X_{i1}+\Delta_1)+T_{2m}(X_{i2}+\Delta_2) + \cdots + T_{mm}(X_{im}+\Delta_m)\}] \quad (4.9)$$

com o vetor médio sendo

$$\left[\frac{\sum_{i=1}^{n}\{T_{11}(X_{i1}+\Delta_1)+T_{21}(X_{i2}+\Delta_2)+\cdots+T_{m1}(X_{im}+\Delta_m)\}}{n}, \ldots, \frac{\sum_{i=1}^{n}\{T_{1m}(X_{i1}+\Delta_1)+T_{2m}(X_{i2}+\Delta_2)+\cdots+T_{mm}(X_{im}+\Delta_m)\}}{n}, \ldots\right] \quad (4.10)$$

$$= \left[\left\{\left(T_{11}\frac{\sum_{i=1}^{n}X_{i1}}{n}+T_{11}\Delta_1\right)+\left(T_{21}\frac{\sum_{i=1}^{n}X_{i2}}{n}+T_{21}\Delta_2\right)+\cdots+\left(T_{m1}\frac{\sum_{i=1}^{n}X_{im}}{n}+T_{m1}\Delta_m\right)\right\}, \ldots, \left\{\left(T_{1m}\frac{\sum_{i=1}^{n}X_{i1}}{n}+T_{1m}\Delta_1\right)+\left(T_{2m}\frac{\sum_{i=1}^{n}X_{i2}}{n}+T_{2m}\Delta_2\right)+\cdots+\left(T_{mm}\frac{\sum_{i=1}^{n}X_{im}}{n}+T_{mm}\Delta_m\right)\right\}\right]$$

Então, o deslocamento da média na variável original também produz um deslocamento da média na variável transformada, cujo deslocamento para a *j-ésima* variável transformada é dado por:

$$\left(T_{1j}\Delta_j + T_{2j}\Delta_2 + \cdots + T_{mj}\Delta_m\right) \quad (4.11)$$

Uma vez que o deslocamento da média não afeta a matriz variância-covariância ou a matriz correlação, o deslocamento da média na variável original não muda a matriz variância-covariância da variável transformada.

4.4 RELAÇÃO ENTRE A VARIÂNCIA DAS VARIÁVEIS ORIGINAIS E AS VARIÁVEIS TRANSFORMADAS

Assumindo o vetor médio com o valor zero para simplificar, e sabendo que as médias não afetam as relações entre variâncias, covariâncias e correlações, os resultados serão gerais.

$$Var\ X_j = \frac{\sum_{i=1}^{n} X_{ij}^2}{n-1}$$

$$Var\ Y_j = \frac{\sum_{i=1}^{n}\left(T_{1j}X_{i1}+T_{2j}X_{i2}+\cdots+T_{mj}X_{im}\right)^2}{n-1}$$

$$Var\ Y_j = \sum_{k=1}^{m} T_{kj}^2\ \frac{\sum_{i=1}^{n} X_{ij}}{n-1} + 2\sum_{k=1}^{m}\sum_{z=k+1}^{m} T_{kj}T_{zj}\ \frac{\sum_{i=1}^{n} X_{ik}X_{iz}}{n-1} \quad (4.12)$$

$$Var\ Y_j = \sum_{k=1}^{m} T_{kj}^2\ Var\ X_j + 2\sum_{k=1}^{m}\sum_{z=k+1}^{m} T_{kj}T_{zj}Cov\ X_kX_z$$

4.5 EFEITO EM DIVIDIR CADA VARIÁVEL ORIGINAL (X_j) POR UMA CONSTANTE Δ_j

Antes de se proceder a divisão, tem-se:

$$Y_j = T_{ij}X_{ij} + T_{2j}X_{i2} + \cdots + T_{mj}X_{im} \quad (4.13)$$

e após a divisão resulta em:

$$Y_j = T_{1j}\frac{X_{i1}}{\Delta_1} + T_{2j}\frac{X_{i2}}{\Delta_2} + \cdots + T_{mj}\frac{X_{im}}{\Delta_m}$$
$$Y_j = \frac{T_{1j}}{\Delta_1}X_{i1} + \frac{T_{2j}}{\Delta_2}X_{i2} + \cdots + \frac{T_{mj}}{\Delta_m}X_{im} \quad (4.14)$$

Dessa forma, a divisão de X_{ij} por Δ_j tem o efeito de dividir a j-ésima linha da matriz de transformação por Δ_j, produzindo uma nova transformação:

$$Q = \Delta_{Diag} \cdot T \quad (4.15)$$

Em que Δ_{Diag}=constante diagonal.

4.6 COVARIÂNCIA DAS VARIÁVEIS ORIGINAIS E TRANSFORMADAS

Novamente, assumindo o vetor médio como sendo zero tem-se:

$$\begin{aligned} Cov\ X_jY_k &= \frac{\sum^n X_{ij}Y_{ik}}{n-1} \\ &= \frac{\sum^n X_{ij}\left(T_{1k}X_{i1} + T_{2k}X_{i2} + \cdots + T_{mk}X_{im}\right)}{n-1} \\ &= \frac{\sum^n \left(T_{1k}X_{ij}X_{i1} + T_{2k}X_{ij}X_{i2} + \cdots + T_{mk}X_{ij}X_{im}\right)}{n-1} \\ &= \frac{T_{1k}\sum X_{ij}X_{i1} + T_{2k}\sum X_{ij}X_{i2} + \ldots + T_{mk}\sum X_{ij}X_{im}}{n-1} \\ &= T_{1k}\frac{\sum X_{ij}X_{i1}}{n-1} + T_{2k}\frac{\sum X_{ij}X_{i2}}{n-1} + \cdots + T_{mk}\frac{\sum X_{ij}X_{im}}{n-1} \\ &= T_{1k}CovX_1X_j + T_{2k}CovX_2X_j + \cdots + T_{mk}CovX_mX_j \end{aligned} \quad (4.16)$$

$$CovX_jX_k = T_{1k}CovX_1X_j + T_{2k}CovX_2X_j + \cdots + T_{jk}VarX_j + \cdots + T_{mk}CovX_mX_j$$

A matriz total das covariâncias é expressa por:

$$C_XT \quad (4.17)$$

4.7 TRANSFORMAÇÕES DE VARIÁVEIS CORRELACIONADAS PARA NÃO CORRELACIONADAS

A matriz de variância-covariância das variáveis X é:

$$S_X = \frac{X'X - \frac{(UX)'(UX)}{n}}{n-1} \quad (4.18)$$

e a matriz de correlação R é:

$$R_X = D_XS_XD_X \quad (4.19)$$

Em que: D é a matriz diagonal da matriz inversa de desvios padrões.

As variáveis serão não correlacionados se os elementos fora da diagonal de R forem zero. Então, o problema é encontrar uma matriz de transformação $T_{(m\cdot m)}$, tal que:

$$Y = XT \quad (4.20)$$

sendo S_Y e R_Y matrizes diagonais.

Para simplificar, assume-se que todas as variáveis X tenham sido ajustadas para média zero, sendo que:

$$S_X = \frac{X'X}{n-1} \quad (4.21)$$

e

$$S_Y = \frac{T'X'XT}{n-1} \quad (4.22)$$

Se os elementos fora da diagonal no numerador da matriz variância-covariância forem zero, as variáveis não são correlacionadas. Então, é necessário T tal que $T'X'XT$ seja diagonal.

Importante observar que as matrizes variância-covariância e correlação são ambas quadradas e simétricas.

A partir da obtenção dos valores e vetores característicos (*eignvalue/eignvector*), deseja-se encontrar um conjunto de escalares λ_i e vetores Z_i, tal que:

$$G_{Z_i} = \lambda_i Z_i \tag{4.23}$$

Em que: G é uma matriz quadrada.

Os escalares λ_i são denominados de valores característicos (*eignvalues*) e os vetores Z_i são denominados vetores característicos (*eignvectors*) da matriz G. Como os vetores característicos têm um multiplicador escalar arbitrário, é necessário que os vetores característicos sejam normalizados.

Os vetores característicos normalizados podem ser escritos como colunas da matriz modal.

$$M = [Z_1, Z_2, \dots, Z_m] \tag{4.24}$$

Então,

$$\begin{gathered} GM = G[Z_1, Z_2, \dots, Z_m] \\ GM = [GZ_1, GZ_2, \dots, GZ_m] \end{gathered} \tag{4.25}$$

E, desde que

$$G_{Z_i} = \lambda_i Z_i \tag{4.26}$$

tem-se que

$$GM = [\lambda_1 Z_1, \lambda_2 Z_2, \dots, \lambda_m Z_m] \tag{4.27}$$

A matriz diagonal dos valores característicos E é denominada de matriz espectral.

Assim,

$$ME = [Z_1, Z_2, \dots, Z_m] \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix} \tag{4.28}$$

$$ME = [\lambda_1 Z_1, \lambda_2 Z_2, \dots, \lambda_m Z_m] = GM$$

Então,

$$GM = ME$$

E, assim

$$M^{-1}GM = M^{-1}ME = E \tag{4.29}$$

Em outras palavras, $M^{-1}GM$ é a matriz diagonal, cujos elementos são os valores característicos.

Para uma matriz simétrica, a matriz modal constituída pelos vetores característicos normalizados também é ortogonal, e assim:

$$M^{-1} = M' \tag{4.30}$$

sendo que

$$M^{-1}GM = M'GM \tag{4.31}$$

Fazendo,

$$G = X'X \text{ e } T = M \tag{4.32}$$

tem-se que

$$T'X'XT \tag{4.33}$$

que é a matriz diagonal. Esta matriz é também o numerador da matriz variância-covariância para o conjunto de variáveis transformadas $Y = XT$. As novas variáveis (Y) são não correlacionadas.

Complementando, quando se adiciona ou se subtrai uma constante de uma variável isto não modificará a variância ou covariância, consequentemente, a simplificação em assumir a média zero não tem nenhum efeito na regra geral.

A segunda interpretação da solução dos componentes principais (e a que lhe confere o nome) é a da rotação de eixos, a qual maximiza a variância ao longo de um primeiro eixo e maximiza as variâncias remanescentes ao longo do segundo eixo perpendicular ao primeiro e, assim, sucessivamente.

Considere vários pontos definidos em duas dimensões com a origem na média. Na Figura 4.1 está ilustrada essa situação.

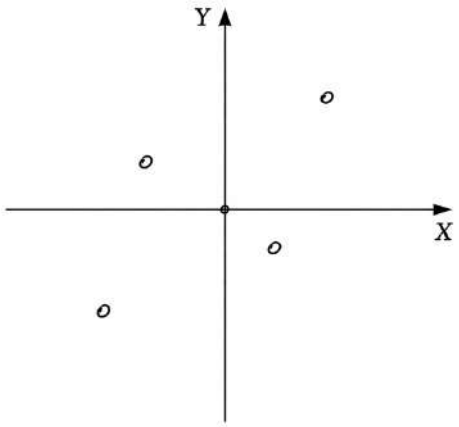


Figura 4.1: Representação de pontos em um eixo ortogonal bidimensional.

Cada ponto pode ser considerado como um vetor, conforme a Figura 4.2.

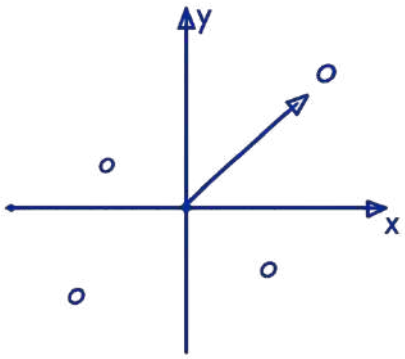


Figura 4.2: Representação de cada ponto considerado como um vetor.

Considere um segundo conjunto de eixos (pontilhados), fazendo ângulos α e β com respeito aos eixos originais. Então, $\cos\alpha$ e $\cos\beta$ são as direções dos cossenos para os novos eixos, conforme Figura 4.3.

Nota: A propriedade geométrica das direções dos cossenos é que a soma dos quadrados das direções dos cossenos é igual a 1.

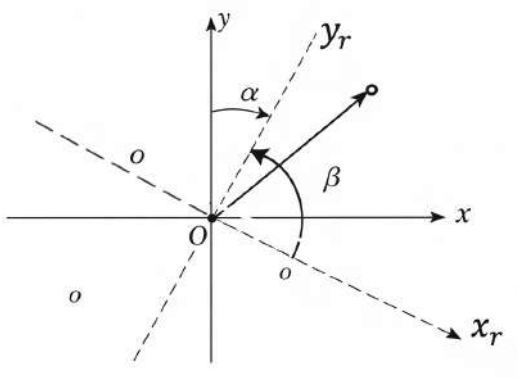


Figura 4.3: Direções dos cossenos α e β para os novos eixos Y_u e X_u.

O comprimento da projeção do vetor sobre o eixo - $Y(Y_u)$ rotacionado é $Y\cos\alpha + X\cos\beta$. A geometria dessa relação está demonstrada na Figura 4.4.

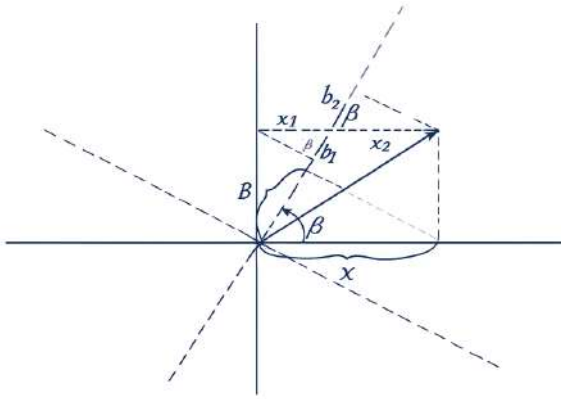


Figura 4.4: Geometria da relação $Y\cos\alpha + X\cos\beta$.

Desenvolvendo as relações na Figura 4.4 tem-se:

$$\frac{B}{X} = \cos\beta \therefore B = X\cos\beta \tag{4.33}$$

$$\frac{b_1}{x_1} = \cos\beta \therefore b_1 = x_1 cos\beta \tag{4.34}$$

$$\frac{b_2}{x_2} = \cos\beta \therefore b_2 = x_2 cos\beta \tag{4.35}$$

então,

$$\begin{gathered} b_1 + b_2 = x_1\cos\beta + x_2\cos\beta \\ b_1 + b_2 = (x_1 + x_2)\cos\beta \\ b_1 + b_2 = X\cos\beta \\ b_1 + b_2 = B \end{gathered} \tag{4.36}$$

Generalizando o que foi demonstrado para m dimensões, faz-se γ_j ser um vetor coluna das direções dos cossesnos para os eixos rotacionados j^{th} com respeito aos eixos originais. Então, o comprimento da projeção de um vetor linha x_i, observado nesse eixo rotacionado é:

$$x_i\gamma_j \tag{4.37}$$

e a variância dessas projeções sobre todos os dados (pontos) é dada por:

$$(X\gamma_j)'(X\gamma_j) \text{ ou } V_r = \frac{\gamma_j' X' X \gamma_j}{(n-1)} \tag{4.38}$$

Em que X é matriz de dados completa, e $\gamma'\gamma = I$.

Para maximizar V_r concernente à direção dos cossenos, e condicionando-se que $\mathfrak{l}'\mathfrak{l} = I$, o multiplicador de Lagrange é empregado e maximiza-se,

$$\mathfrak{l}_j' X'X \mathfrak{l}_j - \lambda(\mathfrak{l}_j'\mathfrak{l}_j - 1) \tag{4.39}$$

Fazendo com que $W = X'X$, isso se expande para:

$$\mathfrak{l}_j' W \mathfrak{l}_j - \lambda\left(\mathfrak{l}_j'\mathfrak{l}_j - 1\right) = \sum_{i=1}^{m} w_{ii}\lambda_{ij}^2 + 2\sum_{i=1}^{m}\sum_{k>i}^{m} W_{ik}\mathfrak{l}_{ij}\mathfrak{l}_{kj} - \lambda\left(\sum_{i=1}^{m}\mathfrak{l}_{ij}^2 - 1\right) \tag{4.40}$$

Tomando-se as derivadas parciais relativas aos elementos de $\mathfrak{l}_j$, tem-se:

$$\frac{\partial\left[\mathfrak{l}_j' W \mathfrak{l}_j - \lambda\left(\mathfrak{l}_j'\mathfrak{l}_j - 1\right)\right]}{\partial \mathfrak{l}_{ij}} = 2W_{ii}\mathfrak{l}_{ij} + 2\sum_{k\neq i}^{m} W_{ik}\mathfrak{l}_{kj} - 2\lambda\mathfrak{l}_{ij} \tag{4.41}$$

Igualando-se o resultado a zero, são eliminados os dois multiplicadores e obtém-se, para o conjunto de m parciais,

$$W\mathfrak{l}_j - \lambda\mathfrak{l}_j = 0 \tag{4.42}$$

Fatorando-se o vetor $\mathfrak{l}_j$, tem-se

$$(W - \lambda I)\mathfrak{l}_j = 0 \tag{4.43}$$

Soluções consistentes existem se e somente se:

$$|W - \lambda I|\mathfrak{l}_j = 0 \tag{4.44}$$

$|W - \lambda I|$ é um polinômio em λ cujas raízes são os valores característicos (*eigenvalues*) de W.

Substituindo cada um dos valores característicos de volta em:

$$(W - \lambda I)\mathfrak{l}_i = 0 \tag{4.45}$$

produz os vetores característicos de W como as possíveis soluções para os vetores das direções dos cossenos $\mathfrak{l}_j$. A condição $\mathfrak{l}_j'\mathfrak{l}_j = I$ exige vetores característicos normalizados.

Substituindo os vetores característicos de volta em:

$$V_r = \frac{\mathfrak{l}' X'X \mathfrak{l}}{(n-1)} \tag{4.46}$$

obtém-se que o vetor com o maior valor característico é aquele que fornece o valor máximo de V_r.

Para examinar as relações entre os valores e os vetores característicos de $W = X'X$, faz-se L uma matriz coluna dos vetores característicos, e Λ ser a matriz diagonal dos valores característicos. Então,

$$WL = L\Lambda \tag{4.47}$$

Se λ_1 e λ_2 são duas soluções com os correspondentes vetores $\mathfrak{l}_1$e $\mathfrak{l}_2$, então

$$\begin{aligned} W\mathfrak{l}_1 &= \lambda_1\mathfrak{l}_1 \\ W\mathfrak{l}_2 &= \lambda_2\mathfrak{l}_2 \end{aligned} \tag{4.48}$$

e

$$\mathfrak{l}_2' W \mathfrak{l}_1 = \lambda_1 \mathfrak{l}_2'\mathfrak{l}_1 \tag{4.49}$$

$$\mathfrak{l}_1' W \mathfrak{l}_2 = \lambda_2 \mathfrak{l}_1'\mathfrak{l}_2 \tag{4.50}$$

Contudo, $W = X'X$ é simétrica, sendo que,

$$\mathfrak{l}_2' W \mathfrak{l}_1 = \mathfrak{l}_1' W \mathfrak{l}_2 \tag{4.51}$$

Consequentemente,

$$\mathfrak{l}_2'\mathfrak{l}_1 = \mathfrak{l}_1'\mathfrak{l}_2 \tag{4.52}$$

Também, se $\lambda_1 \neq \lambda_2$ é necessário que

$$\mathfrak{l}_1'\mathfrak{l}_2 = 0 \tag{4.53}$$

Isso é verdadeiro para quaisquer dois dos vetores $\mathfrak{l}$, sendo que

$$L'L = I \tag{4.53}$$

Em que: L é ortogonal, sendo que $WL = L\Lambda$ e $L'L = I$.

Outra notação que poderia ser aplicada é:

$$C_{XY} = C_X T$$
$$R_{XY} = \Delta_{\frac{1}{S_X}} C_{XY} T \Delta_{\frac{1}{S_Y}} \tag{4.54}$$

Se a matriz de correlação for usada, então $X = Z$ e $\Delta_{\frac{1}{S_Z}} = I$, e

$$R_{ZY} = C_{ZY} T \Delta_{\frac{1}{S_Y}}$$
$$R_{ZY} = L\Lambda\Delta_{\frac{1}{\sqrt{\lambda}}} \tag{4.55}$$
$$R_{ZY} = L\Delta_{\sqrt{\lambda}}, \text{ pois } L\Lambda = WL$$
$$L'WL = \Lambda \text{ e } W = L\Lambda L'$$

Desde que os outros vetores dos cossenos direcionais são ortogonais para o primeiro eixo, e a forma do problema é a mesma para os eixos remanescentes, o conjunto dos vetores característicos definem os eixos principais.

4.8 COMPONENTES PRINCIPAIS COMO O MELHOR AJUSTE $k-$ DIMENSIONAL EM m DIMENSÕES

No processo para maximizar a variância ao longo dos eixos principais, também foram minimizados os resíduos sobre os eixos remanescentes. Assim, componentes principais também podem ser considerados como um ajuste de mínimos quadrados nas primeiras $k-$ dimensões, com os resíduos sendo medidos perpendicularmente para os eixos rotacionados (Figura 4.5).

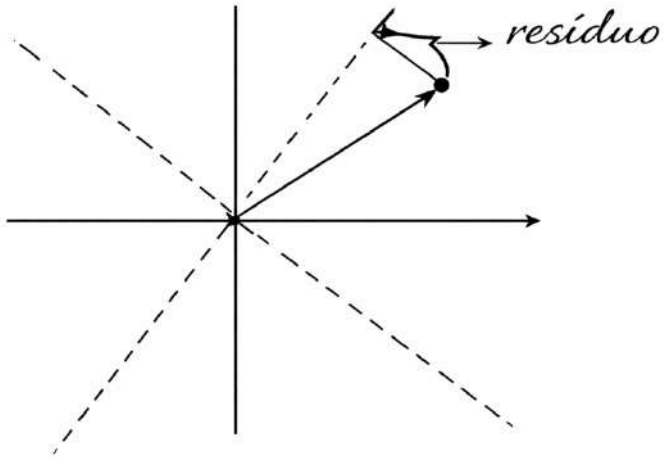


Figura 4.5: Resíduos medidos perpendicularmente para o eixo rotacionado.

4.9 USOS E INTERPRETAÇÃO DA ANÁLISE DE COMPONENTES PRINCIPAIS

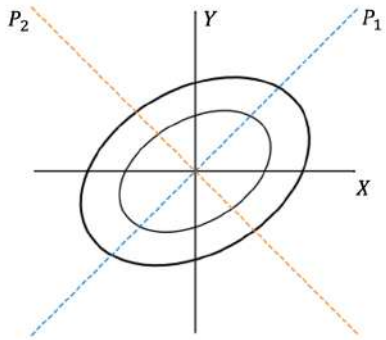
4.9.1 Introdução

A discussão da análise de componentes principais é feita por duas razões. A primeira, servirá como modelo básico, com o qual o modelo de fator comum pode ser comparado e contrastado. A segunda, introduzirá os conceitos, um tanto estranhos, como valores característicos (raízes características, *eingvalues*) e vetores característicos (*eingvectors)*, de maneira mais fácil e seus envolvimentos nos algoritmos analíticos de fatores.

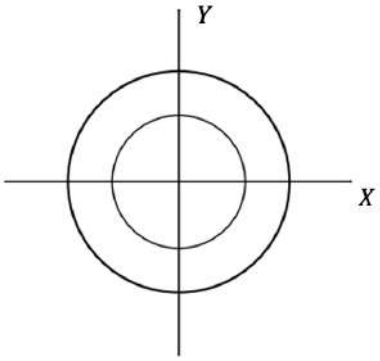
Análise de componentes principais é um método para transformar um conjunto de variáveis observadas em um outro conjunto de variáveis. A maneira mais fácil para ilustrar suas características baseia-se na sua lógica, ou seja, seu exame dentro do contexto

bivariado, assumindo que existem duas variáveis, X e Y, e que suas distribuições são bivariadas normais.

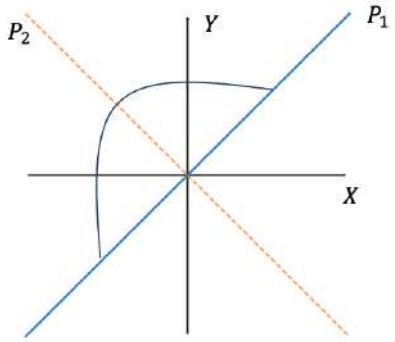
Uma relação normal bivariada com uma associação positiva moderada é ilustrada pela Figura 4.6, por meio de um *grid* bivariado.



a) Alguma correlação entre X e Y



b) Nenhuma correlação entre X e Y



c) Correlação perfeita entre X e Y

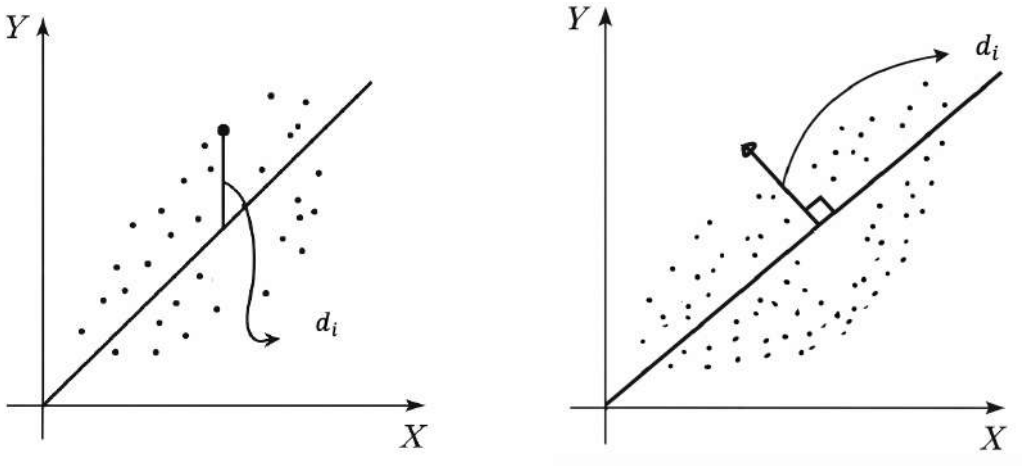
Figura 4.6: Alguns exemplos de eixos principais para distribuições bivariadas.

Esses gráficos mostram que, devido a relação positiva entre X e Y, o conglomerado de pontos (dados) que tem os maiores valores para X apresentam a tendência de se associarem com os maiores valores de Y (e vice-versa) e, então, mais casos são agrupados ao longo do primeiro e terceiro quadrantes do que ao longo do segundo e quarto quadrantes. Esses gráficos formam elipses, os dois eixos representados pelas linhas hachuradas. O eixo principal (P_1) estende-se ao longo da linha aonde a maioria dos pontos está localizada; o segundo eixo (P_2) estende-se ao longo da linha aonde a menor quantidade de pontos está localizada.

Para representar a posição relativa de cada caso em termos de uma única dimensão ou eixo, a escolha lógica seria o eixo referência P_1, pois nesse sentido essa linha é mais próxima para os dados (pontos) como um todo. A primeira componente principal não é mais a representação de casos ao longo do eixo principal. Como um exemplo, um caso com valor 1 em ambos X e Y será representado por um valor maior do que 1 em P_1 e um valor menor em P_2. Se cada caso for descrito tanto para P_1 e P_2 (em termos de pontos num sistema de coordenadas), nenhuma informação é perdida e é possível descrever cada caso exatamente independentemente da relação entre X e Y. Contudo, o primeiro eixo (a primeira componente) é mais informativo em descrever casos assim que a associação entre X e Y fique mais forte. No caso extremo quando X e Y são funções lineares de cada

um, a primeira componente conterá toda a informação necessária para descrever cada caso. Se X e Y são independentes, não haverá eixo principal e o uso da análise de componentes principais não propiciará qualquer vantagem.

Apesar da ilustração do eixo principal ser componente de uma elipse e considerar uma distribuição normal bivariada, o conceito de eixos principais não é restrito para relações que sejam normais. Em geral, o eixo principal é dado por uma linha, na qual a soma das distâncias ao quadrado de cada ponto é um valor mínimo. A comparação com o princípio dos mínimos quadrados ajuda a explicar isso. Quando se ajusta uma linha de regressão por mínimos quadrados $\left(\hat{Y} = a + bX\right)$, minimiza-se a soma das distâncias ao quadrado entre Y e $\hat{Y}$ (em que $Y =$ observado e $\hat{Y} =$ estimado), ou seja, minimiza-se $\left(Y - \hat{Y}\right)^2$, onde a distância é medida pela linha paralela ao eixo Y e perpendicular ao eixo X. No cálculo do eixo principal, minimiza-se a distância perpendicular entre o ponto (dado) e o eixo (a distância é do ponto perpendicular ao eixo principal, não para X). Essa diferença está demonstrada pela Figura 4.7.



a) Mínimos quadrados ($\sum d_i^2$ é minimizada)

b) Eixo principal ($\sum d_i^2$ é minimizada)

Figura 4.7: Comparação entre a linha de regressão por mínimos quadrados e o eixo principal dos componentes principais.

Assim que a primeira componente principal é definida, de tal maneira que a maior parte da informação está contida nela (ela explica a maior parte da variância), a segunda componente é definida da mesma maneira com a condição de que o seu eixo seja perpendicular ao primeiro. No caso bivariado, entretanto, assim que o primeiro componente é especificado, o segundo é determinado automaticamente. Também deve-se considerar que, a não ser que Y seja uma função linear de X e vice-versa, existirão dois componentes principais (são necessários dois eixos para descrever completamente a distribuição conjunta).

Na obtenção dos componentes principais, não é necessário assumir a existência de fatores hipotéticos. Os novos eixos são funções matemáticas das variáveis observadas. Mesmo quando a análise de componentes principais é usada como uma maneira de otimizar a representação (esse é o caso de se examinar somente os primeiros componentes), o objetivo não é explicar as correlações entre as variáveis, mas em se obter o máximo de variância possível dos dados. Por outro lado, a decomposição analítica do fator exigirá somente um fator (nesse caso bivariado), e o objetivo primário é verificar a

correlação entre as variáveis. Assim, o primeiro é orientado em explicar a variância e o segundo a covariância.

Quando existem mais do que duas variáveis, a base para a definição dos componentes principais é a mesma. A ferramenta matemática primária de onde tal decomposição hierárquica ou transformação surge é então denominada de equação característica ou equação *eigen*. A solução dessa equação origina os valores e os vetores característicos associados com a matriz.

4.9.2 Propriedades básicas

A análise de componentes principais é um procedimento analítico para transformar um conjunto de variáveis em um outro de variáveis componentes que tem as seguintes propriedades:

- são funções lineares das variáveis originais;
- são ortogonais, isto é, independentes entre si;
- a variação total entre elas é igual a variação total existente nas variáveis originais. Desta forma, a variação existente entre diferentes variáveis observadas não é perdida pela transformação;
- a variância associada com cada componente decresce em ordem, a primeira variável representará a maior proporção da variação total, a segunda representará a maior proporção remanescente e assim por diante.

Com a compreensão dessas propriedades, a comparação entre a Análise de Componentes Principais (ACP) com a Análise de Fatores é apropriada. Existe uma confusão, conflitante, com relação a esta terminologia. Por exemplo, onde o termo "análise de fatores" algumas vezes é aplicado para todos os procedimentos multivariados que tratam da redução da dimensionalidade e identificação de fatores comuns, ACP é geralmente apresentada com uma técnica dentro da "análise de fatores". Em outros casos, como nas rotinas científicas do programa STATISTICA, SPSS da IBM, ACP é rotulada como "análise de fatores de componentes principais".

Duas importantes distinções existem entre ACP e análise de fatores (para detalhes consultar kendall, 1957; Pearce e Holland, 1960; Cattel, 1965 e Pearce, 1965):

1 - Na análise de fatores, as p variáveis originais são reduzidas em $m < p$ fatores não correlacionados que têm um componente residual não correlacionado; na ACP as p variáveis correlacionadas são transformadas em p variáveis não correlacionadas, sendo que nem todas necessariamente são significantes.

2 - Ao contrário da ACP, na análise de fatores é possível rotacionar os eixos ortogonais que representam "fatores" para novas posições oblíquas, sendo que postulações teóricas inerentes no modelo podem ser testadas.

Uma hipótese básica para a ACP é que a variação observada é causada pelos efeitos que os fatores casuais detém em cada uma das variáveis originais. ACP, então, é um modelo fechado, sem considerar o erro aleatório ou variação externa do sistema. Toda a variação nas variáveis originais é absorvida pelas variáveis derivadas. Na análise de fatores, contudo, somente uma porção da variação total é absorvida pelas $m < p$ variáveis transformadas (essa porção é denominada de "comunalidades") e a variância remanescente é considerada um erro de variância.

Apesar de o modelo de análise de fatores parecer melhor para aplicações biológicas, a necessidade em estimar as comunalidades impõe um problema, pois exige o conhecimento a priori do sistema. Estimativas iniciais das comunalidades geralmente são um pouco melhores do que tentativas arbitrárias; assim, uma série de iterações são necessárias antes que o pesquisador se considere satisfeito. Como resultado, o modelo é

desenvolvido para que se ajuste aos dados (Kendall, 1957). Na ACP, contudo, o processo é reverso: se trabalha a partir dos dados para um modelo hipotético. Iniciando-se com as observações, o pesquisador desenvolve um modelo que reduz as dimensões de variação, o que, consequentemente, auxilia nas interpretações dos dados.

Deve ser enfatizado que as variáveis, componentes principais, derivadas usando ACP podem não ter nenhum significado. Técnicas multivariadas não devem ser consideradas como uma maneira para a geração automática de hipóteses e, sim, como um passo inicial em que conjunto de dados complexos sejam simplificados, objetivando ser mais fácil de interpretar. Qualquer hipótese desenvolvida usando-se ACP, que seja plausível, só pode ser considerada subjetiva até que seja confirmada da existência de conhecimento apropriado ou estudos adicionais (Pearce, 1965).

4.9.3 Terminologia e representação

Não é necessário que o usuário entenda todos os aspectos da derivação da ACP. O que é preciso, contudo, é ter uma visão da sua terminologia e representação, se desejar efetivamente usar ACP.

Por exemplo, suponha que $x_1, x_2, x_3, \dots, x_p$ são variáveis aleatórias e que $\underline{x}$ é um vetor linha composto de valores das variáveis aleatórias $x's$. Desta população, uma amostra de m observações independentes, pode ser retirada, sendo que:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

em que X é uma matriz de dados $n \cdot p$ de ordem completa (por exemplo, linhas e colunas independentes). A matriz variância-covariância de X é definida como S e matriz de correlação de X como R.

As componentes principais são definidas como combinações lineares das varáveis originais x_i em consideração e simbolizadas por $\mathfrak{z}$, $i = 1 \dots p$. A $\mathfrak{z}$ tem a forma

$$\mathfrak{z} = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ij}x_p \qquad (4.56)$$

$\underline{A}$ é definida como um vetor coluna e é relacionada como um "vetor característico" (ou vetor latente), tendo como coeficientes a_{ij}. O subscrito i do coeficiente refere-se ao número do vetor característico e o subscrito j refere-se ao número da variável original (x_i, sendo $i = 1 \dots p$).

Cada vetor característico tem uma variância associada denominada valor característico e representado por λ_i, $i = 1 \dots p$.

Geometricamente, tem-se uma dispersão de dados de n pontos em p dimensões, e a ACP é uma rotação de eixos, tal que os pontos no primeiro eixo são o máximo (o primeiro componente principal). O segundo eixo (segunda componente principal) é ortogonal ao primeiro e absorve o máximo possível da variância remanescente. Cada eixo adicional é também ortogonal e absorve a porção máxima da variação remanescente conforme está apresentado na Figura 4.7.

Sendo os eixos ortogonais entre si, cada um representa uma dimensão no espaço. Graficamente pode-se representar até 3 eixos (3 dimensões) e, além disso, somente matematicamente.

Com o advento do sensoriamento remoto e satélites que produzem n bandas espectrais, então, nesse caso é possível se visualizar tantas dimensões quantas forem o número de bandas. Para maiores detalhes ver Schovengerdt (1997).

As combinações lineares $\mathfrak{z}_i$ são os comprimentos das projeções sobre os novos eixos e projeção dos cossenos direcionais são os coeficientes a_{ij} dos vetores característicos. A variância da projeção é o valor característico λ_i (Figura 4.3).

As propriedades da ACP são importantes para a sua apropriada interpretação. Como especificado anteriormente, a variância de cada componente principal ($\mathfrak{Z}_i$) é o seu valor característico (λ_i). Além disso, os valores de λ_i somam a variância total no experimento. Dessa forma,

$$\lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_p = s_1^2 + s_2^2 + s_3^2 + \cdots + s_p^2 \qquad (4.57)$$

Isso significa que a variação total das variáveis derivadas é igual a variação total das variáveis observadas, sendo que nenhuma informação é perdida pela transformação linear.

Em complemento, tem-se a quantidade

$$\left(\frac{\lambda_i}{\sum \lambda_i}\right) 100 \qquad (4.58)$$

Em que: $\sum \lambda_i$ corresponde ao traço de S, soma dos elementos da diagonal da matriz de correlação ou da matriz de variância-covariância. A expressão (4.58) fornece a porcentagem da variância total explicada pelo $inésimo$ componente principal (Tabela 01).

4.9.4 Sequência operacional

As operações algébricas da Análise de Componentes Principais são importantes, mas elas representam somente um aspecto da análise. O espectro total da sequência operacional é a seguir descrita:

1 - Seleção das variáveis originais;
2 - Se necessário, transformação das variáveis originais para atender as condições - ε's distribuídos normalmente e independentes com média 0 (zero) e variância σ^2;
3 - Cálculo das matrizes variância-covariância ou matriz de correlação;
4 - Determinação dos valores (raízes características) e dos vetores característicos da matriz variância-covariância ou da matriz de correlação; e
5 - Interpretação dos componentes derivados.

O primeiro passo, seleção de variáveis, é extremamente importante. Estas variáveis devem ser caracteres quantitativos e, preferencialmente, serem medidas em uma escala contínua, apesar de que muitas variáveis discretas (ex., números de pinhões de uma pinha) adequadamente se aproximam de variáveis contínuas (Jeffers, 1964).

O segundo passo exige determinar se os dados deverão ser transformados, o que é uma decisão subjetiva. Na maioria das análises estatísticas, a hipótese de normalidade é geralmente negligenciada. Testes de significância somente têm significado para dados que são multivariados-normais em suas distribuições e transformações podem ser necessárias se a normalidade não está detectada. Além disto, testes podem ser usados para testar a homogeneidade da matriz variância-covariância. Contudo, Jeffers (1964) recomenda o uso de transformações somente quando os dados violam severamente as hipóteses, pois as transformações tornam, eventualmente, a interpretação da ACP mais difícil.

O terceiro passo exige uma outra decisão: Qual tipo de matriz será utilizada, se variância-covariância ou a matriz de correlação? Normalmente, se todas as unidades estiverem na mesma escala (ex.: todas são unidades de comprimento), o uso da matriz variância-covariância é recomendado. O uso da matriz variância-covariância tem um grande apelo estatístico, pois a teoria de amostragem é menos complexa do que outras (Anderson, 1958). Contudo, se as unidades são diferentes (ex.: diâmetro, volume, peso), a padronização é necessária e a matriz correlação pode ser usada. Os valores característicos (variância) associados com um vetor característico de uma matriz correlação é a variância padronizada.

O quarto passo envolve a transformação linear das p variáveis originais em variáveis artificiais, também chamadas variáveis latentes. Isto é, a equivalência matemática em se determinar os vetores e valores característicos relacionados da matriz variância-covariância ou matriz de correlação.

O quinto passo envolve a interpretação das componentes principais derivadas. Primeiro, é necessário decidir o número de componentes que tem significância biológica, econômica, ou de qualquer área que esteja sendo estudada. Existem vários critérios para ajudar nesta decisão. Em geral, a eliminação daqueles vetores característicos que não atendem aos critérios pode ser feita por convicção, admitindo-se que alguma subjetividade pode estar envolvida neste processo, mas isto é inerente em todas as decisões estatísticas. Por exemplo, qual seria o número apropriado de componentes a ser extraída para a análise obtida na Tabela 4.1?

Tabela 4.1: Valores característicos e porcentagem cumulativa da variação associada com os valores característicos da análise de componentes principais de medições em um viveiro de mudas de *Pinus taeda* com 4 anos.

Valores característicos	Percentagem cumulativa da variação
10,07	0,534
2,25	0,653
1,63	0,739
1,08	0,797
1,05	0,852
0,56	0,882
0,5	0,908
0,49	0,934
0,31	0,951
0,29	0,966
0,23	0,978
0,18	0,988
0,09	0,993
0,06	0,996
0,03	0,997
0,03	0,999
0,01	0,999
0,01	1,000

O próximo passo será efetuar a análise dos vetores característicos, que é extremamente relevante, porém é necessário atenção, dado que, mesmo depois desta sequência operacional, ainda permanece a questão se as interpretações biológicas, econômicas e outras envolvidas podem ser derivadas de artifícios matemáticos. Para interpretar variáveis derivadas, é necessário ter a capacidade de relacioná-las com as variáveis observadas.

4.9.5 Obtenção dos valores e vetores característicos de uma matriz

Em muitas das aplicações importantes na Engenharia Florestal e em outras áreas de estudo, a equação característica e suas raízes associadas precisam ser consideradas.

Dada uma matriz A de ordem n, o problema é a determinação dos escalares λ e os seus correspondentes vetores e diferentes de zero, e que simultaneamente satisfaçam a equação:

$$Ae = \lambda e \tag{4.59}$$

Isto é conhecido como valores ou problemas característicos. A questão pode ser posta assim: Existe um número escalar, que quando multiplicado pelo vetor e, produz o vetor idêntico resultado do produto Ae?

Subtraindo-se λe em ambos os lados da equação (4.59), obtém-se:

$$Ae - \lambda e = 0 \tag{4.60}$$

Assim, colocando-se *e* em evidência na equação (4.60), por fatoração, obtém-se a equação homogênea

$$(A - \lambda I)e = 0 \tag{4.61}$$

em que I é uma matriz identidade.

A equação homogênea (4.61) possui uma solução não trivial se e somente se

$$det(A - \lambda I) = 0 \tag{4.62}$$

A equação polinomial em λ, resultado de

$$det(A - \lambda I) \tag{4.63}$$

é denominado de polinômio característico.

Exemplificando o que foi acima desenvolvido, um exemplo será mostrado a seguir.

Exemplo 4.1

Dada a matriz

$$A = \begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix}$$

obtém-se o polinômio característico fazendo-se

$$det(A - \lambda I) = det\left(\begin{bmatrix} 3 & 2 \\ 4 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \begin{bmatrix} (3-\lambda) & 2 \\ 4 & (1-\lambda) \end{bmatrix} = (3-\lambda)(1-\lambda) - 8$$
$$= \lambda^2 - 4\lambda - 5$$

Exemplo 4.2

Determinar o polinômio característico para a matriz:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$
$$det(A - \lambda I) = \lambda^2 - 2\lambda - 3$$

tal que

$$det(A - \lambda I) = \begin{vmatrix} (1-\lambda) & 2 \\ 2 & (1-\lambda) \end{vmatrix} = (1-\lambda)^2 - 4 = \lambda^2 - 2\lambda - 3$$

As raízes, ou fatores, da equação característica são chamadas de raízes características ou valores característicos (*eingvalues*). No caso, as raízes características são:

$$\lambda_1 = -1, \lambda_2 = 3$$

sendo $(\lambda + 1)(\lambda - 3) = \lambda^2 - 2\lambda - 3 = 0$

Exemplo 4.3

Encontrar a equação característica e as raízes características da matriz

$$A = \begin{bmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

$$det(A - \lambda I) = \begin{vmatrix} (1-\lambda) & 1 & -2 \\ -1 & (2-\lambda) & 1 \\ 0 & 1 & (-1-\lambda) \end{vmatrix}$$

$$= (1-\lambda)[(2-\lambda)(-1-\lambda) - 1] + (-1)(-1)^3[(-1-\lambda) + 2]$$

$$= (1-\lambda)(\lambda^2 - \lambda - 3) + (-\lambda + 1)$$
$$= -\lambda^3 + 2\lambda^2 + \lambda - 2 = 0$$

Fatorando $det(A - \lambda I) = 0$, tem-se

$$\lambda_1 = -1, \lambda_2 = 2, \lambda_3 = 1$$

e

$$-(\lambda + 1)(\lambda - 2)(\lambda - 1) = \lambda^3 + 2\lambda^2 + \lambda - 2 = 0$$

Em geral, pode ser demonstrado que para uma matriz quadrada de ordem n:

$$det(A - \lambda I) = \lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_1\lambda + (-1)^n|A|$$

de onde obtém-se o polinômio de ordem n^{it} e c_R é o coeficiente do polinômio R^{it}. O último termo é uma constante igual a $(-1)^n$ vezes o valor do determinante de A.

Se as raízes da equação característica não forem inteiras, o trabalho para extraí-las é muito mais difícil. Para tanto é usual a utilização de métodos numéricos, como o método de Newton-Raphson. Quando a ordem da equação característica é maior do que 4, o uso de programas computacionais para encontrar as raízes do polinômio se faz necessário.

Correspondendo a cada raiz característica, λ_i, da equação geral existe um vetor característico e_i (*eingvectors*). Este vetor característico é a solução para as equações homogêneas que são obtidas de:

$$(A - \lambda_i I)e_i = 0$$

O vetor e_i é o vetor que corresponde a raiz i^{ti}.

$$e_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Considerando a matriz

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

com as raízes características $\lambda_1 = -1$ e $\lambda_2 = 3$, como mostrado no Exemplo 02. Associado com a primeira raiz tem-se:

$$(A - \lambda_1 I)e_1 = 0$$

ou

$$\left(\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} - (-1)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

o que implica que:

$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$e_1 = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$ é o vetor característico associado com a raiz $\lambda_1 = -1$.

$e_1 = -e_2$ pois $2\ e_1 + 2\ e_2 = 0$ é obtido de cada linha da equação $(A - \lambda_1 I)e_1 = 0$.

Fazendo-se $e_1 = \beta$, uma constante arbitrária. Dessa forma, $e_1 = \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

O subscrito 1 em e_1 implica que se tem o vetor característico e associado com a primeira raiz $\lambda_1 = -1$. Importante observar que a matriz $(A - \lambda_1 I)$ possui ordem 1 e existe um valor arbitrário e um valor único em e.

A segunda raiz característica é $\lambda_2 = 3$. Então, $(A - \lambda_2 I)e_2 = 0$ e

$$\left(\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} - (3)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

ou

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

lembrando-se que

$e_2 = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$, assim $e_1 = e_2$.

$$e_2 = \gamma \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Novamente com λ_2

$$(A - \lambda_2 I) = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}$$

é de ordem 1 e possui um valor arbitrário em e_2 e um único valor determinado.

O procedimento de obtenção das raízes e dos vetores característicos, como foi demonstrado, exige o entendimento matemático e também a interpretação do que significam as raízes e os vetores característicos. Isto vai depender da estrutura dos dados e o significado das variáveis envolvidas.

Como já foi descrito no item 4.9.4, pode ser usada a matriz variância-covariância ou a matriz de correlação. Esta decisão dependerá do usuário.

Cada área de atuação é específica e exige que o usuário compreenda a estrutura dos dados perfeitamente. Caso contrário, interpretações errôneas serão produzidas com consequências desastrosas.

4.9.6 Teste de esfericidade de Bartlett

Bartlett (1950) apresentou o teste de esfericidade. Ele é, provavelmente, o exemplo mais simples baseado no procedimento de Wilks (1932), ao usar o determinante da matriz de dispersão como a analogia multivariada de variância. Wilks denominou o determinante da matriz variância-covariância como índice de variância generalizada, fazendo com que P seja a matriz de correlação da população, para variáveis padronizadas z, então $|P|$ é a variância generalizada e $|R|$ é a sua estimativa. O que Bartlett desenvolveu foi uma aproximação do teste do Qui-quadrado, que testa H_0: $|P| = 1$. É claro, $|P| = 1$ somente se $P = I$. O critério do teste é:

$$\chi^2_{0,5(p^2-p)} = -\left[(N-1) - \frac{1}{6}(2\,p + 5)\right] \log_e(|R|) \qquad (4.64)$$

Para verificar a racionalidade do critério do teste de Bartlett, a equação (4.64) será então examinada:

- O $|R|$ formado por $z'z\left(\frac{1}{N}\right)$ estará sempre na amplitude de 1,00 até 0,00. Assim que $|R| \to 1$, $R \to I$; como $R \to I$, a dispersão se aproxima da esfericidade $(0 \leq |R|1)$.

- O $\log_e$ de frações decimais (na amplitude de 0 a 1) são negativos e tem a amplitude de $-\infty$ até 0; assim que $|R| \to 1$, então o $\log_e(|R|) \to 0$ e o $\chi^2 \to 0$.

- Um N grande com respeito a p aumentará o χ^2 para um dado $|R|$; então uma razão $\frac{N}{p}$ alta, aumentará o poder do teste.

Um χ^2 não significante para um nível α razoável indica não ser apropriada a obtenção dos valores e vetores característicos da matriz, pois os vetores característicos serão tratados como um conjunto de elementos não correlacionados.

4.9.7 Um exemplo prático de ACP

Para esclarecer o uso de Análise de Componentes Principais, um exemplo da área florestal foi escolhido.

Antes de se iniciar os procedimentos, uma breve descrição dos dados será feita, para o entendimento do que é a ACP, qual é sua importância, sua aplicação e interpretação.

O conjunto de variáveis e as suas respectivas observações são as seguintes:

- Variáveis utilizadas

x_1 = Diâmetro (dap em m) da árvore

$x_2 =$ Altura (em m)
$x_3 =$ Área transversal média (m²)
$x_4 =$ Fator de forma (adimensional)

O conjunto de variáveis e observações está apresentado na Tabela 4.2. Corresponde a um conjunto de dados de pinheiros mensurados nos Estados Unidos. Foram mensurados o diâmetro à altura do peito (DAP, em m) e a altura total (H, em m). As árvores também foram cubadas por meio da mensuração dos diâmetros a cada metro ao longo do fuste. A partir desses dados, foi calculada a área transversal em cada ponto de medição e, posteriormente, obtida a sua média. Essa variável foi denominada área transversal média da árvore (G, em m²). Também foi calculado o fator de forma (FF, adimensional).

- x_1: Diâmetro da árvore, DAP em m

Uma das variáveis usadas no cálculo do volume das árvores é o seu diâmetro medido a 1,30 metros do solo. O DAP corresponde o diâmetro à altura do peito.

- x_2: Altura em m

Esta altura refere-se a altura total e é obtida por meio de instrumentos específicos.

- x_3: Área transversal média, G em m²

A área transversal é obtida por meio da seguinte fórmula:

$$g = \frac{\pi d^2}{4}$$

Os dados no presente exercício referem-se a área transversal média de uma árvore, ou seja, calcula-se a área transversal em cada posição de medição da cubagem e, posteriormente, obtém-se a sua média. Essa variável foi denominada área transversal média da árvore (G, em m²).

- x_4: fator de forma, FF adimensional

O fator de forma é o fator de afilamento da árvore e é calculado em relação a um cilindro perfeito. Assim, o fator de forma é calculado pela seguinte fórmula:

$$FF = \frac{volume\ da\ árvore}{volume\ de\ um\ sólido\ geométrico\ com\ base\ no\ diâmetro\ e\ altura}$$

sendo sua fórmula resumida dada por:

$$FF = \frac{v}{g\ h}$$

Em que: v é o volume da árvore (m³); g é a área transversal na posição do DAP (m²); h é a altura (m)

Para maiores detalhes das variáveis descritas, consultar Machado e Figueiredo Filho (2003).

Havendo este entendimento de que estas variáveis são responsáveis pela obtenção do volume individual da árvore, é possível então por meio da ACP estudar a estrutura destas variáveis e sua importância. Importante frisar que a explicação das variáveis a serem utilizadas na ACP é uma tentativa em demonstrar o quanto o entendimento das variáveis é importante.

4.9.7.1 *Introdução*

Como já discutido, o objetivo da ACP é transformar um conjunto de dados que sejam correlacionados em um outro conjunto não correlacionado (ortogonais entre si). A ideia é, tendo-se organizado ou coletado um conjunto de dados, observar o padrão de variação entre as variáveis.

Desta forma, a variável que tem a maior variação (componente principal) terá também a maior variância, a segunda variável tem a maior variância associada com a

primeira, a terceira variável tem a maior variância associada com a primeira e segunda variáveis, e assim por diante.

Tabela 4.2: Conjunto de variáveis e observações de Pinheiros.

	Variável			
Observação	DAP	H	G	FF
1	1,4062	28,5000	1,6648	0,3475
2	1,1735	32,3336	1,1590	0,3285
3	1,3444	30,4000	1,6537	0,3598
4	1,2646	37,6472	1,3128	0,3215
5	1,2894	34,1056	1,4112	0,3526
6	0,9857	31,0856	0,7885	0,3319
7	1,0793	34,6664	0,9682	0,3370
8	1,3833	46,4000	1,6144	0,3303
9	1,4999	52,0000	2,1217	0,3387
10	1,3110	40,0000	1,4922	0,3381
11	1,3420	40,8424	1,6707	0,3221
12	1,3690	36,0000	1,5474	0,3374
13	1,7499	40,0000	1,6000	0,3267
14	1,5999	33,6000	1,0000	0,3402
15	1,0833	32,6664	2,4000	0,3261
16	1,7272	37,8184	2,2000	0,3339
17	1,2000	35,2000	1,0000	0,3224
18	1,3939	39,2728	2,2000	0,3142
19	1,8333	50,6664	1,2000	0,3021
20	1,5166	44,0000	2,0000	0,3195
21	1,9523	45,7144	1,4000	0,3234
22	1,4333	40,0000	1,0000	0,3141
23	1,3333	34,6664	0,6000	0,3337
24	1,5909	42,9088	2,2000	0,3216
25	1,3541	41,5000	1,6089	0,3374

4.9.7.2 *Procedimentos de cálculo*

- Média para a matriz de dados

1	2	3	4
1,40865	38,47978	1,51254	0,33043

- Desvio padrão para a matriz de dados

1	2	3	4
0,23633	6,12347	0,48468	0,01269

- Matriz variância-covariância da matriz de dados

1	2	3	4
0,05585	0,85205	0,02654	-0,00083
0,85205	37,49685	0,84174	-0,04019
0,02654	0,84174	0,23492	-0,00014
-0,00083	-0,04019	-0,00014	0,00016

$\sum diagonal = 37{,}78778$

- Matriz de correlação da matriz de dados

1	2	3	4
1,00000	0,58878	0,23168	-0,27761
0,58878	1,00000	0,28361	-0,51734
0,23168	0,28361	1,00000	-0,02311
-0,27761	-0,51734	-0,02311	1,00000

$\sum diagonal = 4$

-Valores característicos (*Eingvalues*) da matriz variância-covariância

1	2	3	4
37,53529	0,21621	0,03617	0,00012

-Vetores característicos (*Eingvector*) da matriz variância-covariância (*Modal matrix* – matriz modal)

1	2	3	4
0,02274	0,04054	0,99892	-0,00152
0,99949	-0,02348	-0,02180	0,00118
0,02257	0,99890	-0,04106	-0,00347
-0,00107	0,00355	0,00140	0,99999

-Valores característicos (*Eingvalues*) da matriz de correlação

1	2	3	4
2,03745	0,98556	0,65001	0,32685

-Vetores característicos (*Eingvector*) da matriz correlação (*Modal matrix* – matriz modal)

1	2	3	4
0,54746	0,08679	0,69353	-0,46018
0,62575	-0,06817	0,02979	0,77647
0,30369	0,82749	-0,44611	-0,15497
-0,46530	0,55052	0,56489	0,40164

-Matrizes de carregamento (*loadings*)

1	2	3	4
0,78144	0,08616	0,55921	-0,26309
0,89319	-0,06768	0,02402	0,44391
0,43348	0,82150	-0,35971	-0,08860
-0,66416	0,54653	0,45548	0,22962

4.9.7.3 *Discussão dos resultados*

O objetivo da ACP é transformar um conjunto de dados correlacionados em outro conjunto não correlacionado em que os eixos são ortogonais entre si. A ideia básica disso é, após estar de posse de um conjunto de dados, verificar o padrão de variação entre as variáveis.

Assim, a variável que representa a maior variação (primeira componente principal) tem a maior variância, a segunda variável tem a maior variância associada com a primeira, a terceira variável tem a maior variância associada com a primeira e com a segunda variável, e assim por diante.

4.9.7.4 *Interpretação do exemplo*

O conjunto de dados utilizado e já descrito anteriormente foi:

$x_1 =$ Diâmetro (DAP, em m) da árvore
$x_2 =$ Altura (H, em m)
$x_3 =$ Área transversal média (G, m²)
$x_4 =$ Fator de forma (FF, adimensional)

O primeiro passo realizado foi calcular a matriz de variância-covariância e a matriz de correlação dos dados originais, resultando nas diagonais:

$$\sum diagonal \text{ variância-covariância} = 37{,}79$$
$$\sum diagonal \text{ correlação} = 4{,}00$$

O segundo passo foi obter os valores e vetores característicos das matrizes de variância-covariância e correlação dos dados originais.

O terceiro passo foi obter a matriz de dados não correlacionados derivados da matriz variância-covariância, obtida da seguinte maneira:

$$y_1 = X_1 T \tag{4.65}$$

Em que: y_1 é a matriz de dados não correlacionados (transformada); X_1 é a matriz de dados originais; T é a matriz de transformação ou matriz modal, obtida transpondo-se os vetores característicos da matriz variância-covariância da matriz de dados originais.

O quinto passo foi obter a matriz de dados não correlacionados derivada da matriz de correlação, obtida da seguinte maneira:

$$y_2 = X_z T \tag{4.66}$$

Em que: y_2 é a matriz de dados não correlacionados; X_z é a matriz de dados padronizada; T é a matriz de transformação ou matriz modal, obtida transpondo-se os vetores característicos da matriz correlação dos dados originais.

O sexto passo foi verificar se estes dois conjuntos de dados eram realmente não correlacionados. Assim, o seguinte procedimento foi efetuado:

- obter a variância-covariância e a correlação de y_1 e, assim, ter a confirmação que eram não correlacionados, devido:

  - Na diagonal têm-se as variâncias cujos valores são os mesmos que os valores característicos obtidos da variância-covariância dos dados originais. Fora da diagonal tem-se zero, como mostra a teoria dos componentes principais para se ter um conjunto de dados não correlacionados. Da teoria também se tem:

$$\sum_j \lambda_j = \sum_j s^2_{p_{cj}} = traço\ S_x = \sum_i S_i^2$$

- da matriz de correlação, se tem 1 na diagonal principal e zero fora da diagonal.

O sétimo passo é o mesmo que o sexto, mas relacionado para y_2 e, novamente, se tem a confirmação que o conjunto de dados derivados da matriz de correlação não são correlacionados.

O oitavo passo foi encontrar a matriz de carregamentos (*loading matrix*), na qual é possível observar a correlação entre os valores originais e os valores transformados. Isto foi feito para o conjunto de dados derivados da matriz variância-covariância e da matriz de correlação.

A matriz de correlação obtida do conjunto de dados derivados da variância-covariância apresenta um problema, pois não é independente da escala das variáveis medidas. Devido a isto, a matriz de carregamento (*loading matrix*, correlação entre os valores originais e os valores transformados) obtidos da matriz de correlação são independentes de escala.

Até este ponto, somente foi descrito como foram obtidos os resultados. Para um melhor esclarecimento, um fluxograma está apresentado, visando esclarecer os passos seguidos (Figura 4.8).

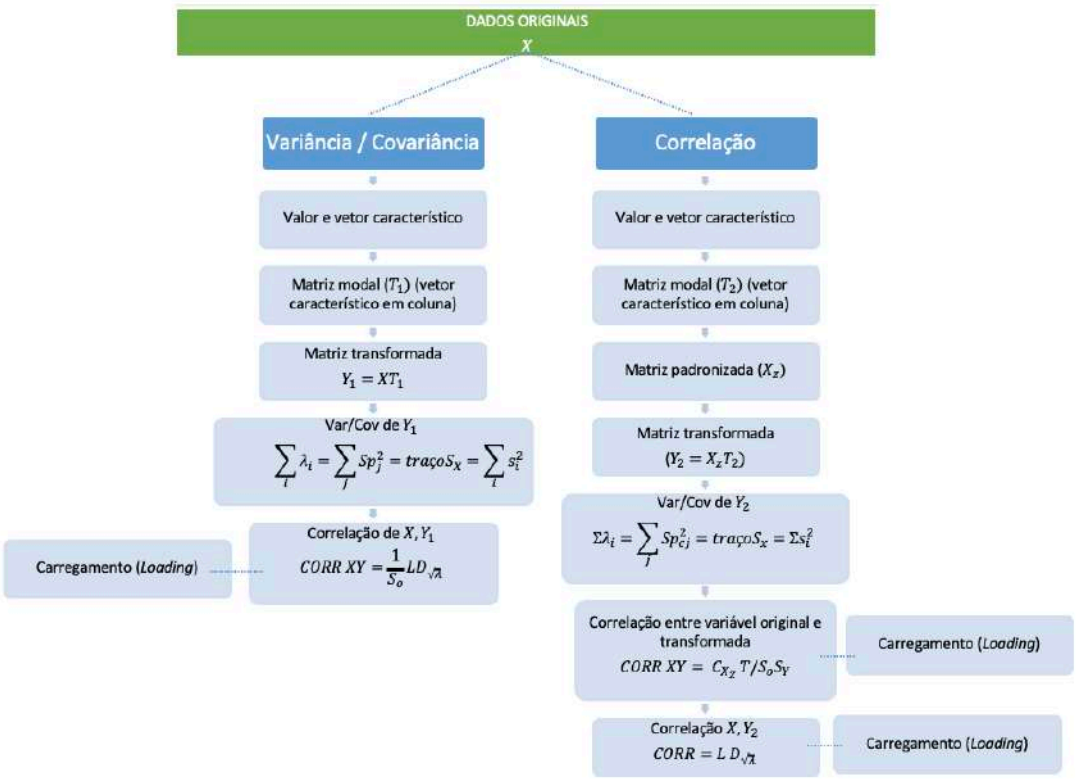


Figura 4.8: Fluxograma dos passos para obtenção dos componentes principais.

Observando-se as médias para verificar a relação entre as médias dos dados originais e os dados transformados da variância-covariância, se tem a teoria, cuja relação é a seguinte:

- as médias das variáveis transformadas são combinações lineares das médias das variáveis originais, multiplicadas pelos coeficientes dos componentes principais (vetores característicos em coluna) em ordem decrescente. O que se quer dizer em ordem decrescente é em função, como explicado nos objetivos dos componentes principais, de que a primeira componente principal absorve a maior variação e, assim, por diante.

Assim, tem-se:

$$\bar{y}_1 = \bar{x}_1 T_{11} + \bar{x}_2 T_{21} + \bar{x}_3 T_{31} + \bar{x}_4 T_{41} \qquad (4.67)$$

Em que: $\bar{y}_1$ é a média da primeira variável transformada; $\bar{x}_1, \ldots, 4$ são as médias das variáveis originais; T_{p1} são os coeficientes da primeira componente principal.

$$\bar{y}_1 = 1{,}40865 \;\; 0{,}02274 + \;\; 38{,}47978 \;\; 0{,}99949. \\ + 1{,}51254 \;\; 0{,}02257 + \;\; 0{,}33043 \;\; (-0{,}00107)$$

$$\bar{y}_1 = 38{,}52581552$$

Assim, observando-se as médias das variáveis transformadas, para a primeira componente tem-se:

$$\bar{y}_1 = 38{,}52582$$

e isto comprova a explicação anteriormente apresentada.

A Tabela 4.3 mostra as variáveis que representam a maior variação em cada componente principal (matriz de carregamento = *loadings*).

Tabela 4.3: Variáveis que contribuem com a maior variação em cada componente principal (matriz de carregamento = *loadings*).

Variável	CP1	CP2	CP3	CP4
x_1	Altura	Área transversal	Diâmetro	Fator de forma
x_2	Diâmetro	Diâmetro	Fator de forma	Altura
x_3	Área transversal	Fator de forma	Altura	Diâmetro
x_4	Fator de forma	Altura	Área transversal	Área transversal

Em que: CP é a componente principal.

Assim, verifica-se que a variável altura representa a maior variação na primeira componente principal e assim sucessivamente.

Para as outras três médias o procedimento é o mesmo.

Concernente ao conjunto derivado da matriz de correlação, as médias das variáveis transformadas serão iguais a zero, pois a matriz foi padronizada.

Observando agora as relações das componentes principais, cujos procedimentos e relações já foram esclarecidas e explicitadas as razões de as variáveis transformadas não serem correlacionadas, pode-se então fazer a seguinte decomposição:

Da teoria sabe-se que

$$\sum_j \lambda_j = \sum_j s_{p_{cj}}^2 = traço\ S_x = \sum_i S_i^2$$

Daqui para frente, somente serão considerados os componentes principais derivados da matriz de correlação devido a independência de escala. Assim,

$$\sum_j \lambda_j = 2{,}03745 + 0{,}98856 + 0{,}65015 + 0{,}32685$$

$$\sum_j \lambda_j \approx 4{,}00 = \sum_j s_{p_{cj}}^2$$

traço $S_x = 2{,}03745 + 0{,}98856 + 0{,}65015 + 0{,}32685$.

Ou seja, aproximadamente equivalente a $\sum_i S_i^2 = 4{,}00$, devido aos erros de arredondamento este valor não é exatamente 4.

Agora, analisando os componentes principais, tem-se a matriz que contém os coeficientes normalizados em colunas das componentes principais (os vetores característicos normalizados de S_x). Para a presente análise, o fator padrão será obtido conforme está apresentado na Tabela 4.4.

Tabela 4.4: Fator padrão dos vetores característicos normalizados.

Variável	CP1	CP2	CP3	CP4
x_1	0,54746	0,08679	0,69353	-0,46530
x_2	0,62575	-0,06817	0,02979	0,77647
x_3	0,30369	0,82749	-0,44611	-0,15497
x_4	-0,46530	0,55052	0,56489	0,40164
$\lambda_j = s^2_{p_{cj}}$	2,03745	0,98556	0,65015	0,32685

A Tabela 4.5 mostra a relação, isto é, a variação (variância-covariância) das variáveis originais e as transformadas derivadas da correlação, pois são independentes de escala. Assim, tem-se:

Tabela 4.5: Relação da variação das variáveis originais e transformadas derivadas da correlação.

Variável	CP1	CP2	CP3	CP4
x_1	Altura	Área transversal	Diâmetro	Altura
x_2	Diâmetro	Fator de forma	Fator de forma	Fator de forma
x_3	Área transversal	Diâmetro	Altura	Área transversal
x_4	Fator de forma	Altura	Área transversal	Diâmetro

Assim, a altura corresponde para a maior variação na primeira componente principal e assim por diante.

Agora, fazendo uma comparação entre a Tabela 4.3 e a Tabela 4.5, tem-se:

Primeiro, a Tabela 4.3 foi usada para explicar as relações sobre as médias, mas também se pode usar para explicar as variâncias derivadas da variância-covariância.

O que é importante ressaltar aqui é o problema de escala, o que realmente é claro comparando-se as duas tabelas:

Componentes principais variância-covariância	:	Componentes principais correlação
CP1	=	CP1
CP2	≠	CP2
CP3	=	CP3
CP4	≠	CP4

Assim, os componentes principais são definidos em termos das variáveis originais da seguinte maneira: (componentes escores – vetores característicos em coluna).

$$CP1 = 0{,}54746\,x_1 + 0{,}62575\,x_2 + 0{,}30686\,x_3 - 0{,}46530\,x_4$$
$$CP2 = 0{,}0868\,x_1 - 0{,}06817\,x_2 + 0{,}82749\,x_3 + 0{,}55052\,x_4$$
$$CP1 = \ 0{,}69353\,x_1 + \ 0{,}02979\,x_2 - 0{,}44611\,x_3 + 0{,}56489\,x_4$$
$$CP1 = \ -0{,}46530\,x_1 + \ 0{,}77647\,x_2 - 0{,}15497\,x_3 + 0{,}40164\,x_4$$

Porém, pela forma em que estas duas equações são apresentadas sugere-se que:

$$x_2 = 0{,}62575\,PC_1 - 0{,}06817\,PC_2 + 0{,}02979\,PC_3 + \ 0{,}77647\,PC_4$$
$$x_1 = 0{,}54746\,PC_1 - 0{,}06817\,PC_2 + 0{,}69353\,PC_3 - 0{,}46530\,PC_4$$
$$x_3 = 0{,}30686\,PC_1 + 0{,}82749\,PC_2 - 0{,}44611\,PC_3 - 0{,}15497\,PC_4$$
$$x_4 = -0{,}46598\,PC_1 + 0{,}55052\,PC_2 + 0{,}56489\,PC_3 + 0{,}40164\,PC_4$$

As variáveis foram listadas em relação ao PC_1, pois foi a altura que contribuiu para a maior variação.

Outro ponto que deve ser observado é que pelo fato de a altura contribuir para a maior variação na $CP1$, $CP2$ a altura foi a última que contribuiu para a maior variação, e assim sucessivamente para as outras variáveis observando-se a Tabela 4.5.

Observando agora para a matriz de carregamento (*loadings*), nota-se que para a correlação entre os valores originais e os valores transformados que a altura é a variável com a maior correlação, conforme demonstrado anteriormente que ela contribuiu para a maior variação. Reafirmando, não está se considerando a matriz de carregamento (*loading matrix*) derivada da matriz variância-covariância, pois é dependente da escala.

Finalmente, observando-se os valores característicos que estão dispostos na diagonal da matriz variância-covariância da matriz padronizada transformada da matriz correlação, é possível determinar as percentagens da variação total que são devidas para cada combinação linear. Assim, tem-se:

Componente principal	% da variação total	% cumulativa
PC1 = 2,03745	50,93%	50,93%
PC2 = 0,98556	24,63%	75,57%
PC3 = 0,65015	16,25%	91,82%
PC4 = 0,32685	8,17%	100%

Complementando as explicações, é importante observar o ordenamento simples das contribuições para os componentes principais, para que se comprove quais são os resultados que mostram o padrão de variação.

Na análise das variáveis padronizadas, observa-se que a primeira componente principal é um fator "geral", que absorve a variação comum entre todas as variáveis (os coeficientes são aproximadamente iguais para todas as variáveis originais). Aparentemente, desta componente depreende-se que o fator de forma decresce com o aumento do tamanho da árvore.

A segunda componente principal indica uma forte relação entre a área transversal e o fator de forma. A terceira componente principal mostra um outro aspecto importante da relação entre o fator de forma e dimensões lineares, que não são devidos à área basal, a qual é uma função quadrada. A relação da quarta componente principal é de alguma forma similar à terceira, porém muito mais difícil de interpretar e menos importante.

4.9.7.5 O teste *scree*

Os valores característicos para fatores sucessivos podem ser mostrados através de uma simples linha num gráfico. Cattell (1966) propôs que este gráfico *scree* pode ser usado para graficamente determinar o número de fatores a serem retidos, conforme apresentado na Figura 4.9.

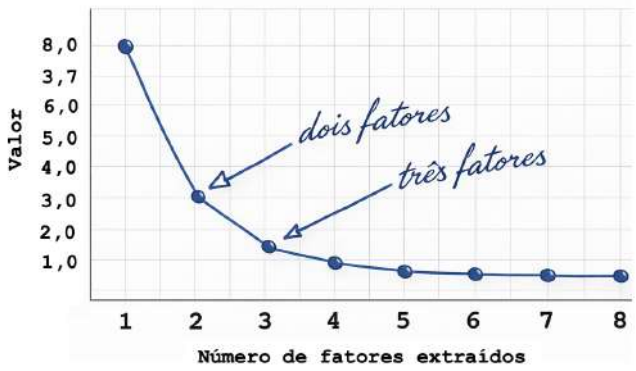


Figura 4.9: Representação gráfica da linha *scree* dos *eingvalues.*

O teste *scree* envolve encontrar o lugar onde os valores característicos começam a ficar uniformes na direção direita do gráfico. A direita deste ponto de uniformidade, provavelmente encontra-se somente "fatores *scree*" – *scree* é o termo geológico que se refere dos fragmentos que são empilhados na parte mais baixa da inclinação de uma rocha. Assim, o número de fatores a serem retidos corresponde à quantidade de fatores localizados à esquerda desse ponto de uniformidade, enquanto os fatores à direita são considerados fatores scree e, portanto, não devem ser retidos.

4.10 EXEMPLO DE ANÁLISE DE COMPONENTES PRINCIPAIS NO MATLAB

```
MATRIZ X COM DADOS ORIGINAIS
X =
    1.4062   28.5000    1.6648    0.3475
    1.1735   32.3336    1.1590    0.3285
    1.3444   30.4000    1.6537    0.3598
    1.2646   37.6472    1.3128    0.3215
    1.2894   34.1056    1.4112    0.3526
    0.9857   31.0856    0.7885    0.3319
    1.0793   34.6664    0.9682    0.3370
    1.3833   46.4000    1.6144    0.3303
    1.4999   52.0000    2.1217    0.3387
    1.3110   40.0000    1.4922    0.3381
    1.3420   40.8424    1.6707    0.3221
    1.3690   36.0000    1.5474    0.3374
    1.7499   40.0000    1.6000    0.3267
    1.5999   33.6000    1.0000    0.3402
    1.0833   32.6664    2.4000    0.3261
    1.7272   37.8184    2.2000    0.3339
    1.2000   35.2000    1.0000    0.3224
    1.3939   39.2728    2.2000    0.3142
    1.8333   50.6664    1.2000    0.3021
    1.5166   44.0000    2.0000    0.3195
    1.9523   45.7144    1.4000    0.3234
    1.4333   40.0000    1.0000    0.3141
    1.3333   34.6664    0.6000    0.3337
    1.5909   42.9088    2.2000    0.3216
    1.3541   41.5000    1.6089    0.3374


COMPONENTES PRINCIPAIS USANDO MATRIZ VARIÂNCIA/COVARIÂNCIA

VETOR DE MÉDIAS
MEDIA=mean(X)
MEDIA =
    1.4087   38.4798    1.5125    0.3304

VETOR DE DESVIOS PADRÃO
DESVIO=sqrt(var(X))
DESVIO =
    0.2363    6.1235    0.4847    0.0127

MATRIZ VARIÂNCIA/COVARIÂNCIA
S=cov(X)
```

```
S =
    0.0558    0.8521    0.0265   -0.0008
    0.8521   37.4969    0.8417   -0.0402
    0.0265    0.8417    0.2349   -0.0001
   -0.0008   -0.0402   -0.0001    0.0002
```

VARIÂNCIA TOTAL

```
TracoS=trace(S)
TracoS =
   37.7878
```

AUTOVALORES E AUTOVETORES DA MATRIZ VARIÂNCIA E COVARIÂNCIA

```
[VETOR,VALOR]=eig(S)
VETOR =

   -0.0015   -0.9989   -0.0405    0.0227
    0.0012    0.0218    0.0235    0.9995
   -0.0035    0.0411   -0.9989    0.0226
    1.0000   -0.0014   -0.0036   -0.0011

VALOR =
    0.0001         0         0         0
         0    0.0362         0         0
         0         0    0.2162         0
         0         0         0   37.5353
```

OBTENÇÃO DOS AUTOVALORES EM ORDEM CRESCENTE

```
[VALORAC , IND] = sort(diag(VALOR))
VALORAC =
    0.0001
    0.0362
    0.2162
   37.5353
IND =
     1
     2
     3
     4
```

OBTENÇÃO DOS AUTOVALORES ORDENADOS

```
VALOR = VALORAC(end:-1:1)
VALOR =
   37.5353
    0.2162
    0.0362
    0.0001
```

AUTOVETORES ORDENADOS POR AUTOVALORES

```
INDICE=IND(end : -1 : 1)
INDICE =
     4
     3
     2
     1

VETOR=VETOR(: , INDICE)
```

```
VETOR =
    0.0227   -0.0405   -0.9989   -0.0015
    0.9995    0.0235    0.0218    0.0012
    0.0226   -0.9989    0.0411   -0.0035
   -0.0011   -0.0036   -0.0014    1.0000
```

AUTOVALORES ACUMULADOS

```
ACUM=cumsum(VALOR)
ACUM =
   37.5353
   37.7515
   37.7877
   37.7878
```

VARIÂNCIA EXPLICADA ACUMULADA PELAS COMPONENTES PRINCIPAIS

```
EXPLICADA=ACUM/trace(S)
EXPLICADA =
    0.9933
    0.9990
    1.0000
    1.0000
```

CÁLCULO DOS LOADINGS

A)RAIZ QUADRADA DOS AUTOVALORES

```
RAIZVALOR=VALOR.^0.5
RAIZVALOR =
    6.1266
    0.4650
    0.1902
    0.0107
```

B)DESVIO PADRAO DA MATRIZ VARIÂNCIA/COVARIÂNCIA

```
DESVIO=diag(S).^0.5
DESVIO =
    0.2363
    6.1235
    0.4847
    0.0127
```

C)NÚMERO DE VARIÁVEIS

```
n=length(DESVIO)
n =
     4
```

D) CORRELAÇÃO DAS COMPONENTES PRINCIPAIS COM AS VARIÁVEIS ORIGINAIS - LOADINGS

```
for i=1 : n
for j=1 :n
LOADINGS(j,i)=VETOR(j,i)*RAIZVALOR(i)/DESVIO(j);
end
end
LOADINGS
LOADINGS =
    0.5895   -0.0798   -0.8038   -0.0001
    1.0000    0.0018    0.0007    0.0000
    0.2853   -0.9583    0.0161   -0.0001
   -0.5171   -0.1303   -0.0210    0.8457
```

ESCORES DAS COMPONENTES PRINCIPAIS

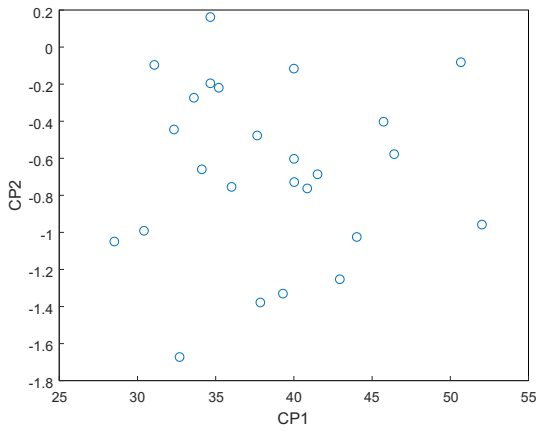
```
ESCORES=X*VETO
ESCORES =
   28.5545   -1.0521   -0.7156    0.3733
   32.3695   -0.4474   -0.4203    0.3610
   30.4519   -0.9940   -0.6129    0.3880
   37.6859   -0.4799   -0.3892    0.3596
   34.1489   -0.6625   -0.4872    0.3861
   31.1095   -0.0990   -0.2752    0.3645
   34.6946   -0.1982   -0.2832    0.3731
   46.4437   -0.5806   -0.3046    0.3775
   52.0549   -0.9606   -0.2782    0.3906
   40.0426   -0.6059   -0.3769    0.3783
   40.8893   -0.7656   -0.3822    0.3626
   36.0472   -0.7572   -0.5198    0.3726
   40.0550   -0.7313   -0.8109    0.3659
   33.6413   -0.2762   -0.8252    0.3741
   32.7281   -1.6755   -0.2720    0.3548
   37.8875   -1.3809   -0.8112    0.3684
   35.2314   -0.2223   -0.3909    0.3588
   39.3336   -1.3332   -0.4465    0.3510
   50.7088   -0.0846   -0.6781    0.3552
   44.0567   -1.0275   -0.4742    0.3624
   45.7666   -0.4056   -0.8967    0.3697
   40.0343   -0.1191   -0.5193    0.3558
   34.6921    0.1593   -0.5521    0.3706
   42.9722   -1.2559   -0.5640    0.3624
   41.5454   -0.6890   -0.3825    0.3789
```

COMPONENTES PRINCIAIS

```
CP1=ESCORES(:,1);
CP2=ESCORES(:,2);
CP3=ESCORES(:,3);
CP4=ESCORES(:,4);
```

GRÁFICO ENTRE AS DUAS PRIMEIRAS COMPONENTES PRINCIPAIS

```
scatter(CP1,CP2)
xlabel('CP1')
ylabel('CP2')
```



COMPONENTES PRINCIPAIS USANDO MATRIZ DE CORRELAÇÃO (DADOS PADRONIZADOS - Z)

TAMANHO DA MATRIZ DE DADOS

```
TAMANHO=size(X)
TAMANHO =
    25     4
LINHA=TAMANHO(1,1)
LINHA =
    25
COLUNA=TAMANHO(1,2)
COLUNA =
     4
```

VETOR DE MÉDIAS

```
MEDIA=mean(X)
MEDIA =
    1.4087   38.4798    1.5125    0.3304
```

VETOR DE DESVIOS PADRÃO

```
DESVIO = sqrt(var(X))
DESVIO =
    0.2363    6.1235    0.4847    0.0127
```

MATRIZ DE DADOS PADRONIZADOS

```
for j=1 :COLUNA
for i= 1:LINHA
Z(i,j)=(X(i,j)-MEDIA(j))/ DESVIO(j);
end
end
Z
Z =
   -0.0104   -1.6298    0.3141    1.3456
   -0.9950   -1.0037   -0.7294   -0.1520
   -0.2719   -1.3195    0.2912    2.3152
   -0.6095   -0.1360   -0.4121   -0.7037
   -0.5046   -0.7143   -0.2091    1.7476
   -1.7897   -1.2075   -1.4938    0.1160
   -1.3936   -0.6227   -1.1231    0.5180
   -0.1073    1.2934    0.2102   -0.0101
    0.3861    2.2079    1.2568    0.6520
   -0.4132    0.2483   -0.0420    0.6047
   -0.2820    0.3858    0.3263   -0.6564
   -0.1678   -0.4050    0.0719    0.5495
    1.4440    0.2483    0.1804   -0.2938
    0.8093   -0.7969   -1.0575    0.7702
   -1.3767   -0.9494    1.8310   -0.3411
    1.3479   -0.1080    1.4184    0.2737
   -0.8829   -0.5356   -1.0575   -0.6328
   -0.0624    0.1295    1.4184   -1.2791
    1.7969    1.9902   -0.6448   -2.2329
    0.4568    0.9015    1.0057   -0.8614
    2.3004    1.1815   -0.2322   -0.5540
    0.1043    0.2483   -1.0575   -1.2870
   -0.3188   -0.6227   -1.8828    0.2579
    0.7712    0.7233    1.4184   -0.6958
   -0.2308    0.4932    0.1988    0.5495
```

MATRIZ DE CORRELAÇÃO DOS DADOS ORIGINAIS

```
CORRX=corr(X)
CORRX =
    1.0000    0.5888    0.2317   -0.2776
    0.5888    1.0000    0.2836   -0.5173
    0.2317    0.2836    1.0000   -0.0231
   -0.2776   -0.5173   -0.0231    1.0000
```

MATRIZ DE CORRELAÇÃO DOS DADOS PADRONIZADOS

```
CORNORNAL=corr(Z)
CORNORNAL =
    1.0000    0.5888    0.2317   -0.2776
    0.5888    1.0000    0.2836   -0.5173
    0.2317    0.2836    1.0000   -0.0231
   -0.2776   -0.5173   -0.0231    1.0000
```

MATRIZ VARIÂNCIA E COVARIÂNCIA DOS DADOS PADRONIZADOS

```
COVS=cov(Z)
COVS =
    1.0000    0.5888    0.2317   -0.2776
    0.5888    1.0000    0.2836   -0.5173
    0.2317    0.2836    1.0000   -0.0231
   -0.2776   -0.5173   -0.0231    1.0000
```

AUTOVALORES E AUTOVETORES DA MATRIZ DE CORRELAÇÃO

```
[VETOR , VALOR ] =eig(CORRX)
VETOR =
    0.5475    0.4602    0.6935    0.0868
    0.6257   -0.7765    0.0298   -0.0682
    0.3037    0.1550   -0.4461    0.8275
   -0.4653   -0.4016    0.5649    0.5505

VALOR =
    2.0374         0         0         0
         0    0.3268         0         0
         0         0    0.6501         0
         0         0         0    0.9856
```

OBTENÇÃO DOS AUTOVALORES EM ORDEM CRESCENTE

```
[VALORAC , IND] = sort(diag(VALOR))
VALORAC =
    0.3268
    0.6501
    0.9856
    2.0374
IND =
     2
     3
     4
     1
```

OBTENÇÃO DOS AUTOVALORES ORDENADOS

```
VALORN = VALORAC(end:-1:1)
VALORN =
    2.0374
    0.9856
    0.6501
    0.3268
```

AUTOVETORES ORDENADOS POR AUTOVALORES

```
INDICE=IND(end : -1 : 1)
INDICE =
     1
     4
     3
     2
VETORN=VETOR(: , INDICE)
VETORN =
    0.5475    0.0868    0.6935    0.4602
    0.6257   -0.0682    0.0298   -0.7765
    0.3037    0.8275   -0.4461    0.1550
   -0.4653    0.5505    0.5649   -0.4016
```

AUTOVALORES ACUMULADOS

```
AVAN=cumsum(VALORN)
AVAN =
    2.0374
    3.0230
    3.6732
    4.0000
```

VARIÂNCIA EXPLICADA ACUMULADA PELAS COMPONENTES PRINCIPAIS

```
VEN=AVAN/trace(CORRX)
VEN =
    0.5094
    0.7558
    0.9183
    1.0000
```

CÁLCULO DOS LOADINGS

A)RAIZ QUADRADA DOS AUTOVALORES

```
RAIZVALOR=VALORN.^0.5
RAIZVALOR =
    1.4274
    0.9928
    0.8063
    0.5717
```

B) NÚMERO DE VARIÁVEIS

```
n=length(RAIZVALOR)
n =
     4
```

C) CORRELAÇÃO DAS COMPONENTES PRINCIPAIS COM AS VARIÁVEIS ORIGINAIS – LOADINGS

```
for i=1 : n
for j=1 :n
LOADINGSN(j,i)=VETORN(j,i)*RAIZVALOR(i);
end
end
LOADINGSN
LOADINGSN =
    0.7814    0.0862    0.5592    0.2631
    0.8932   -0.0677    0.0240   -0.4439
    0.4335    0.8215   -0.3597    0.0886
   -0.6642    0.5465    0.4555   -0.2296
```

ESCORES DAS COMPONENTES PRINCIPAIS PADRONIZADAS

```
ESCORESN=Z*VETORN
ESCORESN =
   -1.5562    1.1110    0.5643    0.7689
   -1.3236   -0.7052   -0.4804    0.2695
   -1.9633    1.5819    0.9501    0.0147
   -0.2165   -0.7721   -0.6405    0.0439
   -1.5999    0.7940    0.7093   -0.4119
   -2.2430   -1.2453   -0.5452   -0.1641
   -1.7347   -0.7227   -0.1915   -0.5399
    0.8191    0.0709   -0.1353   -1.0170
    1.6713    1.2820    0.1412   -1.6038
   -0.3650    0.2454    0.0811   -0.6323
    0.4916   -0.1421   -0.7005   -0.1152
   -0.5791    0.3751    0.1499    0.0277
    1.1374    0.0959    0.7624    0.6177
   -0.7352   -0.3265    1.4444    0.5179
   -0.6330    1.2726   -1.9926    0.5244
    0.9737    1.4487    0.4535    0.8140
   -0.8452   -1.2635   -0.5140    0.0999
    1.0728    0.4553   -1.3947    0.6043
    3.0722   -1.7425    0.3318    0.0785
    1.5204    0.3362   -0.5916    0.0120
    2.1859   -0.3780    1.4213    0.3277
    0.4902   -1.5914   -0.1756    0.2083
   -1.2560   -1.4012    0.7459   -0.0585
    1.6293    0.8083   -0.4694    0.2926
   -0.0131    0.4134    0.0763   -0.6791
CP1N=ESCORESN(:,1);
CP2N=ESCORESN(:,2);
CP3N=ESCORESN(:,3);
CP4N=ESCORESN(:,4);
```

GRÁFICO ENTRE AS DUAS PRIMEIRAS COMPONENTES PRINCIPAIS PADRONIZADAS

```
scatter(CP1N,CP2N)
xlabel('CP1')
ylabel('CP2')
```

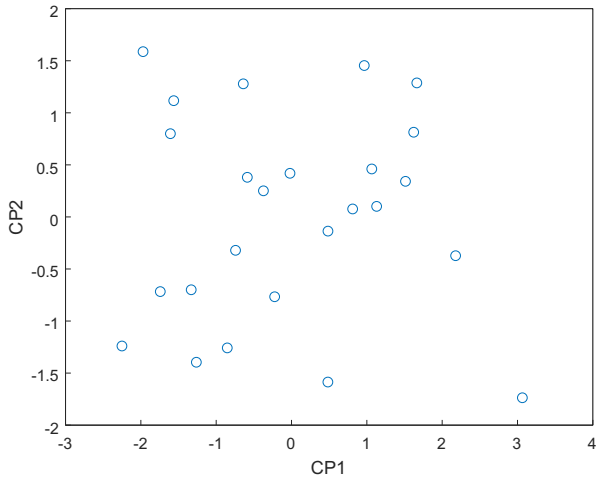
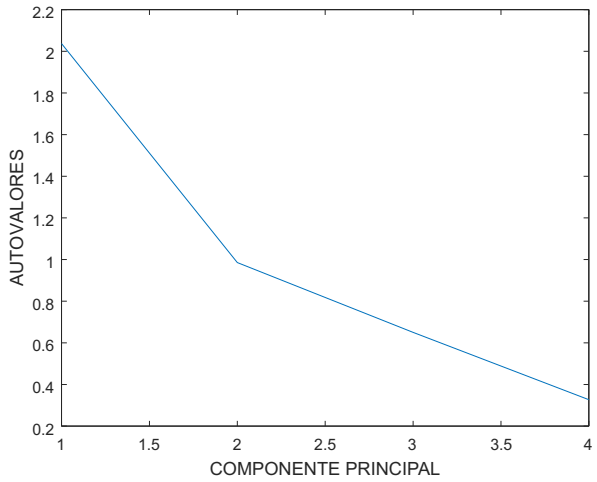


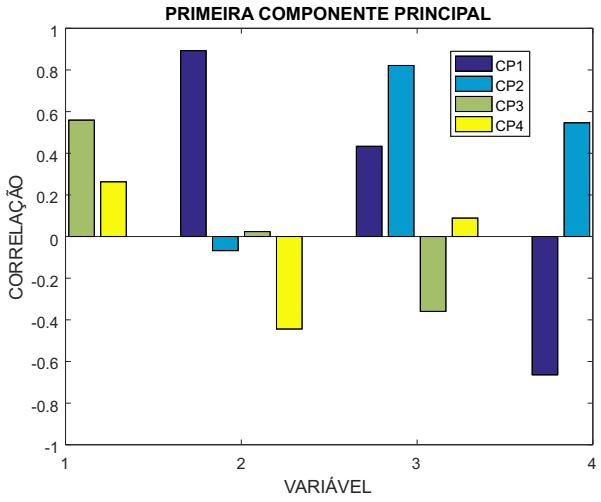
GRÁFICO DO TESTE SCREE

```
plot(VALORN)
xlabel('COMPONENTE PRINCIPAL')
ylabel('AUTOVALORES')
```



IMPORTÂNCIA DAS VARIÁVEIS NA PRIMEIRA COMPONENTE PRINCIPAL PADRONIZADA

```
bar(LOADINGSN)
axis([1 n -1 1])
xlabel('VARIÁVEL')
ylabel('CORRELAÇÃO')
title('PRIMEIRA COMPONENTE PRINCIPAL')
legend('CP1','CP2','CP3','CP4')
```



4.11 EXEMPLO DE ANÁLISE DE COMPONENTES PRINCIPAIS NO JULIA

```
using DataFrames, CSV

#Dados originais
X=CSV.read("Z:\\Alexandre\\Julia\\Multivariada\\dendrometria.csv",
DataFrame)

#Matriz X com dados originais
X=[X.d X.h X.g X.f];
```

```
#Componentes principais usando matriz variância-covariância

using Statistics, LinearAlgebra

#Vetor de médias
Média=mean(X,dims=1)

#Vetor de desvios-padrão
Desvio=std(X,dims=1)

#Matriz variância-covariância
S=cov(X)

#Variância total
TraçoS=tr(S)

#Autovalores e autovetores da matriz variância-covariância
AV = eigen(S)

#Autovalores ordenados
Valor = reverse(AV.values, dims=1)

#Autovetores ordenados por autovalores
Vetor = reverse(AV.vectors, dims=2)

#Autovalores acumulados
AVA = cumsum(Valor)

#Variância explicada acumulada pelas componentes principais
VE = AVA ./ AVA[end]

#Cálculo dos loadings
#a)Raiz quadrada dos autovalores
RaizValor=Valor.^0.5

#b)Desvio padrão da matriz variância-covariância
Desvio=diag(S).^0.5
#c)Número de variáveis
n=length(Desvio)

#d)Correlação das componentes principais com as variáveis originais -
loadings
Loadings = [Vetor[c,l]*RaizValor[l]/Desvio[c] for l in 1:n, c in 1:n]'

#Escores das componentes principais
Escores = X * Vetor

#Componentes principais
CP1=Escores[:,1];
```

```
CP2=Escores[:,2];
CP3=Escores[:,3];
CP4=Escores[:,4];

#Organização dos scores das duas primeiras componentes principais em uma
tabela
using JuliaDB

t = table((pc1 = CP1, pc2 = CP2))

#Gráfico entre as duas primeiras componentes principais
using StatsPlots

scatter(CP1,CP2)

xlabel!("CP1")
ylabel!("CP2")

#Componentes principais usando a matriz de correlação (dados padronizados
- Z)

#Matriz de dados padronizados
Z = (X .- mean(X, dims=1)) ./ std(X, dims=1)

#Matriz de correlação dos dados originais
CORRX=cor(X)

#Matriz de correlação dos dados padronizados
CORRNORMAL=cor(Z)

#Matriz variância-covariância dos dados padronizados
COVS=cov(Z)

#Autovalores e autovetores da matriz de correlação
AVN = eigen(CORRX)

#Autovalores ordenados
ValorN = reverse(AVN.values, dims=1)

#Autovetores ordenados por autovalores
VetorN = reverse(AVN.vectors, dims=2)

#Autovalores acumulados
AVAN = cumsum(ValorN)

#Variância explicada acumulada pelas componentes principais
VEN = AVAN ./ tr(CORRX)

#Cálculo dos loadings
```

```
#a)Raiz quadrada dos autovalores
RaizValor=ValorN.^0.5

#b)Número de variáveis
n=length(RaizValor)

#c)Correlação das componentes principais com as variáveis originais - loadings
LoadingsN = [VetorN[c,l]*RaizValor[l] for l in 1:n, c in 1:n]'

# Escores das componentes principais padronizadas
EscoresN = Z * VetorN

#Componentes principais
CP1N=EscoresN[:,1];
CP2N=EscoresN[:,2];
CP3N=EscoresN[:,3];
CP4N=EscoresN[:,4];

#Gráfico entre as duas primeiras componentes principais padronizadas
using StatsPlots
scatter(CP1N,CP2N)
xlabel!("CP1")
ylabel!("CP2")

#Gráfico do teste Scree
plot(ValorN, legend = false)
xlabel!("Componente Principal")
ylabel!("Autovalor")

# Importância das variáveis na primeira componente principal padronizada
plot(LoadingsN[:,1], seriestype = :bar, legend = false)
ylims!((-1,1))
xlabel!("Variável")
ylabel!("Correlação")
title!("Primeira componente principal")

# Importância das variáveis em cada componente principal padronizada
groupedbar(Loadings, bar_position = :dodge, bar_width=0.7, legend=:top)
ylims!((-1,1))
xlabel!("Variável")
ylabel!("Correlação")
```

5 ANÁLISE DE FATORES

5.1 INTRODUÇÃO

Se várias variáveis são medidas, as correlações entre os seus pares podem ser arranjadas no que é conhecida como matriz de correlação R. Os elementos da diagonal da matriz R são todos um e os elementos fora da diagonal são as correlações entre os pares de variáveis.

A existência de grupos com coeficientes de correlação grandes entre os subgrupos de variáveis sugere que essas variáveis podem estar medindo aspectos de uma mesma dimensão implícita. Essas são conhecidas como fatores ou variáveis latentes. Pela redução de um conjunto de dados de um grupo de variáveis correlacionadas em um conjunto menor de fatores, a análise de fatores consegue explicar a quantidade máxima de variância comum numa matriz de correlação usando menos variáveis (fatores). Na análise de fatores busca-se reduzir a matriz R para as suas dimensões latentes (implícitas) pelo exame de quais variáveis se agrupam de modo relevante. Esta redução é alcançada analisando as variáveis que mais se correlacionam dentro de um grupo delas, sem se correlacionar com as variáveis fora deste grupo.

5.2 ANÁLISE DE FATORES

O objetivo da análise de fatores é reduzir a redundância entre as variáveis usando menos fatores (variáveis latentes). A ênfase da análise de fatores é centrada na modelagem das covariâncias dentre as variáveis $x's$. Deseja-se expressar as $\frac{1}{2}p(p-1)$ covariâncias e as p variâncias de $x_1, x_2, \dots, x_p$ em termos de uma estrutura simplificada envolvendo a extração de carregamentos e variâncias específicas.

A análise de fatores é relacionada à análise de componentes principais, pois ambas procuram um conjunto menor de variáveis, mas elas diferem em alguns aspectos. Por exemplo, duas diferenças básicas são que os componentes principais são definidos como combinações lineares das variáveis originais, enquanto na análise de fatores as variáveis são expressas como combinações lineares dos fatores (variáveis latentes). Na análise de componentes principais a maior parte da variância total das variáveis é explicada (traço da matriz variância-covariância), enquanto na análise de fatores o objetivo é considerar as covariâncias ou correlações entre variáveis.

A análise de fatores é basicamente um procedimento aplicado a uma amostra aleatória de $x_1, x_2, \dots, x_p$ de uma população homogênea, com vetor de médias μ e matriz de variância-covariância Σ. Na análise de fatores as variáveis $x_1, x_2, \dots, x_p$ são representadas como combinações lineares de poucas variáveis aleatórias $f_1, f_2, \dots, f_m$ $(m < p)$ chamadas fatores, com um erro associado que considera a parte das variáveis que são únicas (chamadas fatores específicos). Para $x_1, x_2, \dots, x_p$ em qualquer vetor de observação y, o modelo é:

$$
\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\
&\vdots \\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p
\end{aligned}
$$

Em que: μ_i= média da i-ésima variável; F_j= j-ésimo fator comum; ε_i= i-ésimo erro ou fator específico; l_{ij}= peso ou carregamento na $i$nésima variável X_i do j-ésimo fator F_j, $i = 1,2, \dots, p$ e $j = 1,2, \dots, m$.

O modelo apresentado acima pode ser representado em forma matricial como:

$$\underline{X} - \underline{\mu} = L\underline{F} + \underline{\varepsilon}$$

$$L=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{bmatrix} \qquad \underline{F}1 = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

Em que: Vetor de $\underline{F}$ tem média zero; A variância de $\underline{F}$ corresponde a uma matriz identidade; $\underline{F}$ e $\underline{\varepsilon}$ são independentes: $cov(\underline{\varepsilon}, \underline{F}) = 0$.

De modo ideal m deve ser substancialmente menor do que p, caso contrário não é conseguido a descrição das variáveis como funções de poucos fatores intrínsecos. Os valores de F podem ser considerados como variáveis aleatórias, que dão origem aos $x's$. Os coeficientes (carregamentos) servem como pesos, mostrando como cada x_i depende do valor de F, e sua magnitude representa a importância das variáveis nos fatores.

A variância de cada variável original é composta por uma parte da variância da variável X_i que é distribuída nos m fatores comuns (h_i^2) e pela porção da variância devido ao fator específico (variância específica, Ψ_i), como segue:

$$\sigma_i^2 = h_i^2 + \Psi_i$$
$$\sigma_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \Psi_i$$

Em que: $h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2$ constitui o que se chama de comunalidade, que é a porção da variância da variável X_i (σ_i^2) que é distribuída pelos m fatores comuns; Ψ_i é a porção da variância da variável X_i devido ao fator específico, também denominada de variância específica; l_{ij} é a covariância da inésima variável X_i com o $j-$ésimo fator comum F_j.

O particionamento da variância de x_i em um componente devido aos fatores comuns é chamado de comunalidade e o componente único de x_i é chamado de variância específica. As comunalidades (h_i^2) são também referidas como variância comum, e a variância específica (Ψ_i) é chamada especificidade, variância única ou variância residual.

Se a matriz de correlação for utilizada ao invés da matriz variância-covariância, isto significa que são utilizadas variáveis padronizadas. Então, l_{ij} será o coeficiente de correlação entre a inésima variável X_i com o $j-$ésimo fator comum.

$$l_{ij} = cov(X_i, F_j)$$
$$l_{ij} = cor(X_i, F_j)$$

Uma vez que as variáveis são especificadas, pode-se, então, aplicar a análise de fatores para identificar a estrutura latente das relações entre as variáveis $x's$. Decisões devem ser tomadas concernentes ao método de extração dos fatores e quanto ao número de fatores selecionados para identificar a estrutura latente dos dados.

A primeira técnica considerada para estimar os carregamentos (l_{ij}) é comumente denominada de métodos dos componentes principais. Apesar dessa denominação que pode causar confusão entre análise de fatores e análise de componentes principais, no método de componentes principais para estimar os carregamentos não são calculados nenhum componente principal, o que a diferencia da análise de fatores.

A partir de uma amostra aleatória $x_1, x_2, \ldots, x_p$ é obtida a matriz variância-covariância S e tenta-se encontrar um estimador de l (carregamentos):

$$S \cong \hat{l}\hat{l}' + \widehat{\Psi}$$

Importante é salientar que na abordagem de componentes principais $\widehat{\Psi}$ não é considerado, tornando $S \cong \hat{l}\hat{l}'$.

A fim de fatorar S é utilizada a decomposição espectral:

$$S = CDC'$$

Em que: C é uma matriz ortogonal construída com vetores normalizados ($c_i' c_i = 1$) de S como colunas e D é a matriz com os autovalores de S na diagonal.

Para concluir a decomposição espectral para a forma $\hat{l}\hat{l}'$ deve-se observar que como os autovalores da matriz S são todos positivos ou zeros, pode-se escrever:

$$D = D^{0,5} D^{0,5}.$$

Em que:

$D^{0,5}$ é a matriz diagonal dos autovalores, $D = \begin{bmatrix} \sqrt{\hat{\lambda}_1} & & \\ & \sqrt{\hat{\lambda}_2} & \\ & & \sqrt{\hat{\lambda}_p} \end{bmatrix}$.

Com esta fatoração de D obtém-se:

$$S = CDC' = CD^{0,5} D^{0,5} C'$$

$$S = (CD^{0,5})(CD^{0,5})'$$

As variâncias na diagonal de S são modeladas exatamente, mas as covariâncias presentes fora da diagonal são apenas aproximações, e isto caracteriza o desafio da análise de fatores. Devido a este fato, é importante analisar a matriz de resíduos obtida ao final do processo.

Sintetizando, sejam considerados os pares de autovalores e autovetores de S (matriz variância-covariância amostral) $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2)$,..., $(\hat{\lambda}_p, \hat{e}_p)$, em que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$, com $m < p$ fatores comuns. O estimador da matriz dos pesos ou carregamentos dos fatores $\hat{l}_{ij}$ é dada por:

$$\hat{L} = \hat{C}\hat{D}_{\lambda}^{1/2}$$

$$\hat{C} = \begin{bmatrix} \hat{e}_{11} & \hat{e}_{12} & \cdots & \hat{e}_{1p} \\ \hat{e}_{21} & \hat{e}_{22} & \cdots & \hat{e}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{e}_{p1} & \hat{e}_{p2} & \cdots & \hat{e}_{pp} \end{bmatrix} \qquad \hat{D}_{\lambda}^{1/2} = \begin{bmatrix} \sqrt{\hat{\lambda}_1} & & \\ & \sqrt{\hat{\lambda}_2} & \\ & & \sqrt{\hat{\lambda}_p} \end{bmatrix}$$

A fonte da solução para os componentes principais tem como fonte a matriz de autovetores de S, de modo que os carregamentos nos fatores são proporcionais aos coeficientes dos componentes principais. Os fatores são então denominados os primeiros m componentes principais.

As observações do conjunto de dados $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$ devem ser centralizadas:

$$\underline{x}_j - \underline{\overline{x}} = \begin{bmatrix} x_{1j} - \overline{x}_1 \\ x_{2j} - \overline{x}_2 \\ \vdots \\ x_{pj} - \overline{x}_p \end{bmatrix}$$

Alternativamente, pode-se padronizar as observações:

$$\underline{z}_j = \begin{bmatrix} \dfrac{x_{1j} - \overline{x}_1}{s_1} \\ \dfrac{x_{2j} - \overline{x}_2}{s_2} \\ \vdots \\ \dfrac{x_{pj} - \overline{x}_p}{s_p} \end{bmatrix}$$

Nesse caso, a matriz de covariância amostral (S) corresponde à matriz de correlação (R).

As variâncias específicas estimadas são fornecidas pelos elementos diagonais da matriz $\widehat{\Psi} = S - \hat{L}\hat{L}'$. Assim:

$$\widehat{\Psi} = \begin{bmatrix} \widehat{\Psi}_1 & & & \\ & \widehat{\Psi}_2 & & \\ & & \ddots & \\ & & & \widehat{\Psi}_p \end{bmatrix} \text{ com } \widehat{\Psi}_i = s_i^2 - \sum_{j=1}^{m} l_{ij}^2$$

As comunalidades e a matriz dos resíduos são estimadas como:

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2$$

$$R - \left(\hat{L}_z\hat{L}_z' + \widehat{\Psi}_z\right) = R - \hat{L}_z\hat{L}_z' - \widehat{\Psi}_z$$

Existem vários critérios para a escolha do número de fatores. Pode-se escolher m que seja igual ao número de fatores necessários para se alcançar uma determinada porcentagem do traço da matriz de variância-covariância (S), ou da matriz de correlação (R). Outro critério é que o valor de m deve ser igual a um número de autovetores maior do que um. Também, é possível escolher m pela análise gráfica dos autovalores de S ou R. A experiência do pesquisador também pode auxiliar na escolha de m. De modo geral, a escolha de m deve ser suficientemente grande, de modo que a soma das comunalidades ou dos autovalores englobem uma porção relativamente grande do traço de S ou R. Quando o conjunto de dados é ajustado com sucesso por um modelo de análise de fatores, diferentes métodos definirão o mesmo valor de m.

Em muitas aplicações o pesquisador deseja somente verificar se uma análise de fatores se ajusta aos dados e identificar quais são esses fatores. Em outras aplicações o pesquisador deseja obter os escores fatoriais que são definidos como estimativas intrínsecas dos valores do fator para cada observação. Existem dois usos potenciais para estes escores, ou sejam, analisar o comportamento das observações em termos dos fatores de interesse, ou usar os escores como entrada para outro tipo de análise, como a MANOVA.

Os escores fatoriais são estimados para variáveis originais centradas para a média ou para variáveis padronizadas, respectivamente, como:

$$\underline{\hat{F}} = \left(\hat{L}'\widehat{\Psi}^{-1}\hat{L}\right)^{-1}L'\widehat{\Psi}^{-1}\left(\underline{x} - \underline{\overline{x}}\right)$$

$$\underline{\hat{F}} = \left(\hat{L}'\hat{L}\right)^{-1}L'\underline{z}$$

Uma dificuldade aparece quando se faz cálculo dos escores fatoriais dos eixos principais, pois sua interpretação biológica torna-se mais difícil. O que se pode fazer é transformar os escores obtidos para uma estrutura mais simples, concernente às variáveis originais. Uma vez que os fatores tenham sido extraídos, é possível calcular em que grau as variáveis estão associadas a estes fatores, calculando os carregamentos das variáveis em cada fator. Geralmente, a maior parte das variáveis tem carregamentos fortes no fator mais importante e carregamentos fracos nos outros fatores. Quando isto não acontece torna a interpretação difícil e uma técnica chamada rotação pode ser usada para a discriminação entre fatores.

Se um fator é um eixo de classificação ao longo do qual variáveis podem ser plotadas, o fator de rotação efetivamente rotaciona esses eixos de fatores de modo que as variáveis carreguem ao máximo em um único fator. Existem dois tipos de rotações que podem ser feitas, a primeira é a rotação ortogonal, na qual os fatores são rotacionados mantendo-os independentes. Antes da rotação, todos os fatores são independentes (não se correlacionam) e a rotação ortogonal assegura que os fatores permaneçam não correlacionados. A outra forma de rotação é a oblíqua, na qual os fatores podem ser correlacionados e os eixos não permanecem perpendiculares.

Na teoria, a escolha exata da rotação vai depender grandemente se os fatores latentes são ou não relacionados. Se os fatores são independentes, deve-se usar a rotação ortogonal (VARIMAX). Se, no entanto, existe base teóricas para se supor que os fatores são correlacionados, então uma rotação oblíqua direta deve ser selecionada.

A rotação conhecida como *VARIMAX* é geralmente usada para a obtenção de uma estrutura mais simples. Esta técnica iterativa é necessária para obter a rotação *varimax*, com o carregamento (*loadings*) da análise fatorial como ponto de partida. A técnica de rotação VARIMAX procura carregamentos que maximizem a variância do quadrado dos carregamentos em cada coluna dos autovetores. Se os carregamentos na coluna são aproximadamente iguais, a variância deverá ser próxima de zero. Dessa forma, o método de VARIMAX procura fazer com que os carregamentos sejam grandes ou pequenos para facilitar a interpretação.

Os novos pesos rotacionados $\hat{l}_{ij}^{*}$ são determinados pela relação:

$$\hat{L}^{*} = \hat{L}\,T$$

Em que: $\hat{L}^{*}$ são os carregamentos rotacionados; $\hat{L}$ é a matriz de carregamentos inicial; T é a matriz de rotação.

Os diversos métodos de rotação de eixos, para explicar melhor o conjunto de dados, são descritos a seguir:

Método *Varimax*

Este é um método de rotação ortogonal para minimizar o número de variáveis que possuem altos carregamentos (*loadings*) em cada fator e, isto, simplifica a sua interpretação.

Kaiser (1956, 1958, 1959) propôs uma medida da simples estrutura da soma das variâncias dos quadrados dos carregamentos dentro de cada coluna da matriz de fatores. Maiores detalhes, quanto ao desenvolvimento teórico do método de rotação VARIMAX, pode ser encontrado na literatura pertinente.

Método de rotação oblíqua direta (*Oblimin*)

É um método para rotação oblíqua (não ortogonal). Quando o parâmetro do modelo delta iguala a zero (*default* = número usado, a não ser que se deseje outro), as soluções são, na maioria das vezes, oblíquas. Assim que o delta vem a ser mais negativo,

os fatores vem a ser menos oblíquos. Para evitar o *default* delta de zero (ϕ), um número menor ou igual a 0,8 deve ser usado.

Método Quartimax

Este é um método de rotação que minimiza o número de fatores necessários para explicar cada variável. Isto simplifica a interpretação das variáveis observadas.

Método *Equamax*

Este é um método de rotação, uma combinação do método *Varimax,* o qual simplifica os fatores, e o método *Quartimax*, que simplifica as variáveis. O número de variáveis que carregam muito em um fator e o número de fatores necessários para explicar a variável são minimizados.

Método *Promax*

Este é um método de rotação oblíqua, que é computacionalmente mais rápido do que o de rotação oblíqua direta e é muito mais útil para grandes conjuntos de dados.

Para a execução da rotação é necessário calcular T, que representa os vetores característicos ordenados em coluna sem o último componente. Assim, tem-se:

$$T'A = T'TQ$$

Como $T'T$ é uma matriz identidade a rotação torna-se:

$$T'A = Q$$

Para validar se a matriz Q (a nova matriz de carregamentos) é ortogonal é necessário que:

$$Q'Q = I$$

Eliminada a última coluna da matriz de dados transformados, esta foi multiplicada pela matriz Q, para se obter a matriz H, a qual engloba o novo conjunto de dados no novo espaço, conjunto este, que facilita a interpretação em termos das variações originais.

5.3 EXEMPLO DE ANÁLISE DE FATORES

O conjunto de dados é o mesmo usado quando da análise de componentes principais, ou seja:

$x_1 =$ Diâmetro (DAP, m)
$x_2 =$ Altura (H, m)
$x_3 =$ Área transversal média (G, m²)
$x_4 =$ Fator de forma (FF, adimensional)

- Visão geral dos dados

Número de observações: 25
Número de variáveis: 4

As variáveis estão em escalas muito diferentes (ver desvios-padrão), o que já antecipa ser análise pela matriz de covariância dominada pela variável de maior variância e que a análise pela matriz de correlação é mais apropriada para interpretação estrutural.

- Análise de fatores pela matriz variância–covariância (dados originais)
- Autovalores e variância explicada

Fator	Autovalor	Variância explicada
F1	37,5353	99,33%
F2	0,2162	0,57%
F3	0,0362	0,10%
F4	0,0001	0%

Um único fator explica praticamente toda a variabilidade dos dados. Isso ocorre porque a segunda variável original (altura) tem variância alta em relação às demais (aproximadamente 37,5 de um total de 37,8).

- *Loadings* do fator selecionado

Variável	*Loading* (F1)
Var 1	0,5895
Var 2	1,0000
Var 3	0,2853
Var 4	-0,5171

O fator 1 é quase totalmente definido pela variável 2 (Altura). As demais variáveis contribuem pouco para a estrutura comum. O sinal negativo da variável 4 (fator de forma) indica relação inversa com o fator dominante. Este fator pode ser interpretado como um fator de escala, fortemente associado à variável de maior magnitude.

- Comunalidades e variâncias específicas

Variável	Comunalidade
Var 1	0,35
Var 2	1,00
Var 3	0,08
Var 4	0,27

Apenas a variável 2 (altura) é bem representada pelo fator comum. As demais têm grande parte da variância específica (não explicada). As variâncias específicas negativas indicam problemas de ajuste, comuns quando poucos fatores são usados e a matriz de covariância é dominada por uma única variável. O fator representa basicamente uma única variável, não uma estrutura latente real.

- Análise de fatores pela matriz de correlação (dados padronizados Z)

- Autovalores e variância explicada

Fator	Autovalor	Variância explicada
F1	2,0374	50,94%
F2	0,9856	24,64%
F3	0,6501	16,25%
F4	0,3268	8,17%

Dois fatores explicam 75,6% da variância total, atendendo ao critério de ≥ 80% de forma próxima.

- *Loadings* dos fatores selecionados

O Fator 1 pode ser interpretado como a relação de magnitude e está correlacionado com as variáveis 1 e 2 (diâmetro e altura), apresentando correlação negativa com a variável 4 (fator de forma). Isto representa um gradiente geral de aumento das variáveis

principais com oposição à variável 4. Este fator absorve as variações ao nível de árvore individual.

O Fator 2 pode ser interpretado como a componente estrutural secundária, dominada pelas variáveis 3 (área basal) e 4 (fator de forma). Este fator é independente da escala principal e capta a variação específica não associada ao fator dominante. Este fator absorve as variações ao nível de povoamento.

Variável	Fator 1	Fator 2
Var 1	0,7814	0,0862
Var 2	0,8932	-0,0677
Var 3	0,4335	0,8215
Var 4	-0,6642	0,5465

- Comunalidades (qualidade de representação) e variância específica

Variável	Comunalidades	Variância específica
Var 1	0,62	0,38
Var 2	0,80	0,20
Var 3	0,86	0,14
Var 4	0,74	0,26

Todas as variáveis estão bem representadas pelo modelo com dois fatores, com comunalidades altas e variância específica baixas.

- Matriz de resíduos

A matriz de resíduos revelou valores pequenos e bem distribuídos, o que indica bom ajuste do modelo fatorial. Não há correlações sistemáticas não explicadas.

- Escores fatoriais

Interpretação do gráfico das funções F1 e F2

A dispersão dos pontos mostra um gradiente claro no Fator 1, com domínio da variável 2 (altura) e uma separação adicional pelo Fator 2, causado pela influência da variável 3 (área transversal média). As observações extremas na função representam valores altos das variáveis 1 e 2 (diâmetro e altura), enquanto os valores extremos da função 2 indicam uma estrutura diferenciada da variável 3.

- Gráfico dos *loadings*

As variáveis 1 e 2 (diâmetro e altura) estão próximas, revelando alta correlação. A variável 3 (área transversal média) é praticamente ortogonal ao Fator 1, revelando uma nova dimensão. A variável 4 (fator de forma) atua como variável de contraste entre os fatores.

- Rotação dos fatores

Os dois fatores selecionados foram rotacionados em relação aos seus carregamentos, utilizando-se a rotação VARIMAX. Como resultado obteve-se coeficientes que permitem uma análise mais fácil em relação ao não rotacionados. A figura dos *loadings* rotacionados descrimina as variáveis em cada eixo fatorial de maneira

clara e sem superposição. Assim, o primeiro fator é formado pelas variáveis diâmetro, altura e fator de forma, enquanto o segundo absorve a área transversal.

- Conclusão do exemplo

Na análise com o uso da matriz variância-covariância, detectou-se um único fator dominante, devido ao forte efeito de escala e possui baixa utilidade interpretativa.

A análise com o uso da matriz de correlação, que é recomendada, revelou dois fatores latentes bem definidos, e um modelo estável e bem ajustado, permitindo uma interpretação conceitual clara.

A estrutura multivariada dos dados é adequadamente descrita por dois fatores independentes, sendo o primeiro associado à magnitude das variáveis principais ao nível de árvore individual e o segundo a uma componente estrutural específica relacionada ao nível de povoamento.

5.4 ANÁLISE DE IMAGEM

Uma outra abordagem para se tratar com as comunalidades, que não introduz a necessidade para estimativas iterativas, é chamada de análise de imagem. Este procedimento lida com comunalidades em efeito, pelo uso de regressão múltipla para estimar cada variável de todas as outras. Os componentes estimados são inseridos na análise e os resíduos são eliminados. Deve ser levado em consideração que a análise de imagem exige que o determinante da matriz correlação não seja igual a zero.

A análise de imagem é computacionalmente semelhante à análise de fatores de componentes principais, exceto que envolve um passo adicional. Este elimina os componentes únicos das variáveis e mantém as porções comuns.

O primeiro passo é padronizar a matriz de dados (Z). O passo adicional pode ser representado como:

$$C_{(n\cdot m)} = Z_{(n\cdot M)} A_{(m\cdot m)}$$

Em que: A é a matriz de ajuste para as comunalidades.

Fazendo R ser a matriz de correlação e D ser a matriz diagonal, obtida quando se substitui os elementos fora da diagonal de R^{-1} por zeros e os elementos da diagonal por seus recíprocos, obtém-se:

$$A_{(m\cdot m)} = I_{(m\cdot m)} - R^{-1}_{(m\cdot m)} D_{(m\cdot m)}$$

Deste ponto em diante, os cálculos são semelhantes aos da análise de fatores de componentes principais.

5.5 EXEMPLO PRÁTICO DA ANÁLISE DE IMAGEM

O tipo de estrutura básica do problema de análise de fatores precisa ser esclarecido considerando-se as correlações entre as variáveis $x's$. Um caminho para extração dos fatores mais importantes nesta situação é aplicar a análise de componentes principais na matriz de correlação, o que corresponde à análise de componentes principais, que implica no uso das variáveis padronizadas.

O princípio de análise de fatores, contudo, não se preocupa com a magnitude dos resíduos, mas com as relações entre as variáveis. É verdadeiro, também, que eliminando as primeiras componentes principais da correlação, usualmente reduz as correlações entre os resíduos e, neste sentido, a análise de componentes principais pode ser considerada como uma aproximação da análise de fatores.

Contudo, considerando a seguinte equação:

$$x_{ij} = \sum_{k=1}^{m} a_{ij}\, g_{ij} + e_{ij}$$

Em que: x_{ij} representa a variável j na observação i; a_{ij} são conhecidos como os fatores de carregamento de cada fator nas diferentes observações; g_{ij} é o valor da variável j e e_{ij} são as variáveis aleatórias não correlacionadas (resíduo).

Como o modelo definitivo, no qual a relação entre as variáveis $x's$ é explicada precisamente por meio de m fatores ($m < p$) e os resíduos são não correlacionados, a análise de componentes principais não é adequada para ajustar o modelo.

A distinção, então, é que a análise de fatores com m fatores assume um modelo matemático definido, e a análise de componentes principais não assume. A análise de componentes principais considera os m fatores mais importantes, sem necessariamente assumir que os resíduos são completamente não correlacionados, ou que são normalmente distribuídos. O ponto crítico em decidir pela aplicação ou não a análise de fatores, é a validade do modelo de fator e este é o ponto que deve ser discutido.

A interpretação do mesmo exemplo usado em componentes principais e aplicado em análise de fatores, considerando que a análise de componentes principais pode ser assumida como uma aproximação da análise de fatores, os passos aplicados são os que seguem.

- O conjunto de dados é o mesmo usado quando da análise de componentes principais, ou seja:

$x_1 =$ Diâmetro (DAP, m)
$x_2 =$ Altura (H, m)
$x_3 =$ Área transversal média (G, m²)
$x_4 =$ Fator de forma (FF, adimensional)

O primeiro passo realizado é a aplicação da análise de componentes principais usando a matriz correlação (matriz variância-covariância padronizada) por razões de independência de escala. Assim, a matriz correlação obtida dos dados originais, tem diagonal da matriz $S = 4{,}00$.

O segundo passo foi obter os valores e vetores característicos da matriz correlação. Quando se transpõe os vetores característicos tem-se, então, a matriz de transformação (matriz modal).

Obtém-se a seguir a matriz de dados padronizados, sendo multiplicada pela matriz de transformação (matriz modal) para se obter a matriz de dados transformada (matriz de dados não correlacionada). Assim,

$$Y = X_z T$$

Em que: Y é a matriz de dados transformada; X_z é a matriz de dados padronizada e T é matriz de transformação que é a matriz modal, obtida transpondo-se os vetores característicos da matriz de correlação, usando-se a matriz de dados originais.

Uma vez obtidas as matrizes variância-covariância e correlação, a partir da matriz de dados transformados, verifica-se se existe correlação para o novo conjunto de variáveis (Y), o que resultou em ausência de correlação. A razão disto se deveu ao fato de que na diagonal se tem as variâncias, que são as mesmas obtidas com os valores característicos derivados da matriz correlação dos dados originais. Como fora da diagonal os valores são iguais a zero, a teoria de componentes principais mostra que se tem um novo conjunto de dados não correlacionados.

Complementando, da teoria também se tem:

$$\sum_j \lambda_j = \sum_j s_{p_{cj}}^2 = traço\ S_x = \sum_i S_i^2$$

Após estes cálculos, foi então obtida a matriz de carregamento (*loading matrix*), na qual é possível se observar a correlação entre os valores originais e os valores transformados (Tabela 5.1).

Tabela 5.1: Matriz de carregamento, na qual se observa a correlação entre os valores originais e os valores transformados.

Valor original	Valor transformado			
	1	2	3	4
x_1	0,78144	0,08616	0,55921	-0,26309
x_2	0,89319	-0,06768	0,02402	0,44391
x_3	0,43348	0,82150	-0,35971	-0,08860
x_4	-0,66416	0,54653	0,45548	0,22962

Observando-se os valores característicos, que estão na diagonal da matriz variância-covariância transformada padronizada a partir da matriz de correlação, é possível determinar as percentagens da variação total, que são devidas a cada combinação linear. Assim, tem-se:

Componente principal	% da variação total	% cumulativa
CP1 = 2,03745	50,93%	50,93%
CP2 = 0,98556	24,63%	75,57%
CP3 = 0,65015	16,25%	91,82%
CP4 = 0,32685	8,17%	100,00%

Em que: CP é a componente principal = valor característico.

A primeira componente principal é um fator geral que absorve a variação comum entre todas as variáveis, pois os coeficientes são aproximadamente iguais para todas as variáveis originais, com exceção da variável x_3 (área transversal média). Desta componente verifica-se que a variável x_4 (fator de forma) decresce com o aumento do tamanho da árvore, pois seu coeficiente é negativo. A segunda componente principal indica uma forte relação entre a área transversal média e o fator de forma. A terceira componente principal mostra um outro aspecto da relação entre o fator de forma e as dimensões lineares que não são responsáveis pela área basal, a qual é uma função quadrada. A relação responsável pela quarta componente principal é de alguma forma similar à terceira, porém muito mais difícil de interpretar e menos importante.

Observando-se a percentagem cumulativa, foi eliminada a última componente principal, a qual representa somente 8,18% do total. Eliminando a última componente ainda se mantém explicada 91,82% da variação total.

A análise fatorial é o procedimento a ser utilizado depois de eliminadas as variáveis menos importantes. Para que ela possa ser empregada, dois testes devem ser aplicados, quais sejam:

- KMO (Kaiser – Meyer – Olkin): que mede a adequabilidade da amostragem, testando se as correlações parciais entre as variáveis são pequenas.

- Teste de esfericidade de Bartlett: testa se a matriz de correlação é uma matriz identidade, o que indica se o modelo fatorial é adequado.

Tendo sido atendidas as condicionantes, eliminado o último componente principal e aplicada a matriz de transformação (vetores característicos da matriz de correlação dos dados originais) por meio da ROTAÇÃO VARIMAX (*varimax rotation*), obtém-se resultados mais fáceis de serem interpretados.

As matrizes variância-covariância e correlação da matriz H, estão apresentadas nas Tabelas 5.2 e 5.3.

Tabela 5.2: Matriz variância-covariância e correlação da matriz *H*

Variáveis transformadas	Variáveis transformadas		
	Diâmetro	Altura	Área transversal
Diâmetro	1,20717	0,09906	0,58669
Altura	0,09906	1,02217	0,35389
Área transversal	0,58669	0,35389	1,44381

Tabela 5.3: Matriz de correlação da matriz *H*

Variáveis transformadas	Variáveis transformadas		
	Diâmetro	Altura	Área transversal
Diâmetro	1	0,089179	0,44440
Altura	0,08918	1	0,29131
Área transversal	0,44440	0,29131	1

É possível observar na Tabela 5.3, que a correlação entre diâmetro e altura é baixa e a análise fatorial mostrou isto muito claramente, pois a relação diâmetro e altura é naturalmente baixa. A correlação diâmetro e área transversal aumenta, pois a área transversal é derivada do diâmetro, enquanto a correlação entre altura e área transversal é baixa. Uma alternativa a este método é a execução de uma análise de imagem.

Na análise de imagem (*image analysis*) a maior parte da variação representada nos dados originais é conhecida ser única para as variáveis individuais. O que se deseja é ajustar a matriz de correlação para que se possa eliminar as porções únicas da variância das variáveis originais. Os procedimentos a seguir descritos devem ser executados.

- Obtenção da matriz de correlação (R) a partir dos dados originais;

- Cálculo do inverso da matriz correlação (R^{-1}) e da recíproca do inverso da matriz de correlação;

- Obtenção da matriz diagonal ($D_{m \cdot m}$), que tem na diagonal os elementos diagonais da recíproca do inverso da matriz correlação;

- Multiplicação da matriz R^{-1} pela matriz $D_{m \cdot m}$:

$$R^{-1}\ D_{m \cdot m}$$

Em seguida, essa matriz é subtraída da matriz identidade (I), obtendo-se a matriz A, que será posteriormente multiplicada pela matriz de dados padronizados. A matriz A corresponde, portanto, à matriz modificada de pesos.

- Após calculada a matriz A, se obtém a matriz comum (C) – *commom matrix*. Esta matriz contém as covariâncias imagens que representam as relações entre somente as porções comuns da variação original. A partir da matriz comum deve ser calculada as matrizes variância-covariância e correlação.

- A partir da matriz correlação da matriz comum, extrai-se os valores e vetores característicos. Em se transpondo os vetores característicos, obtém-se a matriz de transformação (matriz modal), sendo então possível determinar as percentagens que são devidas a cada combinação linear. Então, tem-se:

Componente principal	% da variação total	% cumulativa
PC1 = 3,28221	82,05%	
PC2 = 0,49309	12,33%	94,38%
PC3 = 0,18661	4,66%	99,04%
PC4 = 0,03808	0,95%	100%

Na Tabela 5.4 está demonstrada esta matriz de transformação (matriz modal) para a análise de imagem aplicada à rotação *varimax,* utilizando o mesmo algoritmo que foi usado no cálculo anterior (quando foram rotados apenas dois fatores).

Tabela 5.4: Matriz transformação (matriz modal) para análise de imagem aplicada à rotação *varimax.*

Variável	CP1	CP2	CP3	CP4
Diâmetro	0,54237	-0,10856	-0,12265	0,82402
Altura	0,43304	0,87780	0,14116	-0,14837
Área transversal	0,49996	-0,40794	0,71203	-0,27684
Fator de forma	-0,51802	0,22642	0,67678	0,47153

Observando-se a matriz de transformação da Tabela 5.4 (matriz modal), a primeira componente principal (CP1) absorveu quase toda a variação comum entre todas as variáveis (82,05%). Aqui se tem uma situação diferente da que ocorreu na primeira parte da análise, pois o diâmetro está contribuindo para a maior variação, e isto é devido ao ajuste feito na matriz correlação, quando foram eliminadas as porções únicas da variância das variáveis originais, o que não ocorreu na primeira análise. Pode-se observar que o comportamento das variáveis é melhor, pois na primeira componente principal o fator decresce com o aumento do diâmetro.

- Em seguida foi padronizada a matriz comum e multiplicada pela matriz de transformação (matriz modal), para se obter a matriz de dados transformados, que é o novo conjunto de dados não correlacionados.

$$Y = C\ Modal$$

- O próximo passo é verificar se o novo conjunto de dados não é correlacionado. Para tanto, foram calculadas as matrizes variância-covariância e a correlação e verificado se fora da diagonal os valores são zero ou muito próximos disto, devido aos arredondamentos. Os resultados revelaram que o novo conjunto de dados não é relacionado.

- A seguir foi calculada a matriz de carregamentos (*loading matrix*):

$$L = Modal\ \ D_{\sqrt{\lambda}}$$

Em que: L é a matriz de carregamento; $Modal$ é a matriz modal; $D_{\sqrt{\lambda}}$ é a diagonal da raiz quadrada dos valores característicos, que representam à correlação entre as variáveis originais e as variáveis transformadas.

Fazendo-se uma comparação entre as duas matrizes de carregamento obtidas na primeira parte e na segunda parte (análise de imagem), é possível observar que a matriz obtida pela análise de imagem se comportou melhor do que a matriz obtida pela análise de fatores componentes principais. Isto deve-se ao ajuste da matriz correlação explicado anteriormente. As Tabelas 5.5 e 5.6 mostram estas diferenças.

Tabela 5.5: Matriz de carregamento (*loading matrix),* correlação entre os valores originais e os valores transformados obtidos pela análise fatorial.

Variável original	Variáveis transformadas			
	DAP	H	G	FF
Diâmetro	0,78144	0,08616	0,55921	-0,26309
Altura	0,89319	-0,06768	0,02402	0,44391
Área transversal	0,43348	0,82150	-0,35971	-0,08860
Fator de forma	-0,66416	0,54653	0,45548	0,22962

Tabela 5.6: Matriz de carregamento (*loading matrix),* correlação entre os valores originais e os valores transformados obtidos pela análise de imagem.

Variável original	Variáveis transformadas			
	DAP	H	G	FF
Diâmetro	0,98261	-0,07623	-0,05299	0,16081
Altura	0,78454	0,61640	0,06098	-0,02896
Área transversal	0,90576	-0,28646	0,30759	-0,05403
Fator de forma	-0,93850	0,15899	0,29236	0,09202

Ao se analisar os valores característicos obtidos na matriz comum tem-se que:

Valor característico 1:	3,2822	82,05%
Valor característico 2:	0,4931	12,33%
Valor característico 3:	0,1866	4,66%
Valor característico 4:	0,0381	0,95%
Total	4,0000	100%

Calculando-se as percentagens cumulativas foi possível eliminar o último componente, que representa somente 0,95%. Eliminando o último componente se tem ainda 99,04% da variação comum explicada. Importante observar que no procedimento de análise fatorial de componentes principais, o último componente representava 8,17%, mas com o procedimento da análise de imagem este último componente representa somente 0,95%. Isto prova que trabalhando-se com análise de imagem se consegue eliminar as variações únicas ou singulares.

Após eliminar o último componente, deve-se então proceder a rotação *varimax* para recalcular as percentagens do traço para cada fator vetor. O que retorna da rotação *varimax* é a matriz $A = T\,Q$. Importante ressaltar que a matriz que foi usada na rotação *varimax* foi a matriz transformação sem o último componente. As Tabelas 5.7 e 5.8 ilustram estas duas matrizes.

Tabela 5.7: Matriz de transformação (matriz modal) sem o último componente.

Variável	CP1	CP2	CP3
Diâmetro	0,54237	-0,10856	-0,12265
Altura	0,43304	0,87780	0,14116
Área transversal	0,49996	-0,40794	0,71203
Fator de forma	-0,51802	0,22642	0,67679

Tabela 5.8: Matriz $A = T\ Q$ obtida pela análise de rotação *varimax*, pela matriz da Tabela 5.7.

Variável	CP1	CP2	CP3
DAP	0,50545	0,11615	0,22810
H	0,02239	0,98853	-0,01699
G	0,02384	-0,02559	0,96028
FF	-0,86224	0,09305	0,15982

Comparando as duas matrizes de transformação dos dois conjuntos, um obtido por meio da análise de fatores componentes principais e outro pela análise de imagem, observa-se que a matriz de transformação obtida pela análise de imagem se comporta melhor. As duas matrizes estão listadas abaixo, nas Tabelas 5.9 e 5.10

.

Tabela 5.9: Matriz de transformação rotacionada obtida por *varimax* da análise de fatores componentes principais sem o último componente.

Variável	CP1	CP2	CP3
DAP	-0,15418	-0,07139	0,87141
H	0,37772	0,12954	0,48749
G	-0,01549	0,98779	-0,00366
FF	-0,91286	0,04890	0,05458

Tabela 5.10: Matriz $A = T\ Q$ obtida pela análise de rotação *varimax* por meio da análise de imagem sem o último componente.

Variável	CP1	CP2	CP3
DAP	0,50545	0,11615	0,22810
H	0,02239	0,98853	-0,01699
G	0,02384	-0,02559	0,96028
FF	-0,86224	0,09305	0,15982

Na Tabela 5.9, na primeira componente (CP1) principal, a altura representa a maior variação (0,37772), na segunda componente principal é a área transversal média (0,98779) pois é uma função ao quadrado e na terceira componente principal é o diâmetro (0,87141).

5.6 EXEMPLO DE ANÁLISE DE FATORES NO MATLAB

```
MATRIZ X COM DADOS ORIGINAIS
X =
    1.4062   28.5000    1.6648    0.3475
    1.1735   32.3336    1.1590    0.3285
    1.3444   30.4000    1.6537    0.3598
    1.2646   37.6472    1.3128    0.3215
    1.2894   34.1056    1.4112    0.3526
    0.9857   31.0856    0.7885    0.3319
    1.0793   34.6664    0.9682    0.3370
    1.3833   46.4000    1.6144    0.3303
    1.4999   52.0000    2.1217    0.3387
    1.3110   40.0000    1.4922    0.3381
    1.3420   40.8424    1.6707    0.3221
    1.3690   36.0000    1.5474    0.3374
    1.7499   40.0000    1.6000    0.3267
    1.5999   33.6000    1.0000    0.3402
```

```
    1.0833   32.6664    2.4000    0.3261
    1.7272   37.8184    2.2000    0.3339
    1.2000   35.2000    1.0000    0.3224
    1.3939   39.2728    2.2000    0.3142
    1.8333   50.6664    1.2000    0.3021
    1.5166   44.0000    2.0000    0.3195
    1.9523   45.7144    1.4000    0.3234
    1.4333   40.0000    1.0000    0.3141
    1.3333   34.6664    0.6000    0.3337
    1.5909   42.9088    2.2000    0.3216
    1.3541   41.5000    1.6089    0.3374
```

ANÁLISE DE FATORES USANDO MATRIZ VARIÂNCIA/COVARIÂNCIA – DADOS ORIGINAIS

VETOR DE MÉDIAS

```
MEDIA=mean(X)
MEDIA =
    1.4087   38.4798    1.5125    0.3304
```

VETOR DE DESVIOS PADRÃO

```
DESVIO=sqrt(var(X))
DESVIO =
    0.2363    6.1235    0.4847    0.0127
```

MATRIZ VARIÂNCIA/COVARIÂNCIA

```
S=cov(X)
S =
    0.0558    0.8521    0.0265   -0.0008
    0.8521   37.4969    0.8417   -0.0402
    0.0265    0.8417    0.2349   -0.0001
   -0.0008   -0.0402   -0.0001    0.0002
```

VARIÂNCIA TOTAL

```
TracoS=trace(S)
TracoS =
   37.7878
```

AUTOVALORES E AUTOVETORES DA MATRIZ VARIÂNCIA E COVARIÂNCIA

```
[VETOR,VALOR]=eig(S)
VETOR =
   -0.0015   -0.9989   -0.0405    0.0227
    0.0012    0.0218    0.0235    0.9995
   -0.0035    0.0411   -0.9989    0.0226
    1.0000   -0.0014   -0.0036   -0.0011
VALOR =
    0.0001         0         0         0
         0    0.0362         0         0
         0         0    0.2162         0
         0         0         0   37.5353
```

OBTENÇÃO DOS AUTOVALORES EM ORDEM CRESCENTE

```
[VALORAC , IND] = sort(diag(VALOR))
VALORAC =
    0.0001
    0.0362
    0.2162
   37.5353
```

```
IND =
     1
     2
     3
     4

OBTENÇÃO DOS AUTOVALORES ORDENADOS
VALOR = VALORAC(end:-1:1)
VALOR =
   37.5353
    0.2162
    0.0362
    0.0001

AUTOVETORES ORDENADOS POR AUTOVALORES
INDICE=IND(end : -1 : 1)
INDICE =
     4
     3
     2
     1

VETOR=VETOR(: , INDICE)
VETOR =
    0.0227   -0.0405   -0.9989   -0.0015
    0.9995    0.0235    0.0218    0.0012
    0.0226   -0.9989    0.0411   -0.0035
   -0.0011   -0.0036   -0.0014    1.0000

AUTOVALORES ACUMULADOS
ACUM=cumsum(VALOR)
ACUM =
   37.5353
   37.7515
   37.7877
   37.7878

VARIÂNCIA EXPLICADA ACUMULADA PELOS FATORES
EXPLICADA=ACUM/trace(S)
EXPLICADA =
    0.9933
    0.9990
    1.0000
    1.0000

CÁLCULO DOS LOADINGS
A)RAIZ QUADRADA DOS AUTOVALORES
RAIZVALOR=VALOR.^0.5
RAIZVALOR =
    6.1266
    0.4650
    0.1902
    0.0107

B)DESVIO PADRAO DA MATRIZ VARIÂNCIA/COVARIÂNCIA
DESVIO=diag(S).^0.5
```

```
DESVIO =
    0.2363
    6.1235
    0.4847
    0.0127
```

C)NÚMERO DE VARIÁVEIS

```
n=length(DESVIO)
n =
     4
```

D) CORRELAÇÃO DOS FATORES COM AS VARIÁVEIS ORIGINAIS - LOADINGS

```
for i=1 : n
for j=1 :n
LOADINGS(j,i)=VETOR(j,i)*RAIZVALOR(i)/DESVIO(j);
end
end
LOADINGS
LOADINGS =
    0.5895   -0.0798   -0.8038   -0.0001
    1.0000    0.0018    0.0007    0.0000
    0.2853   -0.9583    0.0161   -0.0001
   -0.5171   -0.1303   -0.0210    0.8457
```

ESCOLHA DO NÚMERO DE FATORES – CRITÉRIO: AUTOVALORES MAIORES DO QUE UM

```
VALORES=VALOR( VALOR > 1)
VALORES =
   37.5353
```

NÚMERO DE FATORES

```
NF=length(VALORES)
NF =
     1
```

VETORES ASSOCIADOS AOS FATORES SELECIONADOS

```
VETORES=VETOR(: , 1:NF)
VETORES =
    0.0227
    0.9995
    0.0226
   -0.0011
```

LOADINGS ASSOCIADOS AOS FATORES SELECIONADOS

```
L = LOADINGS(: , 1:NF)
L =
    0.5895
    1.0000
    0.2853
   -0.5171
```

COMUNALIDADES

```
L2 = L.^2
L2 =
    0.3475
    1.0000
    0.0814
    0.2674
```

```
COMUNALIDADES=sum(L2 , 2)
COMUNALIDADES =
    0.3475
    1.0000
    0.0814
    0.2674
```

VARIÂNCIAS ESPECÍFICAS

```
s=diag(S)
s =
    0.0558
   37.4969
    0.2349
    0.0002
VESPECIFICA=s-COMUNALIDADES
VESPECIFICA =
   -0.2916
   36.4969
    0.1535
   -0.2672
```

MATRIZ DE RESÍDUOS

```
MATRIZRESIDUOS=S-(L*L'+diag(VESPECIFICA(: , :)))
MATRIZRESIDUOS =
   -0.0000    0.2626   -0.1416    0.3040
    0.2626         0    0.5564    0.4769
   -0.1416    0.5564   -0.0000    0.1474
    0.3040    0.4769    0.1474    0.0000
```

ESCORES PARA OS FATORES SELECIONADOS

```
FE = X*L
FE =
   29.6242
   33.1861
   31.4782
   38.6009
   35.0859
   31.7199
   35.4045
   47.5051
   53.3143
   41.0236
   41.9435
   37.0739
   41.3190
   34.6524
   33.8210
   39.2915
   36.0259
   40.5596
   51.9331
   45.2993
   47.0973
   40.9677
   35.4509
   44.3079
   42.5827
```

ANÁLISE DE FATORES USANDO MATRIZ DE CORRELAÇÃO - DADOS PADRONIZADOS Z

TAMANHO DA MATRIZ DE DADOS

```
TAMANHO=size(X)
TAMANHO =
    25     4
LINHA=TAMANHO(1,1)
LINHA =
    25
COLUNA=TAMANHO(1,2)
COLUNA =
     4
```

VETOR DE MÉDIAS

```
MEDIA=mean(X)
MEDIA =
    1.4087   38.4798    1.5125    0.3304
```

VETOR DE DESVIOS PADRÃO

```
DESVIO = sqrt(var(X))
DESVIO =
    0.2363    6.1235    0.4847    0.0127
```

MATRIZ DE DADOS PADRONIZADOS

```
for j=1 :COLUNA
for i= 1:LINHA
Z(i,j)=(X(i,j)-MEDIA(j))/ DESVIO(j);
end
end
Z
Z =
   -0.0104   -1.6298    0.3141    1.3456
   -0.9950   -1.0037   -0.7294   -0.1520
   -0.2719   -1.3195    0.2912    2.3152
   -0.6095   -0.1360   -0.4121   -0.7037
   -0.5046   -0.7143   -0.2091    1.7476
   -1.7897   -1.2075   -1.4938    0.1160
   -1.3936   -0.6227   -1.1231    0.5180
   -0.1073    1.2934    0.2102   -0.0101
    0.3861    2.2079    1.2568    0.6520
   -0.4132    0.2483   -0.0420    0.6047
   -0.2820    0.3858    0.3263   -0.6564
   -0.1678   -0.4050    0.0719    0.5495
    1.4440    0.2483    0.1804   -0.2938
    0.8093   -0.7969   -1.0575    0.7702
   -1.3767   -0.9494    1.8310   -0.3411
    1.3479   -0.1080    1.4184    0.2737
   -0.8829   -0.5356   -1.0575   -0.6328
   -0.0624    0.1295    1.4184   -1.2791
    1.7969    1.9902   -0.6448   -2.2329
    0.4568    0.9015    1.0057   -0.8614
    2.3004    1.1815   -0.2322   -0.5540
    0.1043    0.2483   -1.0575   -1.2870
   -0.3188   -0.6227   -1.8828    0.2579
    0.7712    0.7233    1.4184   -0.6958
   -0.2308    0.4932    0.1988    0.5495
```

MATRIZ DE CORRELAÇÃO DOS DADOS ORIGINAIS

```
CORRX=corr(X)
CORRX =
    1.0000    0.5888    0.2317   -0.2776
    0.5888    1.0000    0.2836   -0.5173
    0.2317    0.2836    1.0000   -0.0231
   -0.2776   -0.5173   -0.0231    1.0000
```

AUTOVALORES E AUTOVETORES DA MATRIZ DE CORRELAÇÃO

```
[VETOR , VALOR ] =eig(CORRX)
VETOR =
    0.5475    0.4602    0.6935    0.0868
    0.6257   -0.7765    0.0298   -0.0682
    0.3037    0.1550   -0.4461    0.8275
   -0.4653   -0.4016    0.5649    0.5505
VALOR =
    2.0374         0         0         0
         0    0.3268         0         0
         0         0    0.6501         0
         0         0         0    0.9856
```

OBTENÇÃO DOS AUTOVALORES EM ORDEM CRESCENTE

```
[VALORAC , IND] = sort(diag(VALOR))
VALORAC =
    0.3268
    0.6501
    0.9856
    2.0374
IND =
     2
     3
     4
     1
```

OBTENÇÃO DOS AUTOVALORES ORDENADOS

```
VALORN = VALORAC(end:-1:1)
VALORN =
    2.0374
    0.9856
    0.6501
    0.3268
```

AUTOVETORES ORDENADOS POR AUTOVALORES

```
INDICE=IND(end : -1 : 1)
INDICE =
     1
     4
     3
     2
VETORN=VETOR(: , INDICE)
VETORN =
    0.5475    0.0868    0.6935    0.4602
    0.6257   -0.0682    0.0298   -0.7765
    0.3037    0.8275   -0.4461    0.1550
   -0.4653    0.5505    0.5649   -0.4016
```

AUTOVALORES ACUMULADOS

```
AVAN=cumsum(VALORN)
```

```
AVAN =
    2.0374
    3.0230
    3.6732
    4.0000
```

VARIÂNCIA EXPLICADA ACUMULADA PELOS FATORES

```
VEN=AVAN/trace(CORRX)
VEN =
    0.5094
    0.7558
    0.9183
    1.0000
```

CÁLCULO DOS LOADINGS

A)RAIZ QUADRADA DOS AUTOVALORES

```
RAIZVALORN=VALORN.^0.5
RAIZVALORN =
    1.4274
    0.9928
    0.8063
    0.5717
```

B) NÚMERO DE VARIÁVEIS

```
n=length(RAIZVALORN)
n =
     4
```

C) CORRELAÇÃO DOS FATORES COM AS VARIÁVEIS ORIGINAIS – LOADINGS

```
for i=1 : n
for j=1 :n
LOADINGSN(j,i)=VETORN(j,i)*RAIZVALORN(i);
end
end
LOADINGSN
LOADINGSN =
    0.7814    0.0862    0.5592    0.2631
    0.8932   -0.0677    0.0240   -0.4439
    0.4335    0.8215   -0.3597    0.0886
   -0.6642    0.5465    0.4555   -0.2296
```

ESCOLHA DO NÚMERO DE FATORES – CRITÉRIO: NÚMERO DE FATORES QUE EXPLICAM ATÉ 80% DA VARIÂNCIA TOTAL

```
VALORESN=VEN(VEN <0.8)
VALORESN =
    0.5094
    0.7558
```

NÚMERO DE FATORES

```
NFN=length(VALORESN)
NFN =
     2
```

VETORES ASSOCIADOS AOS FATORES SELECIONADOS

```
VETORESN=VETORN(: , 1:NFN)
VETORESN =
```

```
    0.5475    0.0868
    0.6257   -0.0682
    0.3037    0.8275
   -0.4653    0.5505
```

LOADINGS ASSOCIADOS AOS FATORES SELECIONADOS

```
LN=LOADINGSN(: , 1:NFN)
LN =
    0.7814    0.0862
    0.8932   -0.0677
    0.4335    0.8215
   -0.6642    0.5465
```

COMUNALIDADES

```
L2N=LN.^2
L2N =
    0.6106    0.0074
    0.7978    0.0046
    0.1879    0.6749
    0.4411    0.2987

COMUNALIDADESN = sum(L2N,2)
COMUNALIDADESN =
    0.6181
    0.8024
    0.8628
    0.7398
```

VARIÂNCIAS ESPECÍFICAS

```
r=diag(CORRX)
r =
     1
     1
     1
     1
VESPECIFICAN=r-COMUNALIDADESN
VESPECIFICAN =
    0.3819
    0.1976
    0.1372
    0.2602
```

MATRIZ DE RESÍDUOS

```
MATRIZRESIDUOSN=CORRX-(LN*LN'+diag(VESPECIFICAN(: , :)))
MATRIZRESIDUOSN =
         0   -0.1034   -0.1778    0.1943
   -0.1034         0   -0.0480    0.1129
   -0.1778   -0.0480         0   -0.1842
    0.1943    0.1129   -0.1842         0
```

ESCORES PADRONIZADOS PARA OS FATORES SELECIONADOS

```
FN=Z*LN
FN =
   -2.2213    1.1029
   -1.8893   -0.7001
   -2.8024    1.5704
   -0.3090   -0.7665
   -2.2837    0.7882
```

```
   -3.2017   -1.2363
   -2.4761   -0.7174
    1.1692    0.0704
    2.3856    1.2727
   -0.5210    0.2436
    0.7017   -0.1411
   -0.8266    0.3724
    1.6235    0.0953
   -1.0494   -0.3241
   -0.9035    1.2634
    1.3899    1.4382
   -1.2064   -1.2544
    1.5313    0.4520
    4.3852   -1.7299
    2.1702    0.3338
    3.1202   -0.3753
    0.6996   -1.5799
   -1.7928   -1.3911
    2.3256    0.8024
   -0.0187    0.4104
```

GRÁFICO DOS ESCORES PADRONIZADOS ENTRE OS FATORES SELECIONADOS

```
scatter(FN(: , 1),FN(: , 2))
xlabel('FATOR 1')
ylabel(' FATOR 2')
axis([-4 4 -4 4])
```

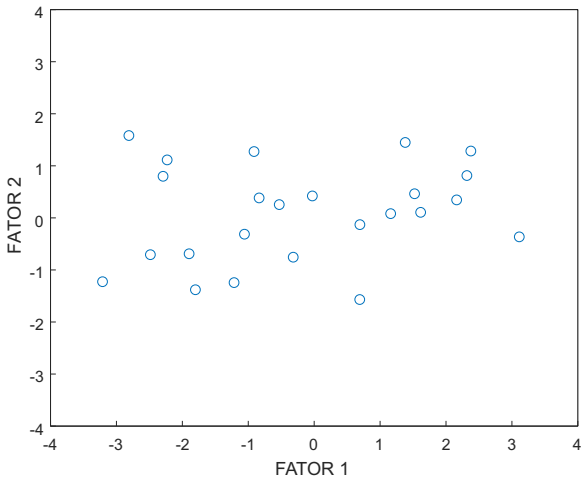
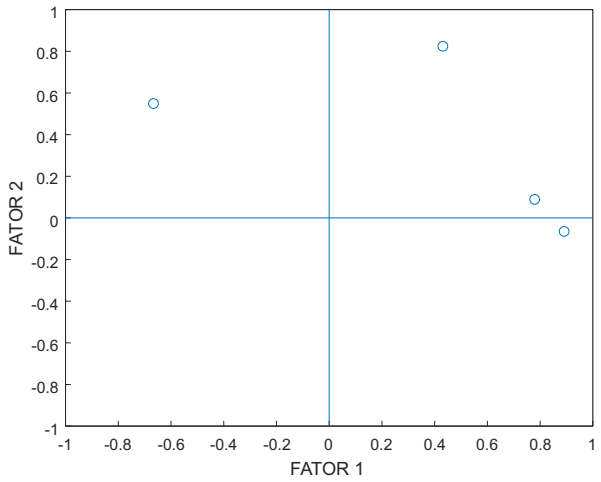


GRÁFICO DOS FATORES SELECIONADOS

```
scatter(LN(: , 1),LN(: , 2))
xlabel ('FATOR 1')
ylabel ('FATOR 2')
axis([-1 1 -1 1])
line(xlim,[0,0])
line([0,0],ylim)
```



ROTAÇÃO DOS FATORES

```
% MATRIZ DE LOADINGS
X=  [  0.7814    0.0862
    0.8932   -0.0678
    0.4335    0.8215
   -0.6642    0.5465]

% DIMENSÕES DA MATRIZ DE LOADINGS
  [NV,NF]=size(X);

% DEFINIÇÃO DA MATRIZ
   c=X;
   XT=X';

% INICIO DO LOOP ITERATIVO
  ITER=0;

% INICIALIZAÇÃO DA MATRIZ DE ROTAÇÃO (T)
   T=eye(NF);

% NÚMERO MÁXIMO DE ITERAÇÕES
   NMI=200;

% INÍCIO DO PROCESSO ITERATIVO
   while ITER<NMI

% LOADINGS ATUAIS ROTACIONADOS
   LR=X*T;

% POTÊNCIA CÚBICA (MAXIMIZA A VARIÂNCIA DOS QUADRADOS DOS LOADINGS)
   PC=LR.^3;
   PC=real(PC);
   c5=LR;

%  MATRIZ DIAGONAL F
F=(1/NV)*diag(diag(LR'*LR));
   F=real(F);
```

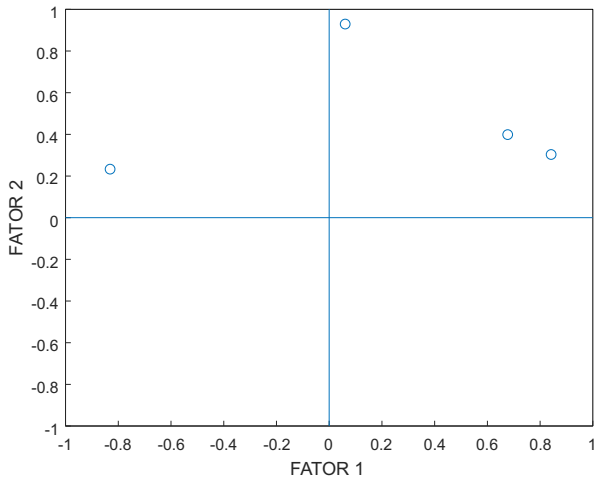
```
%  MATRIZ GR ( GRADIENTE DA ROTAÇÃO )
   GR=XT*(PC-LR*F);
   GR=real(GR);

%  ATUALIZAÇÃO DA MATRI DE ROTAÇÃO
   T=(inv(sqrtm(GR*GR')))*GR;
   T=real(T);

% ATUALIZAÇÃO DO CONTADOR
ITER=ITER+1;
end
T
T =
    0.9140    0.4057
   -0.4057    0.9140
LR
LR =

    0.6792    0.3958
    0.8439    0.3004
    0.0630    0.9267
   -0.8288    0.2301

% GRÁFICO DOS FATORES SELECIONADOS ROTACIONADOS
scatter(LR(:,1),LR(:,2))
xlabel ('FATOR 1')
ylabel ('FATOR 2')
axis ([-1 1 -1 1])
line(xlim,[0,0])
line([0,0],ylim)
```



5.7 EXEMPLO DE ANÁLISE DE IMAGEM NO MATLAB

MATRIZ DE DADOS

```
X =
    1.4062   28.5000    1.6648    0.3475
    1.1735   32.3336    1.1590    0.3285
    1.3444   30.4000    1.6537    0.3598
    1.2646   37.6472    1.3128    0.3215
    1.2894   34.1056    1.4112    0.3526
    0.9857   31.0856    0.7885    0.3319
    1.0793   34.6664    0.9682    0.3370
    1.3833   46.4000    1.6144    0.3303
    1.4999   52.0000    2.1217    0.3387
    1.3110   40.0000    1.4922    0.3381
    1.3420   40.8424    1.6707    0.3221
    1.3690   36.0000    1.5474    0.3374
    1.7499   40.0000    1.6000    0.3267
    1.5999   33.6000    1.0000    0.3402
    1.0833   32.6664    2.4000    0.3261
    1.7272   37.8184    2.2000    0.3339
    1.2000   35.2000    1.0000    0.3224
    1.3939   39.2728    2.2000    0.3142
    1.8333   50.6664    1.2000    0.3021
    1.5166   44.0000    2.0000    0.3195
    1.9523   45.7144    1.4000    0.3234
    1.4333   40.0000    1.0000    0.3141
    1.3333   34.6664    0.6000    0.3337
    1.5909   42.9088    2.2000    0.3216
    1.3541   41.5000    1.6089    0.3374
```

VETOR DE MÉDIAS

```
MEDIA=mean(X)
MEDIA =
    1.4087   38.4798    1.5125    0.3304
```

VETOR DE DESVIOS PADRÃO

```
DESVIO=sqrt(var(X))
DESVIO =
    0.2363    6.1235    0.4847    0.0127
```

PADRONIZAÇÃO DOS DADOS ORIGNAIS

```
[LINHA,COLUNA]=size(X)
LINHA =
    25
COLUNA =
     4
for j=1 :COLUNA
for i= 1:LINHA
Z(i,j)=(X(i,j)-MEDIA(j))/ DESVIO(j);
end
end
```

MATRIZ DE CORRELAÇÃO DOS DADOS ORIGINAIS

```
R=corr(X)
R =
    1.0000    0.5888    0.2317   -0.2776
    0.5888    1.0000    0.2836   -0.5173
    0.2317    0.2836    1.0000   -0.0231
   -0.2776   -0.5173   -0.0231    1.0000
```

INVERSA DA MATRIZ DE CORRELAÇÃO R

```
RINV=inv(R)
RINV =
    1.5425   -0.8993   -0.1032   -0.0394
   -0.8993    2.0429   -0.3526    0.7991
   -0.1032   -0.3526    1.1196   -0.1852
   -0.0394    0.7991   -0.1852    1.3982
```

MATRIZ RECÍPROCA DA INVERSA RINV

```
REC=RINV.^-1
REC =
    0.6483   -1.1120   -9.6890  -25.3610
   -1.1120    0.4895   -2.8363    1.2515
   -9.6890   -2.8363    0.8932   -5.4003
  -25.3610    1.2515   -5.4003    0.7152
```

DIAGONAL DA RECÍPROCA DA INVERSA

```
D=diag(diag(REC))
D =
    0.6483         0         0         0
         0    0.4895         0         0
         0         0    0.8932         0
         0         0         0    0.7152
B=RINV*D
B =
    1.0000   -0.4402   -0.0922   -0.0282
   -0.5830    1.0000   -0.3149    0.5715
   -0.0669   -0.1726    1.0000   -0.1324
   -0.0256    0.3911   -0.1654    1.0000
II=eye(COLUNA)
II =

     1     0     0     0
     0     1     0     0
     0     0     1     0
     0     0     0     1
```

MATRIZ MODIFICADA DE PESOS – MATRIZ A

```
A=II-B
A =
         0    0.4402    0.0922    0.0282
    0.5830         0    0.3149   -0.5715
    0.0669    0.1726         0    0.1324
    0.0256   -0.3911    0.1654         0
```

MATRIZ COMUM

```
CC=Z*A
CC =
   -0.8948   -0.4767   -0.2916    0.9727
   -0.6379   -0.5045   -0.4329    0.4490
   -0.6906   -0.9750   -0.0577    0.7850
   -0.1248   -0.0642   -0.2154    0.0059
   -0.3858   -0.9418    0.0176    0.3663
   -0.8010   -1.0911   -0.5260    0.4418
   -0.4250   -1.0099   -0.2389    0.1679
    0.7679   -0.0070    0.3957   -0.7144
    1.3881    0.1318    0.8387   -1.0845
```

```
    0.1574   -0.4257    0.1401   -0.1591
    0.2300    0.1889   -0.0131   -0.1852
   -0.2172   -0.2764   -0.0521    0.2362
    0.1493    0.7817    0.1627   -0.0773
   -0.5157   -0.1275   -0.0490    0.3382
   -0.4397   -0.1566   -0.4823    0.7462
    0.0389    0.7311    0.1355    0.2876
   -0.3992   -0.3237   -0.3547    0.1412
    0.1377    0.7176   -0.1765    0.1121
    1.0601    1.5531    0.4230   -1.1721
    0.5709    0.7116    0.1835   -0.3691
    0.6591    1.1893    0.4925   -0.6411
    0.0411    0.3668   -0.1251   -0.2790
   -0.4825   -0.5662   -0.1828    0.0976
    0.4988    0.8564    0.1838   -0.2038
    0.3149   -0.2823    0.2249   -0.2621
```

MATRIZ VARIÂNCIA-COVARIÂNCIA DA MATRIZ COMUM

```
VARCOV_CC =
    0.3517    0.3034    0.1719   -0.2959
    0.3034    0.5105    0.1295   -0.2376
    0.1719    0.1295    0.1068   -0.1414
   -0.2959   -0.2376   -0.1414    0.2848
```

MATRIZ DE CORRELAÇÃO DA MATRIZ COMUM

```
R_CC =
    1.0000    0.7160    0.8869   -0.9350
    0.7160    1.0000    0.5544   -0.6231
    0.8869    0.5544    1.0000   -0.8106
   -0.9350   -0.6231   -0.8106    1.0000
```

AUTOVALORES E AUTOVETORES DA MATRIZ DE CORRELAÇÃO DA MATRIZ COMUM

```
VETOR =
    0.5424    0.1086    0.8240   -0.1227
    0.4330   -0.8778   -0.1484    0.1412
    0.5000    0.4079   -0.2768    0.7120
   -0.5180   -0.2264    0.4715    0.6768

VALOR =
    3.2822         0         0         0
         0    0.4931         0         0
         0         0    0.0381         0
         0         0         0    0.1866
```

AUTOVALORES ORDENADOS

```
[VALOR , IND] = sort(diag(VALOR))
VALOR =

    0.0381
    0.1866
    0.4931
    3.2822
IND =

     3
     4
     2
     1
```

```
VALOR = VALOR(end:-1:1)
VALOR =
    3.2822
    0.4931
    0.1866
    0.0381
```

AUTOVETORES ORDENADOR POR AUTOVALORES – MATRIZ MODAL

```
INDICE=IND(end : -1 : 1)
INDICE =
     1
     2
     4
     3
MODAL=VETOR(: , INDICE)
MODAL =
    0.5424    0.1086   -0.1227    0.8240
    0.4330   -0.8778    0.1412   -0.1484
    0.5000    0.4079    0.7120   -0.2768
   -0.5180   -0.2264    0.6768    0.4715
```

AUTOVALORES ACUMULADOS

```
AVA=cumsum(VALOR)
AVA =
    3.2822
    3.7753
    3.9619
    4.0000
```

VARIÂNCIA EXPLICADA ACUMULADA PELOS FATORES

```
VE=AVA/trace(R)
VE =
    0.8206
    0.9438
    0.9905
    1.0000
```

MATRIZ DE DADOS TRANSFORMADOS

```
Y = CC * MODAL
Y =
   -1.3414   -0.0179    0.4932   -0.1272
   -1.0134    0.0953    0.0026   -0.1192
   -1.2323    0.5796    0.4373   -0.0383
   -0.2063   -0.0464   -0.1431   -0.0309
   -0.7981    0.7091    0.1748   -0.0103
   -1.3988    0.5561   -0.1313   -0.1442
   -0.8742    0.7049   -0.1469   -0.0551
    0.9814    0.4127   -0.2969    0.1874
    1.7911    0.6226   -0.2884    0.3807
    0.0535    0.4839   -0.0873    0.0791
    0.2960   -0.1043   -0.1362    0.0778
   -0.3859    0.1443    0.1104   -0.0122
    0.5409   -0.5861    0.1556   -0.0744
   -0.5346   -0.0406    0.2393   -0.2330
   -0.9340   -0.2760    0.1935    0.1463
    0.2565   -0.6474    0.3896    0.0217
   -0.6071    0.0641   -0.1538   -0.1162
    0.2391   -0.7124    0.0346    0.1087
```

```
    2.0662   -0.8103   -0.4028   -0.0267
    0.9007   -0.4042   -0.0887    0.1400
    1.4508   -0.6263    0.0038   -0.0720
    0.2631   -0.3054   -0.2311   -0.1175
   -0.6488    0.3479   -0.0849   -0.2169
    0.8389   -0.5765    0.0527    0.1370
    0.2968    0.4330   -0.0957    0.1155
```

MATRIZ VARIÂNCIA-COVARIÂNCIA DA MATRIZ DE DADOS TRANSFORMADOS - Y

```
VARCOV_Y=cov(Y)
VARCOV_Y =
    0.9402   -0.2208   -0.1173    0.0740
   -0.2208    0.2408   -0.0095    0.0017
   -0.1173   -0.0095    0.0530   -0.0085
    0.0740    0.0017   -0.0085    0.0198
```

MATRIZ DE CORRELAÇÃO DA MATRIZ DE DADOS TRANSFORMADOS - Y

```
R_Y=corr(Y)
R_Y =
    1.0000   -0.4640   -0.5256    0.5427
   -0.4640    1.0000   -0.0838    0.0253
   -0.5256   -0.0838    1.0000   -0.2619
    0.5427    0.0253   -0.2619    1.0000
```

RAIZ QUADRADA DOS AUTOVALORES

```
RAIZVALOR=VALOR.^0.5
RAIZVALOR =
    1.8117
    0.7022
    0.4320
    0.1952
```

CÁLCULO DOS CARREGAMENTOS – LOADINGS

```
n=length(RAIZVALOR)
n =
     4
for i=1 : n
for j=1 :n
LOADINGS(j,i)=MODAL(j,i)*RAIZVALOR(i);
end
end
LOADINGS =
    0.9826    0.0762   -0.0530    0.1608
    0.7845   -0.6164    0.0610   -0.0290
    0.9058    0.2865    0.3076   -0.0540
   -0.9385   -0.1590    0.2924    0.0920
```

5.8 EXEMPLO DE ANÁLISE DE FATORES NO JULIA

```
using DataFrames, CSV

#Dados originais
X=CSV.read("Z:\\Alexandre\\Julia\\Multivariada\\dendrometria.csv",
DataFrame)
```

```
#Matriz X com dados originais
X=[X.d X.h X.g X.f];

#Análise de fatores usando matriz variância-covariância - dados originais

using Statistics, LinearAlgebra

#Vetor de médias
Média=mean(X,dims=1)

#Vetor de desvios-padrão
Desvio=std(X,dims=1)

#Matriz variância-covariância
S=cov(X)

#Variância total
TraçoS=tr(S)

#Autovalores e autovetores da matriz variância-covariância
AV = eigen(S)

#Autovalores ordenados
Valor = reverse(AV.values, dims=1)

#Autovetores ordenados por autovalores
Vetor = reverse(AV.vectors, dims=2)

#Autovalores acumulados
AVA = cumsum(Valor)

#Variância explicada acumulada pelos fatores
VE = AVA ./ AVA[end]

#Cálculo dos loadings
#a)Raiz quadrada dos autovalores
RaizValor=Valor.^0.5

#b)Desvio padrão da matriz variância-covariância
Desvio=diag(S).^0.5

#c)Número de variáveis
n=length(Desvio)

#d)Correlação dos fatores com as variáveis originais - loadings
Loadings = [Vetor[c,l]*RaizValor[l]/Desvio[c] for l in 1:n, c in 1:n]'

#Escolha do número de fatores - critério -> autovalores maiores do que um
Valores=filter(x -> x > 1, Valor)
```

```
#Número de fatores
NF=length(Valores)

#Vetores associados aos fatores selecionados
Vetores=Vetor[:,1:NF]

#Loadings associados aos fatores selecionados
L=Loadings[:,1:NF]

#Comunalidades
L2=L.^2
Comunalidades= sum(L2, dims=2)

#Variâncias específicas
s=diag(S)
VEspecifica=s-Comunalidades

#Matriz de resíduos
MatrizResíduos=S-(L*L'+diagm(VEspecifica[:]))

#Escores para os fatores selecionados
FE=X*L

#Análise de fatores usando a matriz de correlação - dados padronizados Z

#Matriz de dados padronizados
Z = (X .- mean(X, dims=1)) ./ std(X, dims=1)

#Matriz de correlação dos dados originais
CORRX=cor(X)

#Autovalores e autovetores da matriz de correlação
AVN = eigen(CORRX)

#Autovalores ordenados
ValorN = reverse(AVN.values, dims=1)

#Autovetores ordenados por autovalores
VetorN = reverse(AVN.vectors, dims=2)

#Autovalores acumulados
AVAN = cumsum(ValorN)

#Variância explicada acumulada pelos fatores
VEN = AVAN ./ tr(CORRX)

#Cálculo dos loadings
#a)Raiz quadrada dos autovalores
RaizValorN=ValorN.^0.5
```

```
#b)Número de variáveis
n=length(RaizValorN)

#c)Correlação dos fatores com as variáveis originais - loadings
LoadingsN = [VetorN[c,l]*RaizValorN[l] for l in 1:n, c in 1:n]'

#Escolha do número de fatores - critério -> número de fatores que
explicam até 80% da variância total
ValoresN=filter(x -> x < 0.8, VEN)

#Número de fatores
NFN=length(ValoresN)

#Vetores associados aos fatores selecionados
VetoresN=VetorN[:,1:NFN]

#Loadings associados aos fatores selecionados
LN=LoadingsN[:,1:NFN]

#Comunalidades
L2N=LN.^2
ComunalidadesN= sum(L2N, dims=2)

#Variâncias específicas
r=diag(CORRX)
VEspecificaN=r-ComunalidadesN

#Matriz de resíduos
MatrizResíduosN=CORRX-(LN*LN'+diagm(VEspecificaN[:]))

#Escores padronizados para os fatores selecionados
FN = Z * LN

#Gráfico dos escores padronizados entre os fatores selecionados
using StatsPlots
scatter(FN[:,1], FN[:,2], legend = false)
ylims!((-4,4))
xlims!((-4,4))
xlabel!("Fator 1")
ylabel!("Fator 2")

#Gráfico dos fatores selecionados
scatter(LN[:,1], LN[:,2], framestyle =:zerolines,legend = false)
ylims!((-1,1))
xlims!((-1,1))
xlabel!("Fator 1")
ylabel!("Fator 2")
```

```
#Rotação dos fatores
#ROTAÇÃO VARIMAX

using LinearAlgebra
X=[0.7814 0.0862
0.8932 -0.06777
0.4335 0.8215
-0.6642 0.5465]

#Passo 1 – Matriz de loadings
X   # matriz de loadings não rotacionados

#Passo 2 – Dimensões da matriz de loadings
NV, NF = size(X)

#Passo 3 – Matriz transposta de loadings
XT = transpose(X)

#Passo 4 – Inicialização da matriz de rotação (T)
T = Matrix{Float64}(I, NF, NF)

#Passo 5 – Definição do número de iterações
#Número máximo de iterações
NMI = 200
ITER = 0

#Passo 6 – Início do processo iterativo
while ITER < NMI

#Passo 7 – Loadings atuais rotacionados
LR = X * T

#Passo 8 – Potência cúbica (maximiza a variância dos quadrados dos loadings)
PC = LR .^ 3

#Passo 9 – Matriz diagonal F
F = (1 / NV) * Diagonal(diag(transpose(LR) * LR))

#Passo 10 – Matriz GR (gradiente da rotação)
#Direção de ajuste da matriz de rotação para maximizar o critério Varimax.
GR = XT * (PC - LR * F)

#Passo 11 – Atualização da matriz de rotação
T = inv(sqrt(GR * transpose(GR))) * GR

#Passo 12 – Atualização do contador
ITER += 1
end

#Passo 13 – Loadings finais rotacionados
LR = X * T
```

```
#Gráfico para representar os fatores e os fatores rotacionados
scatter(LN[:,1], LN[:,2], framestyle =:zerolines, label = "L")
scatter!(LR[:,1], LR[:,2], framestyle =:zerolines, label = "L*")
ylims!((-1,1))
xlims!((-1,1))
xlabel!("Fator 1")
ylabel!("Fator 2")
```

5.9 EXEMPLO DE ANÁLISE DE IMAGEM NO JULIA

```
#Análise de imagem

using Statistics, LinearAlgebra

#Padronização dos dados originais
Z = (X .- mean(X, dims=1)) ./ std(X, dims=1)

#Matriz de correlação dos dados originais
R=cor(X)

#Inversa da matriz de correlação R
RINV=inv(R)

#Matriz recíproca da inversa RINV
REC=RINV.^-1

#Diagonal da recíproca da inversa
D=diag(REC)
D=diagm(D)

B=RINV*D

#Matriz identidade
II=Matrix(I, 4, 4)

#Matriz modificada de pesos – matriz A
A=II-B
#Matriz comum
CC=Z*A

#Matriz variância-covariância da matriz comum
VARCOV_CC=cov(CC)

#Matriz correlação da matriz comum
R_CC=cor(CC)

#Autovalores e autovetores da matriz de correlação da matriz comum
ValVet=eigen(R_CC)

#Autovalores ordenados
Valor = reverse(ValVet.values, dims=1)
```

```
#Autovetores ordenados por autovalores – matriz modal
Modal = reverse(ValVet.vectors, dims=2)

#Autovalores acumulados
AVA = cumsum(Valor)

#Variância explicada acumulada pelos fatores
VE = AVA ./ AVA[end]

#Matriz de dados transformados
Y = CC * Modal

#Matriz variância-covariância da matriz de dados transformados – Y
VarCov_Y=cov(Y)

#Matrix de correlação da matriz de dados transformados – Y
R_Y=cor(Y)

#Raíz quadrada dos autovalores
Raiz_Valor=Valor.^0.5

#Cáluclo dos carregamentos – loadings
n=length(R[:,1])
L = [Modal[c,l]*Raiz_Valor[l] for l in 1:n, c in 1:n]'
```

6 ANÁLISE DE VARIÂNCIA MULTIVARIADA - MANOVA

6.1 INTRODUÇÃO

Na análise de variância univariada que analisa uma única variável resposta, a soma dos quadrados total é dividida em fatores de variação dentro e entre grupos, com o intuito de se testar a hipótese de não haver diferenças entre as médias dos grupos, normalmente referenciados como tratamentos, sendo a hipótese testada pelo teste F de Fisher. As variâncias dentro dos grupos são assumidas serem homogêneas.

A extensão multivariada deste procedimento ocorre quando existem várias variáveis resposta. A hipótese da homogeneidade de variâncias também deve ser considerada, mas neste caso utiliza as matrizes de variância-covariância dentro dos grupos, que devem ser homogêneas.

O teste multivariado da hipótese da nulidade, agora entre os vetores de médias dos grupos, considera todas as variáveis simultaneamente, comparando-se os vetores de médias dos g grupos.

H_0: $\underline{\mu}_1 = \underline{\mu}_2 = \cdots = \underline{\mu}_g$.

H_1: pelo menos um dos vetores de médias $\underline{\mu}_i$ $(i = 1,\ 2,\ \ldots, g)$ é diferente.

6.2 OBTENÇÃO DA MANOVA POR MEIO DO CÁLCULO DOS DESVIOS

A execução de uma análise de variância multivariada pode ser feita de várias formas. Dentre elas, aquela em que a solução é feita pelo cálculo das somas de quadrado total, entre grupos e resíduos, calculando-se os desvios de forma semelhante aos métodos univariados. A outra forma de abordagem para solução de uma MANOVA é por meio do ajuste de modelos de regressão lineares múltiplos.

O objetivo é investigar se as populações têm o mesmo vetor médio e se não tem, e identificar quais variáveis diferem significativamente. Esse procedimento pressupõe que as amostras aleatórias (grupos) são independentes e que todas as populações têm a mesma matriz de variância-covariância $\sum$ e são normalmente distribuídas. Esquematicamente tem-se:

Amostra 1	**Amostra 2**	⋯	**Amostra g**
Tamanho n_1	Tamanho n_2		Tamanho n_g
Média $\underline{\overline{X}}_1$	Média $\underline{\overline{X}}_2$		Média $\underline{\overline{X}}_g$
Var/Cov: S_1	Var/Cov: S_2		Var/Cov: S_g
$\underline{X}_1 \sim N_p\left(\underline{\mu}_1, \sum_1\right)$	$\underline{X}_2 \sim N_p\left(\underline{\mu}_2, \sum_2\right)$	⋯	$\underline{X}_g \sim N_p\left(\underline{\mu}_g, \sum_g\right)$

As hipóteses testadas são:

H_0: $\underline{\mu}_1 = \underline{\mu}_2 = \cdots = \underline{\mu}_g$.

H_1: pelo menos uma das médias $\underline{\mu}_i$ $(i = 1,\ 2,\ \ldots, g)$ é diferente.

No método, que é a extensão dos procedimentos univariados, as somas de quadrados e produtos cruzados total, dentro e entre os grupos são calculadas da seguinte maneira:

A matriz de soma de quadrados e produtos cruzados total (em alguns livros denominada de S) é a matriz de ordem p (número de variáveis resposta) obtida ignorando-se os grupos. Essa matriz equivale a soma de quadrados total do caso univariado.

A matriz de soma de quadrados e produtos cruzados corrigida dentro dos grupos (W) é calculada a partir dos desvios dos valores de cada grupo em relação ao vetor de médias. Esta matriz equivale à soma dos quadrados dos resíduos do caso univariado. A matriz entre grupos (B) pode ser calculada por subtração e equivale a soma de quadrados entre grupos (tratamentos) quando se trata de análise variância univariada. Essa matriz também pode ser calculada por meio dos desvios do vetor médias de cada tratamento em relação ao vetor de médias geral, corrigindo-se para o tamanho das amostras.

A avaliação da hipótese da nulidade pode ser efetuada pelo critério de Wilks, $\Lambda^* = \frac{|W|}{|S|}$, que substitui o teste F de Fischer do caso univariado (Marriott, 1974). A estatística Λ^* varia no intervalo $0 \leq \Lambda^* \leq 1$, em que vetores de médias iguais retornam um valor da matriz B, que tende a zero e $\Lambda^*= 1$, aceitando a hipótese da nulidade. A existência de vetores de médias diferentes resulta em uma matriz B grande e, consequentemente, um valor que tende a $\Lambda^*= 0$, rejeitando a hipótese da nulidade, ou seja, indicando diferenças entre pelo menos um vetor de médias em relação aos demais.

A MANOVA para um delineamento inteiramente casualizado, em que p variáveis são analisadas simultaneamente, indica se existem ou não diferenças nos grupos (tratamentos). Se existirem diferenças entre grupos, o próximo passo é analisar a natureza das diferenças entre grupos. A análise discriminante multivariada propicia uma maneira de executar essa análise. A análise de variância multivariada, que considera a matriz de variância-covariância entre e dentro dos grupos pode ser conduzida de maneira similar ao caso univariado. A soma de quadrados e produtos cruzados, $\sum(x_i - \bar{x}_i)(x_j - \bar{x}_j)$ pode ser dividida em componentes dentro e entre grupos, permitindo a análise sumarizada na Tabela 6.1.

Tabela 6.1: Resumo da análise de variância multivariada pelo método de cálculo das somas de quadrados de modo análogo ao caso univariado.

Fonte de Variação	Graus de Liberdade	Somas de quadrados	Wilks Λ^*
Tratamentos	$g-1$	$B = \sum_{i=1}^{g} n_i \left(\underline{\overline{X}}_i - \underline{\overline{X}}\right)\left(\underline{\overline{X}}_i - \underline{\overline{X}}\right)'$	$\Lambda^* = \frac{\lvert W\rvert}{\lvert S\rvert}$
Resíduo	$\sum_{i=1}^{g} n_i - g$	$W = \sum_{i=1}^{g}\sum_{j=1}^{n_i} \left(\underline{X}_{ij} - \underline{\overline{X}}_i\right)\left(\underline{X}_{ij} - \underline{\overline{X}}_i\right)'$	
Total	$\sum_{i=1}^{g} n_i - 1$	$S = B + W$	

Adaptado de Marriott (1974).

Em que: a matriz S é expressa como a soma das matrizes B e W da soma de quadrados entre e dentro dos grupos, respectivamente. Isto é conhecido como análise de variância multivariada ou também como análise de dispersão. Isto é exatamente a mesma forma de uma análise de variância e covariância comum, mas o propósito é diferente. Aqui o objetivo não é efetuar uma análise de variância em uma única variável com o efeito eliminado das outras, mas sim pela análise simultânea de todas as variáveis $x's$. $\underline{X}_{ij}$= j-ésima observação do i-ésimo tratamento; $\underline{\overline{X}}_i$ = média da iésimo tratamento; $\underline{\overline{X}}$ = média global (todas amostras).

6.3 OBTENÇÃO DA MANOVA POR MEIO DE MODELOS DE REGRESSÃO LINEARES MÚLTIPLOS

Como observado previamente, a medição de várias variáveis resposta em situações experimentais é bem frequente. O procedimento comum em se efetuar uma análise de variância separada de cada variável resposta ignora a presença de correlações entre elas. A natureza multivariada de tais dados não pode ser ignorada.

A simples analogia multivariada da análise de variância univariada trata somente da detecção de diferenças entre categorias experimentais. Se for de interesse explorar a natureza de tais diferenças, então, os métodos da análise de discriminantes são apropriados. Para detectar diferenças entre respostas multivariadas, o vetor de respostas é substituído no modelo de ANOVA pela matriz de respostas. Então, procede-se ajustando os modelos máximo e reduzidos.

Um procedimento similar para obtenção dos resultados da MANOVA utiliza as covariáveis, nas quais os grupos de matrizes total e dentro são calculadas por ajustamento de funções para as covariáveis (sobre os resíduos). Se existem p variáveis respostas e q covariáveis, com V representando a soma de quadrados e produtos cruzados corrigidos, então,

$$V_{p\cdot p}^{*} = V_{p\cdot p} - V_{p\cdot q}V_{q\cdot q}^{-1}V_{q\cdot p}$$

Em que: V^{*} é a matriz da Soma de Quadrados e Produtos Cruzados (SQPC) ajustada.

Os graus de liberdade para os testes também precisam ser modificados como descrito por Marriott (1974). A situação é paralela como para regressão, com o termo na direita (termo subtrativo) sendo análogo à Soma de Quadrados de Regressão (SQR).

A classificação obtida pela MANOVA indica, para um delineamento inteiramente casualizado, se existem ou não diferenças entre grupos. Se as diferenças entre grupos são evidentes, o próximo passo é determinar a natureza das diferenças dos grupos. A análise discriminante multivariada propicia uma maneira de resolver isto (ver capítulo VII).

A relação entre o uso de matrizes usando-se a SQPC corrigida (Soma de Quadrados e Produtos Cruzados corrigida) e de matriz de covariância, quando se multiplica uma matriz por uma constante, multiplica-se também os valores característicos (*eignvalues*) pelo mesmo valor, mas isso não afeta os vetores característicos (*eignvectors*) normalizados.

$$\left[\frac{W}{(n-q)}\right]^{-1}\frac{B}{(q-1)} = [W^{-1}B]\left[\frac{(n-q)}{(q-1)}\right]^{(1),(2)}$$

(1) escalar multiplicador

(2) as raízes da matriz de SQPC (Soma de Quadrados e Produtos Cruzados corrigida) são maiores devido a esta multiplicação.

A análise de variância multivariada entre e dentro dos grupos pode ser conduzida na maneira usual para o conjunto de variáveis analisadas. O resumo da MANOVA está apresentado na Tabela 6.2.

Tabela 6.2: Resumo da análise de variância multivariada.

Fonte de variação	GL	Matriz da Soma de Produtos (S.P.)	Matriz de dispersão estimada
Entre grupos	$q-1$	B	
Dentro dos grupos	$n-q+1$	W	$V = \frac{W}{n-q+1}$
Total	n	S	

Adaptado de Marriott (1974). Em que: a matriz S é expressa como a adição das matrizes

B e W da soma de quadrados entre e dentro dos grupos, respectivamente. Isto é conhecido como análise de variância multivariada ou também como análise de dispersão.

Outro aspecto, refere-se em ajustar os modelos máximo e reduzido. O modelo linear multivariado toma a seguinte forma:

$$\hat{Y}^{(k\cdot t)} = B^{(k\cdot m)}X^{(m\cdot t)} + E^{(k\cdot t)}$$

Em que: $\hat{Y}$ é a matriz das respostas estimada, sendo que cada coluna corresponde a um vetor observado; X é a matriz de variáveis categóricas e das covariáveis que representam os grupos (tratamentos); B é a matriz de coeficientes estimada; e E é a matriz de erros.

Para encontrar a matriz resposta estimada, utiliza-se analogamente as equações normais multivariadas. Com isto obtém-se:

$$(X'X)B' = XY'$$

Para estimar a matriz B, a solução exigida é:

$$(XX')^{-1}(X'X)B' = (XX')^{-1}(XY')$$
$$IB = (XX')^{-1}(XY')$$
$$B' = (XX')^{-1}(XY')$$
$$B = (YX')(XX')^{-1}$$

Após ajustar os modelos máximo e reduzido apropriados, o próximo passo é comparar as suas adequabilidades. Na ANOVA univariada, o teste F seria o critério adequado e único. Na análise de variância multivariada (MANOVA), contudo, existem vários critérios possíveis. Um dos testes que é semelhante ao teste univariado F é a estatística Qui-quadrado baseada no determinante da matriz resultante da diferença entre os modelos máximo e reduzido, conforme segue:

$$|YY' - BXX'B|$$

Uma alternativa de ajuste do modelo $B = (X'X)^{-1}(X'Y)$, visando a obtenção da matriz inversa de $X'X$ é por meio do método de Moore Penrose. Assim, para o cálculo da MANOVA os seguintes passos são necessários:

$$S = X'X$$
$$B = (X'X)^{g}(X'Y)$$

Em que: S é a soma de quadrados e produtos não corrigidos; B é a solução de quadrados de produtos mínimos; $(X'X)^{g}$ é a inversa generalizada obtida pelo método de Moore Penrose. Destaca-se que existem alternativas para obtenção da inversa de $X'X$, sendo uma delas apresentada nos exemplos.

A partir disso, procede-se o cálculo das demais somas de quadrados. A soma dos quadrados dos resíduos corrigida (SQ_{Res}) é obtida por:

$$SQ_{Res} = Y'Y - B'X'Y$$

As matrizes de somas de quadrados dos tratamentos e total, não corrigidas, são dadas por:

$$SQ_{Trat} = B'X'Y$$
$$SQ_{Total} = Y'Y$$

As somas de quadrados total e dos tratamentos corrigidas são:

$$SQ_{TotalCor} = a'a$$
$$SQ_{TratCor} = SQ_{TotalCor} - SQ_{Res}$$

Em que: $a = Y - [1,1]\ Y\ \left(\frac{1}{m}\right)$; a é uma matriz de desvios; $[1,1]$ é uma matriz de uns de tamanho $m \cdot m$; m é o número de observações.

Finalmente, após estes ajustes pode-se calcular as estatísticas para a avaliação da hipótese de nulidade da MANOVA.

6.4 CRITÉRIOS DE AVALIAÇÃO DA MANOVA - TESTE ESTATÍSTICO DE WILKS

Wilks lambda compara a matriz de soma de quadrados e produtos dentro dos grupos com a matriz de soma de quadrados e produtos total. Isto é similar a estatística F no caso univariado que compara as somas de quadrados entre e dentro. O uso de determinantes faz com que a estatística seja reduzida a um escalar. A informação multivariada contida nas matrizes dentro e entre grupos, quando da separação dos vetores de médias, é canalizada em uma simples escala, na qual se pode determinar se a separação dos vetores de médias é significante. Este é um teste multivariado típico realizado na área florestal.

A estatística lambda de Wilks é dada por:

$$\Lambda = \frac{|W|}{|B+W|}$$

O valor de Wilks lambda está dentro do intervalo entre $0 \leq \Lambda \leq 1$. O teste baseado no Wilks lambda é um teste inverso, uma vez que se rejeita H_0 para valores pequenos de lambda. Se vetores de médias são iguais, deve-se ter a variância com baixo valor e o valor de lambda próximo a um. Por outro lado, se os vetores de médias se tornam mais dispersos, a variância entre torna-se maior que a variância dentro e o valor de lambda se aproxima de zero.

A transformação do valor de Wilks lambda para o valor de F é dada por meio das equações apresentadas na Tabela 6.3.

Tabela 6.3: Transformação do valor de Wilks lambda para o valor de F.

p	g	Distribuições amostrais
$p = 1$	$g > 2$	$\left(\frac{\sum_{i=1}^{g} n_i - g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1,n-g}$
$p = 2$	$g \geq 2$	$\left(\frac{\sum_{i=1}^{g} n_i - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(n-g-1)}$
$p \geq 1$	$g = 2$	$\left(\frac{\sum_{i=1}^{g} n_i - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p-1,n-p-1}$
$p \geq 1$	$g = 3$	$\left(\frac{\sum_{i=1}^{g} n_i - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(n-p-2)}$

Em que: p é o número de variáveis; g é o número de grupos; n é dado por $n_i (i = 1,2, \ldots, g)$, em que $n = \sum_{i=1}^{g} n_i$, sendo F o valor tabelado de F.

Quando se deseja transformar o valor de Wilks lambda para a distribuição de Qui-quadrado (para amostras em que n é grande), utiliza-se a equação:

$$-\left(n - 1 - \frac{p+g}{2}\right) ln\, \Lambda^* = -\left(n - 1 - \frac{p+g}{2}\right) \ln\left(\frac{|\mathrm{W}|}{|\mathrm{B+W}|}\right) \sim \chi^2_{p(g-1)}$$

As matrizes W e B são as definidas anteriormente.

Além do teste descrito, existem os testes estatísticos Maior Raiz de Roy, Pillai e Lawley-Hotelling. Os quatros testes são comparados em termos dos autovalores $\lambda_1 > \lambda_2 > \cdots > \lambda_s$ da matriz $W^{-1}B$, em que s é o valor mínimo entre graus de liberdade de tratamento e o número de variáveis. Os testes são calculados pelas seguintes equações:

Pillai: $V^s = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$

Lawley-Hotelling: $U^s = \sum_{i=1}^{s} \lambda_i$

Maior raiz de Roy: $\theta = \frac{\lambda_1}{1+\lambda_1}$

Lambda de Wilks: $\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$

6.5 Exemplo de MANOVA com variáveis energéticas

A seguir está apresentado um exemplo de MANOVA, na qual se comparou quatro idades de acácia-negra (grupos/tratamentos), cada idade com cinco variáveis energéticas usualmente utilizadas em análises de qualidade da biomassa para energia. O experimento foi conduzido na região de Piratini (RS), sendo os dados apresentados na Tabela 6.4.

Tabela 6.4: Resultados obtidos para quatro idades (anos) e cinco variáveis energéticas, com 4 repetições cada.

IDADE	REP	Y1	Y2	Y3	Y4	Y5
1	1	0,50	69,33	27,43	3,24	5113,53
1	2	0,53	69,01	27,81	3,18	5115,39
1	3	0,51	68,94	27,88	3,19	5043,65
1	4	0,50	69,07	27,61	3,32	5081,57
3	1	0,51	75,24	22,58	2,18	4708,60
3	2	0,51	73,24	24,28	2,49	4553,28
3	3	0,51	73,89	23,89	2,21	4479,64
3	4	0,48	73,93	23,75	2,33	4664,67
5	1	0,52	73,44	23,91	2,65	4754,10
5	2	0,54	72,15	24,97	2,88	4909,00
5	3	0,50	73,81	23,48	2,71	4704,58
5	4	0,49	73,80	23,71	2,49	4690,54
10	1	0,55	74,63	23,50	1,88	4620,83
10	2	0,61	73,63	24,13	2,24	4647,88
10	3	0,60	72,24	25,23	2,53	4657,54
10	4	0,59	72,69	24,87	2,45	4693,50

Em que: $Y1 =$ densidade básica (kg/m^3); $Y2 =$ teor de materiais voláteis (%); $Y3 =$ teor de carbono fixo (%); $Y4 =$ teor de cinzas (%); $Y5 =$ poder calorífico superior (kcal/g).

Uma análise de variância multivariada (MANOVA) foi efetuada com o objetivo de avaliar a existência de diferenças estatisticamente significantes entre os quatro tratamentos (idades), considerando simultaneamente cinco variáveis resposta, com quatro repetições por tratamento. A MANOVA testou a hipótese de igualdade dos vetores de médias multivariadas entre os tratamentos, isto é, assume-se inicialmente que os tratamentos não diferem quando as variáveis são analisadas em conjunto.

A avaliação multivariada foi conduzida por meio das estatísticas clássicas de Wilks, Pillai, Hotelling–Lawley e Roy. Na área florestal comumente é utilizado a estatística de Lambda de Wilks, a qual quantifica a proporção da variabilidade total não explicada pelos tratamentos, sendo que valores próximos de zero indicam maior diferença entre os grupos. No presente caso, o valor de Lambda de Wilks foi de 0,00149, o qual foi comparado com o valor de tabela de Qui-quadrado, ao nível de 1% de significância. O valor de Lambda de Wilks corrigido foi de 68,34 e o valor de tabela de Qui-quadrado foi de 30,58. Quando o valor calculado supera o valor crítico, rejeita-se a hipótese nula, indicando que os tratamentos apresentam diferenças multivariadas significantes,

considerando o conjunto das cinco variáveis avaliadas. As estatísticas obtidas também podem ser comparadas utilizando a distribuição de F de Fischer. Para tal é necessário converter os valores das estatísticas para os valores de F.

Diante da significância multivariada observada, recomenda-se a realização de análises complementares, como a análise de discriminante, com o objetivo de identificar quais variáveis e quais tratamentos são responsáveis pelas diferenças observadas entre os grupos.

6.6 Exemplo de MANOVA com variáveis de produção e de tecnologia da madeira

Em um experimento com acácia-negra foram testados três tipos de adubação, com quatro repetições cada, em delineamento de simples classificação. Foram medidas as variáveis biomassa do fuste (BF, em $t.ha^{-1}$), biomassa da copa (BC, em t/ha), volume (V, em m^3ha^{-1}), massa específica da copa (MEC, em $g.cm^{-3}$) e massa específica do fuste (MEF, em $g.cm^{-3}$) e os dados estão apresentados na Tabela 6.5. Deseja-se saber se existem diferenças entre os tipos de adubação para todas as variáveis avaliadas simultaneamente.

Tabela 6.5: Valores das variáveis mensuradas em um experimento com três tipos de adubação.

Tratamento	Rep	BF	BC	V	MEC	MEF
1	1	43,48	8,42	73,70	0,56	0,49
1	2	67,96	12,79	121,04	0,56	0,50
1	3	65,19	11,61	114,27	0,54	0,47
1	4	70,50	14,36	119,43	0,59	0,50
1	5	47,40	8,53	78,21	0,59	0,51
1	6	52,90	10,58	95,22	0,57	0,49
1	7	61,30	12,87	109,73	0,53	0,46
1	8	59,70	11,04	98,51	0,57	0,48
2	1	72,10	17,46	143,16	0,52	0,53
2	2	70,87	20,93	138,80	0,51	0,50
2	3	85,31	22,18	197,15	0,51	0,50
2	4	66,04	20,70	124,67	0,52	0,54
2	5	74,06	17,77	216,39	0,51	0,51
2	6	89,53	18,80	222,53	0,54	0,54
2	7	69,15	18,58	119,72	0,58	0,55
2	8	87,99	21,13	146,65	0,58	0,56
3	1	110,08	19,43	178,07	0,58	0,58
3	2	159,37	33,56	278,88	0,58	0,57
3	3	127,09	17,30	222,87	0,60	0,56
3	4	111,52	14,38	201,16	0,58	0,54
3	5	143,13	24,24	257,03	0,56	0,56
3	6	137,43	27,21	236,47	0,58	0,58
3	7	115,42	20,30	190,76	0,58	0,57
3	8	152,91	24,84	246,90	0,58	0,53

As hipóteses testadas são:

$$\begin{cases} H_0: \begin{bmatrix}\mu_{11}\\ \mu_{12}\\ \vdots\\ \mu_{1p}\end{bmatrix} = \begin{bmatrix}\mu_{21}\\ \mu_{22}\\ \vdots\\ \mu_{2p}\end{bmatrix} = \begin{bmatrix}\mu_{31}\\ \mu_{32}\\ \vdots\\ \mu_{3p}\end{bmatrix} \\ H_1: \begin{bmatrix}\mu_{11}\\ \mu_{12}\\ \vdots\\ \mu_{1p}\end{bmatrix} \neq \begin{bmatrix}\mu_{21}\\ \mu_{22}\\ \vdots\\ \mu_{2p}\end{bmatrix} \neq \begin{bmatrix}\mu_{31}\\ \mu_{32}\\ \vdots\\ \mu_{3p}\end{bmatrix} \end{cases} \qquad \begin{cases} H_o: \underline{\mu}_1 = \underline{\mu}_2 = \underline{\mu}_3 \\ H_1: \underline{\mu}_1 \neq \underline{\mu}_2 \neq \underline{\mu}_3 \end{cases}$$

A análise de variância multivariada revelou existirem diferenças significantes entre os tipos de adubação, ao nível de 5% de significância. A análise de discriminantes pode ser efetuada visando identificar a separação dos tratamentos testados, bem como a variável mais importante.

6.7 SOLUÇÃO DO EXEMPLO 6.5 DE ANÁLISE DE VARIÂNCIA MULTIVARIADA (MANOVA) NO MATLAB

MÉTODO 1: OBTENÇÃO DA MANOVA POR MEIO DO CÁLCULO DOS DESVIOS

```
% MATRIZ DAS VARIÁVEIS RESPOSTAS
X =
          0.5        69.33        27.43         3.24       5113.5
         0.53        69.01        27.81         3.18       5115.4
         0.51        68.94        27.88         3.19       5043.6
          0.5        69.07        27.61         3.32       5081.6
         0.51        75.24        22.58         2.18       4708.6
         0.51        73.24        24.28         2.49       4553.3
         0.51        73.89        23.89         2.21       4479.6
         0.48        73.93        23.75         2.33       4664.7
         0.52        73.44        23.91         2.65       4754.1
         0.54        72.15        24.97         2.88         4909
          0.5        73.81        23.48         2.71       4704.6
         0.49         73.8        23.71         2.49       4690.5
         0.55        74.63         23.5         1.88       4620.8
         0.61        73.63        24.13         2.24       4647.9
          0.6        72.24        25.23         2.53       4657.5
         0.59        72.69        24.87         2.45       4693.5

% MATRIZ DE DADOS PARA O GRUPO 1 (TRATAMENTO 1)
X1=X(1:4,:)
X1 =
          0.5        69.33        27.43         3.24       5113.5
         0.53        69.01        27.81         3.18       5115.4
         0.51        68.94        27.88         3.19       5043.6
          0.5        69.07        27.61         3.32       5081.6

% MATRIZ DE DADOS PARA O GRUPO 2 (TRATAMENTO 2)
X2=X(5:8,:)
X2 =
         0.51        75.24        22.58         2.18       4708.6
         0.51        73.24        24.28         2.49       4553.3
         0.51        73.89        23.89         2.21       4479.6
         0.48        73.93        23.75         2.33       4664.7
```

```
% MATRIZ DE DADOS PARA O GRUPO 3 (TRATAMENTO 3)
X3=X(9:12,:)
X3 =
         0.52        73.44        23.91         2.65       4754.1
         0.54        72.15        24.97         2.88         4909
          0.5        73.81        23.48         2.71       4704.6
         0.49         73.8        23.71         2.49       4690.5

% MATRIZ DE DADOS PARA O GRUPO 4 (TRATAMENTO 4)
X4=X(13:16,:)
X4 =
         0.55        74.63         23.5         1.88       4620.8
         0.61        73.63        24.13         2.24       4647.9
          0.6        72.24        25.23         2.53       4657.5
         0.59        72.69        24.87         2.45       4693.5

% TAMANHO DA AMOSTRA DO GRUPO 1 (n1)E NÚMERO DE VARIÁVEIS (p)
[n1,p]=size(X1)
n1 =
     4
p =
     5

% TAMANHO DA AMOSTRA DO GRUPO 2 (n2)E NÚMERO DE VARIÁVEIS (p)
[n2,p]=size(X2)
n2 =
     4
p =
     5

% TAMANHO DA AMOSTRA DO GRUPO 3 (n3)E NÚMERO DE VARIÁVEIS (p)
[n3,p]=size(X3)
n3 =
     4
p =
     5

 % TAMANHO DA AMOSTRA DO GRUPO 4 (n4)E NÚMERO DE VARIÁVEIS (p)
[n4,p]=size(X4)
n4 =
     4
p =
     5

% TAMANHO TOTAL DAS AMOTRAS
n=n1+n2+n3
n =
    12

% VETOR DO TAMANHO DAS AMOSTRAS DOS GRUPOS
ng=[n1;n2;n3]
ng =
     4
     4
     4
ng=[n1;n2;n3;n4]
```

```
ng =
     4
     4
     4
     4

% TAMANHO TOTAL DAS AMOTRAS
n=n1+n2+n3+n4
n =
    16

% VETOR DO TAMANHO DAS AMOSTRAS DOS GRUPOS
ng=[n1;n2;n3;n4]
ng =
     4
     4
     4
     4

% NÚMERO DE GRUPOS
g=size(ng,1)
g =
     4

% VETOR DE MÉDIAS PARA O GRUPO 1
MEDIAX1=mean(X1)
MEDIAX1 =
         0.51       69.088       27.682       3.2325       5088.5

% VETOR DE MÉDIAS PARA O GRUPO 2
MEDIAX2=mean(X2)
MEDIAX2 =
       0.5025       74.075       23.625       2.3025       4601.5

% VETOR DE MÉDIAS PARA O GRUPO 3
MEDIAX3=mean(X3)
MEDIAX3 =
       0.5125         73.3       24.017       2.6825       4764.6

% VETOR DE MÉDIAS PARA O GRUPO 4
MEDIAX4=mean(X4)
MEDIAX4 =
       0.5875       73.297       24.433        2.275       4654.9

% VETOR DE MÉDIAS DOS GRUPOS
MEDIAS=[MEDIAX1;MEDIAX2;MEDIAX3;MEDIAX4]
MEDIAS =
         0.51       69.088       27.682       3.2325       5088.5
       0.5025       74.075       23.625       2.3025       4601.5
       0.5125         73.3       24.017       2.6825       4764.6
       0.5875       73.297       24.433        2.275       4654.9

% VETOR DE MÉDIAS PARA GERAL
MEDIAGERAL=mean(X)
MEDIAGERAL =
      0.52812        72.44       24.939       2.6231       4777.4
```

```
% MATRIZ DE DIFERENÇAS DAS MÉDIAS DOS GRUPOS E A MÉDIA GERAL
DIFMEDIA=MEDIAS-MEDIAGERAL

DIFMEDIA =
    -0.018125      -3.3525       2.7431      0.60937       311.14
    -0.025625        1.635      -1.3144     -0.32063      -175.85
    -0.015625         0.86     -0.92188     0.059375      -12.839
     0.059375       0.8575     -0.50688     -0.34813      -122.46

% MATRIZ DE DIFERENÇAS DAS MÉDIAS DOS GRUPOS E A MÉDIA GERAL
CORRIGIDAS PELO TAMANHO DOS GRUPOS
DIFMEDIACOR=DIFMEDIA.*ng
DIFMEDIACOR =
      -0.0725       -13.41       10.972       2.4375       1244.6
      -0.1025         6.54      -5.2575      -1.2825      -703.38
      -0.0625         3.44      -3.6875       0.2375      -51.355
       0.2375         3.43      -2.0275      -1.3925      -489.82

% MATRIZ DE SOMA DOS QUADRADOS DOS GRUPOS
B=DIFMEDIACOR' * DIFMEDIA
B =
     0.019019      0.22537     -0.12692    -0.097706      -32.814
      0.22537        61.55      -50.291      -11.258      -5786.6
     -0.12692      -50.291       41.436       8.8589       4634.1
    -0.097706      -11.258       8.8589       2.3954       1151.4
      -32.814      -5786.6       4634.1       1151.4   5.7156e+05

% MATRIZ DE DIFERENÇAS DOS VALORES DE CADA GRUPO E A MÉDIA DE SEU
RESPECTIVO GRUPO
DIFX1=X1-MEDIAX1
DIFX1 =
        -0.01       0.2425      -0.2525       0.0075       24.995
         0.02      -0.0775       0.1275      -0.0525       26.855
            0      -0.1475       0.1975      -0.0425      -44.885
        -0.01      -0.0175      -0.0725       0.0875       -6.965
DIFX2=X2-MEDIAX2
DIFX2 =
       0.0075        1.165       -1.045      -0.1225       107.05
       0.0075       -0.835        0.655       0.1875      -48.268
       0.0075       -0.185        0.265      -0.0925      -121.91
      -0.0225       -0.145        0.125       0.0275       63.122
DIFX3=X3-MEDIAX3
DIFX3 =
       0.0075         0.14      -0.1075      -0.0325      -10.455
       0.0275        -1.15       0.9525       0.1975       144.44
      -0.0125         0.51      -0.5375       0.0275      -59.975
      -0.0225          0.5      -0.3075      -0.1925      -74.015
DIFX4=X4-MEDIAX4
DIFX4 =
      -0.0375       1.3325      -0.9325       -0.395      -34.108
       0.0225       0.3325      -0.3025       -0.035      -7.0575
       0.0125      -1.0575       0.7975        0.255       2.6025
       0.0025      -0.6075       0.4375        0.175       38.563
DIF=[DIFX1;DIFX2;DIFX3;DIFX4]
DIF =
        -0.01       0.2425      -0.2525       0.0075       24.995
         0.02      -0.0775       0.1275      -0.0525       26.855
            0      -0.1475       0.1975      -0.0425      -44.885
```

```
        -0.01      -0.0175      -0.0725       0.0875       -6.965
       0.0075        1.165       -1.045      -0.1225       107.05
       0.0075       -0.835        0.655       0.1875      -48.268
       0.0075       -0.185        0.265      -0.0925      -121.91
      -0.0225       -0.145        0.125       0.0275       63.122
       0.0075         0.14      -0.1075      -0.0325      -10.455
       0.0275        -1.15       0.9525       0.1975       144.44
      -0.0125         0.51      -0.5375       0.0275      -59.975
      -0.0225          0.5      -0.3075      -0.1925      -74.015
      -0.0375       1.3325      -0.9325       -0.395      -34.108
       0.0225       0.3325      -0.3025       -0.035      -7.0575
       0.0125      -1.0575       0.7975        0.255       2.6025
       0.0025      -0.6075       0.4375        0.175       38.563

% MATRIZ DE SOMAS DE QUADRADOS DOS RESÍDUOS
W=DIF' * DIF
W =
     0.004825     -0.10487       0.0803        0.024       6.0212
     -0.10487       7.4222       -5.921      -1.5034         -120
       0.0803       -5.921       4.8219       1.1013       67.443
        0.024      -1.5034       1.1013      0.40252       53.522
       6.0212         -120       67.443       53.522        68801

% CÁLCULO DO VALOR DE LAMBDA DE WILKS
L=det(W)/ det(B+W)
L =
    0.0014901

% COMPARAÇÃO DO VALOR DE LAMBDA DE WILKS CONTRA O VALOR CRÍTICO DE
% QUI-QUDRADO

% CÁLCULO DO VALOR DE LAMBDA DE WILKS CORRIGIDO

LCORRIGIDO = -(n-1-((p+g)/2))* log(L)
LCORRIGIDO =
       68.343

% VALOR CRÍTICO DE QUI-QUADRADO
QUI2=chi2inv(0.99,(p*(g-1)))
QUI2 =
       30.578

% CÁLCULO DA ESTATÍSTICA DE PILLAI
PILLAI=trace(B*(inv(B+W)))
PILLAI =
        2.321

% CÁLCULO DA ESTATÍSTICA DE HOTELLING
HOTELLING=trace(inv(W)*B)
HOTELLING =
       49.787

% CÁLCULO DA ESTATÍSTICA DE ROY
% CÁLCULO DOS AUTO VALORES DA MATRIZ INVERSA DE W * B
INVWB=inv(W)*B;
[VETOR,VALOR]=eig(INVWB);
[VALORAC,IND]=sort(diag(VALOR),'descend'
ROY=VALORAC(1,1)/(1+VALORAC(1,1))
```

```
ROY =
      0.97643
```

MÉTODO 2: MANOVA POR MEIO DE MODELOS DE REGRESSÃO LINEARES MÚLTIPLOS. ALTERNATIVA PARA A INVERSA DE X'X

```
% MATRIZ DAS VARIÁVEIS RESPOSTAS
Y=[0.50    69.33 27.43 3.24  5113.53
0.53  69.01 27.81 3.18  5115.39
0.51  68.94 27.88 3.19  5043.65
0.50  69.07 27.61 3.32  5081.57
0.51  75.24 22.58 2.18  4708.60
0.51  73.24 24.28 2.49  4553.28
0.51  73.89 23.89 2.21  4479.64
0.48  73.93 23.75 2.33  4664.67
0.52  73.44 23.91 2.65  4754.10
0.54  72.15 24.97 2.88  4909.00
0.50  73.81 23.48 2.71  4704.58
0.49  73.80 23.71 2.49  4690.54
0.55  74.63 23.50 1.88  4620.83
0.61  73.63 24.13 2.24  4647.88
0.60  72.24 25.23 2.53  4657.54
0.59  72.69 24.87 2.45  4693.50
]

% MATRIZ DE VARIÁVEIS CATEGÓRICAS PARA REPRESENTAR OS TRATAMENTOS
X=[1  1     0     0     0
1     1     0     0     0
1     1     0     0     0
1     1     0     0     0
1     0     1     0     0
1     0     1     0     0
1     0     1     0     0
1     0     1     0     0
1     0     0     1     0
1     0     0     1     0
1     0     0     1     0
1     0     0     1     0
1     0     0     0     1
1     0     0     0     1
1     0     0     0     1
1     0     0     0     1]

% SOMAS DE QUADRADOS E PRODUTOS NÃO CORRIGIDAS
S=X'*X
S =
    16     4     4     4     4
     4     4     0     0     0
     4     0     4     0     0
     4     0     0     4     0
     4     0     0     0     4

% DETERMINAÇÃO DO NÚMERO DE REPETIÇÕES
T=sum(X(:,2:end))
T =
     4     4     4     4
T=[0 T]
```

```
T =
     0     4     4     4     4
[m,n]=size(X)
m =
    16
n =
     5

% MATRIZ NULA
NULA=zeros((n-1),n)
NULA =
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0

% FATOR DE CORREÇÃO
A=[T;NULA]
A =
     0     4     4     4     4
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0

% CORREÇÃO DA MATRIZ S PARA ADMITIR INVERSA
M=S-A
M =
    16     0     0     0     0
     4     4     0     0     0
     4     0     4     0     0
     4     0     0     4     0
     4     0     0     0     4

% MATRIZ DOS COEFICIENTES
B=inv(M)*(X'*Y)
B =
     0.52812        72.44       24.939       2.6231       4777.4
    -0.018125      -3.3525       2.7431      0.60937       311.14
    -0.025625        1.635      -1.3144     -0.32063      -175.85
    -0.015625         0.86     -0.92188     0.059375      -12.839
     0.059375       0.8575     -0.50688     -0.34813      -122.46


% MATRIZ DE SOMAS DE QUADRADOS DOS RESÍDUOS CORRIGIDA
SQRES=(Y'*Y) - (B'*(X'*Y))
SQRES =
     0.004825     -0.10487       0.0803        0.024       6.0212
     -0.10487       7.4222      -5.9211      -1.5034         -120
       0.0803      -5.9211       4.8219       1.1013       67.443
        0.024      -1.5034       1.1013      0.40252       53.522
       6.0212         -120       67.443       53.522        68801

% MATRIZ DE SOMAS DE QUADRADOS DOS TRATAMENTOS NÃO CORRIGIDA
SQTRATNC=B' * (X'*Y)
SQTRATNC =
```

```
% MATRIZ DE SOMAS DE QUADRADOS TOTAIS NÃO CORRIGIDA
SQTOTNC=Y'*Y
SQTOTNC =
       4.4817       612.34       210.61       22.068        40336
       612.34        84022        28855         3029    5.5314e+06
       210.61        28855         9993       1055.6     1.911e+06
       22.068         3029       1055.6       112.49    2.0166e+05
        40336    5.5314e+06    1.911e+06    2.0166e+05    3.6575e+08

% SOMAS DE QUADRADOS CORRIGIDAS
% MATRIZ DE DESVIOS
UNS=ones(m);
a=Y-((UNS*Y).*(1/m))
a =
    -0.028125        -3.11       2.4906      0.61687       336.14
     0.001875        -3.43       2.8706      0.55687          338
    -0.018125         -3.5       2.9406      0.56687       266.26
    -0.028125        -3.37       2.6706      0.69687       304.18
    -0.018125          2.8      -2.3594     -0.44313      -68.794
    -0.018125          0.8     -0.65938     -0.13313      -224.11
    -0.018125         1.45      -1.0494     -0.41313      -297.75
    -0.048125         1.49      -1.1894     -0.29313      -112.72
    -0.008125            1      -1.0294     0.026875      -23.294
     0.011875        -0.29     0.030625      0.25687       131.61
    -0.028125         1.37      -1.4594     0.086875      -72.814
    -0.038125         1.36      -1.2294     -0.13313      -86.854
     0.021875         2.19      -1.4394     -0.74313      -156.56
     0.081875         1.19     -0.80938     -0.38313      -129.51
     0.071875         -0.2      0.29062    -0.093125      -119.85
     0.061875         0.25    -0.069375     -0.17313      -83.894

% MATRIZ DE SOMAS DE QUADRADOS TOTAIS CORRIGIDA
SQTOTCORR=a'*a
SQTOTCORR =
     0.023844       0.1205    -0.046619    -0.073706      -26.793
       0.1205        68.972      -56.212      -12.762      -5906.6
    -0.046619      -56.212       46.258       9.9602       4701.6
    -0.073706      -12.762       9.9602       2.7979       1204.9
      -26.793      -5906.6       4701.6       1204.9    6.4037e+05

% MATRIZ DE SOMAS DE QUADRADOS DOS TRATAMENTOS CORRIGIDA
SQTRATCORR=SQTOTCORR-SQRES
SQTRATCORR =
     0.019019      0.22537     -0.12692    -0.097706      -32.814
      0.22537        61.55      -50.291      -11.258      -5786.6
     -0.12692      -50.291       41.436       8.8589       4634.1
    -0.097706      -11.258       8.8589       2.3954       1151.4
      -32.814      -5786.6       4634.1       1151.4    5.7156e+05

% TESTES ESTATÍTICOS
B=SQTRATCORR;
W=SQRES;
% CÁLCULO DE LAMBDA DE WILKS
L=det(W)/ det(B+W)
L=
      0.00149
```

```
% CÁLCULO DA ESTATÍSTICA DE PILLAI
PILLAI=trace(B*(inv(B+W)))
PILLAI =
      2.32101

% CÁLCULO DA ESTATÍSTICA DE HOTELLING
HOTELLING=trace(inv(W)*B)
HOTELLING =
      49.78697

% CÁLCULO DA ESTATÍSTICA DE ROY

% CÁLCULO DOS AUTO VALORES DA MATRIZ INVERSA DE W * B

INVWB=inv(W)*B;
[VETOR,VALOR]=eig(INVWB);
[VALORAC,IND]=sort(diag(VALOR),'descend');
ROY=VALORAC(1,1)/(1+VALORAC(1,1))
ROY=
      0.976443
```

6.8 SOLUÇÃO DO EXEMPLO 6.5 DE ANÁLISE DE VARIÂNCIA MULTIVARIADA (MANOVA) NO JULIA

MÉTODO 1: OBTENÇÃO DA MANOVA POR MEIO DO CÁLCULO DOS DESVIOS

```
using DataFrames, CSV, Statistics, LinearAlgebra

#Matriz de dados geral
X=[0.50 69.33   27.43   3.24    5113.53
0.53    69.01   27.81   3.18    5115.39
0.51    68.94   27.88   3.19    5043.65
0.50    69.07   27.61   3.32    5081.57
0.51    75.24   22.58   2.18    4708.60
0.51    73.24   24.28   2.49    4553.28
0.51    73.89   23.89   2.21    4479.64
0.48    73.93   23.75   2.33    4664.67
0.52    73.44   23.91   2.65    4754.10
0.54    72.15   24.97   2.88    4909.00
0.50    73.81   23.48   2.71    4704.58
0.49    73.80   23.71   2.49    4690.54
0.55    74.63   23.50   1.88    4620.83
0.61    73.63   24.13   2.24    4647.88
0.60    72.24   25.23   2.53    4657.54
0.59    72.69   24.87   2.45    4693.50]

#Matriz de dados para o grupo 1 (Tratamento 1)
X1=X[1:4,:]

#Matriz de dados para o grupo 2 (Tratamento 2)
X2=X[5:8,:]
```

```
#Matriz de dados para o grupo 3 (Tratamento 3)
X3=X[9:12,:]

#Matriz de dados para o grupo 4 (Tratamento 4)
X4=X[13:16,:]

#Tamanho da amostra do grupo 1 (n1) e número de variáveis (p)
(n1,p)=size(X1)

#Tamanho da amostra do grupo 2 (n2) e número de variáveis (p)
(n2,p)=size(X2)

#Tamanho da amostra do grupo 3 (n3) e número de variáveis (p)
(n3,p)=size(X3)

#Tamanho da amostra do grupo 4 (n4) e número de variáveis (p)
(n4,p)=size(X4)

#Tamanho total das amostras
n=n1+n2+n3+n4

#Vetor do tamanho das amostras dos grupos
ng=[n1;n2;n3;n4]

#Número de grupos
g=size(ng,1)

#Vetor de médias para o grupo 1
MEDIAX1=mean(X1,dims=1)

#Vetor de médias para o grupo 2
MEDIAX2=mean(X2,dims=1)

#Vetor de médias para o grupo 3
MEDIAX3=mean(X3,dims=1)

#Vetor de médias para o grupo 4
MEDIAX4=mean(X4,dims=1)

#Vetor de médias dos grupos
MEDIAS=[MEDIAX1;MEDIAX2;MEDIAX3;MEDIAX4]
#Vetor de médias geral
MEDIAGERAL=mean(X,dims=1)

#Matriz de diferenças das médias dos grupos e a média geral
DIFMEDIA=MEDIAS.-MEDIAGERAL

#Matriz de diferenças das médias dos grupos e a média geral corrigida pelo
tamanho dos grupos
DIFMEDIACOR=DIFMEDIA.*ng
```

```
#Matriz de somas de quadrados dos grupos
B=DIFMEDIACOR'*DIFMEDIA

#Matriz de diferenças dos valores de cada grupo e a média de seu respectivo
grupo
DIFX1=X1.-MEDIAX1
DIFX2=X2.-MEDIAX2
DIFX3=X3.-MEDIAX3
DIFX4=X4.-MEDIAX4
DIF=[DIFX1;DIFX2;DIFX3;DIFX4]

#Matriz de somas de quadrados dos resíduos
W=DIF'*DIF

#Cálculo do valor de Lambda de Wilks
L=det(W)/det(B+W)

#Cálculo da estatística de Pillai
Pillai=tr(B*inv(B+W))

#Cálculo da estatística de Hotteling
Hotteling=tr(inv(W)*B)

#Cálculo da estatística de Roy
#Autovalores da matriz inversa de W * B
INVWB=inv(W)*B
AV = eigen(INVWB)

#Autovalores ordenados
Valor = reverse(AV.values, dims=1)

ROY=Valor[1,1]/(1+Valor[1,1])

#Comparação do valor de Lambda de Wilks contra o valor crítico de Qui-quadrado
#Cálculo do valor de Lambda de Wilks corrigido
Lcorrigido=-(n-1-((p+g)/2))*log(L)

#Valor crítico de Qui-Quadrado
using Distributions
QUIQUADRAINV = quantile(Chisq(p * (g - 1)), 0.99)
```

MÉTODO 2: MANOVA POR MEIO DE MODELOS DE REGRESSÃO LINEARES MÚLTIPLOS. ALTERNATIVA PARA A INVERSA DE X'X

```
using DataFrames, CSV, Statistics, LinearAlgebra

#Matriz das variáveis respostas
Y=[
0.50    69.33   27.43   3.24    5113.53
0.53    69.01   27.81   3.18    5115.39
0.51    68.94   27.88   3.19    5043.65
```

```
0.50    69.07   27.61   3.32    5081.57
0.51    75.24   22.58   2.18    4708.60
0.51    73.24   24.28   2.49    4553.28
0.51    73.89   23.89   2.21    4479.64
0.48    73.93   23.75   2.33    4664.67
0.52    73.44   23.91   2.65    4754.10
0.54    72.15   24.97   2.88    4909.00
0.50    73.81   23.48   2.71    4704.58
0.49    73.80   23.71   2.49    4690.54
0.55    74.63   23.50   1.88    4620.83
0.61    73.63   24.13   2.24    4647.88
0.60    72.24   25.23   2.53    4657.54
0.59    72.69   24.87   2.45    4693.50]

#Matriz de variáveis categóricas para representar os tratamentos
X=[
1   1   0   0   0
1   1   0   0   0
1   1   0   0   0
1   1   0   0   0
1   0   1   0   0
1   0   1   0   0
1   0   1   0   0
1   0   1   0   0
1   0   0   1   0
1   0   0   1   0
1   0   0   1   0
1   0   0   1   0
1   0   0   0   1
1   0   0   0   1
1   0   0   0   1
1   0   0   0   1]

#Somas de quadrados e produtos não corrigidas
S=X'X

#Determinação do número de repetições
T=sum(X[:,2:end], dims=1)
T=[0 T]
(m,n)=size(X)

#Matriz nula
Nula=zeros((n-1),n)

#Fator de correção
A=[
T
Nula]

#Correção da matriz S para admitir inversa
M=S-A
```

```
#Matriz dos coeficientes
B=inv(M)X'Y

#Matriz de somas de quadrados dos resíduos corrigida
SQRes=Y'Y-B'X'Y

#Matriz de somas de quadrados dos tratamentos não corrigida
SQTrat_NC=B'X'Y

#Matriz de soma de quadrados totais não corrigida
SQTot_NC=Y'Y

#Somas de quadrados corrigidas
#Matriz de desvios
a=Y-((ones(m)*ones(m)')*Y).*(1/m)

#Matriz de somas de quadrados totais corrigida
SQTot_Corr=a'*a

#Matriz de somas de quadrados dos tratamentos corrigida
SQTrat_Corr=SQTot_Corr-SQRes

#Testes estatísticos
B=SQTrat_Corr
W=SQRes

#Cálculo de Lambda de Wilks
L=det(W)/det(B+W)

#Cálculo da estatística de Pillai
Pillai=tr(B*inv(B+W))

#Cálculo da estatística de Hotteling
Hotteling=tr(inv(W)*B)

#Cálculo da estatística de Roy
#Autovalores da matriz inversa de W * B
INVWB=inv(W)*B
AV = eigen(INVWB)
#Autovalores ordenados
Valor = reverse(AV.values, dims=1)

ROY=Valor[1,1]/(1+Valor[1,1])
```

6. 9 SOLUÇÃO DO EXEMPLO 6.6, MANOVA PELO MÉTODO DE REGRESSÃO LINEAR, NO MATLAB E JULIA. USO DO MÉTODO DE MOORE PENROSE PARA A OBTENÇÃO DA INVERSA DE X'X.

```
% SOLUÇÃO DO EXEMPLO 6.6 NO MATLAB
% MATRIZ DE DADOS
% A PRIMEIRA COLUNA DA MATRRIZ DEVE CONTER OS TRATAMENTOS

DADOS =
    1.0000   43.4800    8.4200   73.7000    0.5600    0.4900
    1.0000   67.9600   12.7900  121.0400    0.5600    0.5000
    1.0000   65.1900   11.6100  114.2700    0.5400    0.4700
    1.0000   70.5000   14.3600  119.4300    0.5900    0.5000
    1.0000   47.4000    8.5300   78.2100    0.5900    0.5100
    1.0000   52.9000   10.5800   95.2200    0.5700    0.4900
    1.0000   61.3000   12.8700  109.7300    0.5300    0.4600
    1.0000   59.7000   11.0400   98.5100    0.5700    0.4800
    2.0000   72.1000   17.4600  143.1600    0.5200    0.5300
    2.0000   70.8700   20.9300  138.8000    0.5100    0.5000
    2.0000   85.3100   22.1800  197.1500    0.5100    0.5000
    2.0000   66.0400   20.7000  124.6700    0.5200    0.5400
    2.0000   74.0600   17.7700  216.3900    0.5100    0.5100
    2.0000   89.5300   18.8000  222.5300    0.5400    0.5400
    2.0000   69.1500   18.5800  119.7200    0.5800    0.5500
    2.0000   87.9900   21.1300  146.6500    0.5800    0.5600
    3.0000  110.0800   19.4300  178.0700    0.5800    0.5800
    3.0000  159.3700   33.5600  278.8800    0.5800    0.5700
    3.0000  127.0900   17.3000  222.8700    0.6000    0.5600
    3.0000  111.5200   14.3800  201.1600    0.5800    0.5400
    3.0000  143.1300   24.2400  257.0300    0.5600    0.5600
    3.0000  137.4300   27.2100  236.4700    0.5800    0.5800
    3.0000  115.4200   20.3000  190.7600    0.5800    0.5700
    3.0000  152.9100   24.8400  246.9000    0.5800    0.5300

% MATRIZ COM OS VALORES DAS VARIÁVEIS RESPOSTAS
Y=DADOS(:,2:end)
Y =
   43.4800    8.4200   73.7000    0.5600    0.4900
   67.9600   12.7900  121.0400    0.5600    0.5000
   65.1900   11.6100  114.2700    0.5400    0.4700
   70.5000   14.3600  119.4300    0.5900    0.5000
   47.4000    8.5300   78.2100    0.5900    0.5100
   52.9000   10.5800   95.2200    0.5700    0.4900
   61.3000   12.8700  109.7300    0.5300    0.4600
   59.7000   11.0400   98.5100    0.5700    0.4800
   72.1000   17.4600  143.1600    0.5200    0.5300
   70.8700   20.9300  138.8000    0.5100    0.5000
   85.3100   22.1800  197.1500    0.5100    0.5000
   66.0400   20.7000  124.6700    0.5200    0.5400
   74.0600   17.7700  216.3900    0.5100    0.5100
   89.5300   18.8000  222.5300    0.5400    0.5400
   69.1500   18.5800  119.7200    0.5800    0.5500
   87.9900   21.1300  146.6500    0.5800    0.5600
  110.0800   19.4300  178.0700    0.5800    0.5800
  159.3700   33.5600  278.8800    0.5800    0.5700
  127.0900   17.3000  222.8700    0.6000    0.5600
  111.5200   14.3800  201.1600    0.5800    0.5400
  143.1300   24.2400  257.0300    0.5600    0.5600
```

```
  137.4300   27.2100  236.4700    0.5800    0.5800
  115.4200   20.3000  190.7600    0.5800    0.5700
  152.9100   24.8400  246.9000    0.5800    0.5300

% NÚMERO DE OBSERVAÇÕES E VARIÁVEIS
[m,p]=size(Y)
m =
    24
p =
     5

% NÚMERO DE REPETIÇÕES POR TRATAMENTOS
[REP,TRAT]=hist(DADOS(:,1),unique(DADOS(:,1)))
REP =
     8     8     8
TRAT =
     1
     2
     3

% NÚMERO DE GRUPOS
g=size(TRAT,1)
g =
     3

% MATRIZ DE VARIÁVEIS CATEGÓRICAS REPRESENTANDO OS GRUPOS
(TRAMENTOS)
X=[ones(m,1) dummyvar(DADOS(:,1))]
X =
     1     1     0     0
     1     1     0     0
     1     1     0     0
     1     1     0     0
     1     1     0     0
     1     1     0     0
     1     1     0     0
     1     1     0     0
     1     0     1     0
     1     0     1     0
     1     0     1     0
     1     0     1     0
     1     0     1     0
     1     0     1     0
     1     0     1     0
     1     0     1     0
     1     0     0     1
     1     0     0     1
     1     0     0     1
     1     0     0     1
     1     0     0     1
     1     0     0     1
     1     0     0     1
     1     0     0     1

% SOMAS DE QUADRADOS E PRODUTOS NÃO CORRIGIDAS
S=X'*X
```

```
S =
    24     8     8     8
     8     8     0     0
     8     0     8     0
     8     0     0     8

% MATRIZ DOS COEFICIENTES/INVERSA DE X'X OBTIDA PELO MÉTODO DE
MOORE PENROSE
B=pinv(X'*X)*(X'*Y)
B =
   66.8884   13.4066  122.8538    0.4194    0.3944
   -8.3347   -2.1316  -21.5900    0.1444    0.0931
    9.9928    6.2872   40.7800    0.1144    0.1344
   65.2303    9.2509  103.6637    0.1606    0.1669

% MATRIZ DE SOMAS DE QUADRADOS DOS RESÍDUOS CORRIGIDA
SQRES=(Y'*Y) - (B'*(X'*Y))
SQRES =
   1.0e+04 *
    0.3805    0.0859    0.7360   -0.0000   -0.0000
    0.0859    0.0312    0.1359   -0.0000    0.0000
    0.7360    0.1359    2.2913   -0.0005   -0.0004
   -0.0000   -0.0000   -0.0005    0.0000    0.0000
   -0.0000    0.0000   -0.0004    0.0000    0.0000

% MATRIZ DE SOMAS DE QUADRADOS DOS TRATAMENTOS NÃO CORRIGIDA
SQTRATNC=B' * (X'*Y)
SQTRATNC =
   1.0e+05 *
    2.1436    0.4134    3.8750    0.0121    0.0115
    0.4134    0.0823    0.7597    0.0024    0.0023
    3.8750    0.7597    7.0672    0.0221    0.0210
    0.0121    0.0024    0.0221    0.0001    0.0001
    0.0115    0.0023    0.0210    0.0001    0.0001

% MATRIZ DE SOMAS DE QUADRADOS TOTAIS NÃO CORRIGIDA
SQTOTNC=Y'*Y
SQTOTNC =
   1.0e+05 *
    2.1816    0.4220    3.9486    0.0121    0.0115
    0.4220    0.0854    0.7733    0.0024    0.0023
    3.9486    0.7733    7.2964    0.0220    0.0210
    0.0121    0.0024    0.0220    0.0001    0.0001
    0.0115    0.0023    0.0210    0.0001    0.0001

% SOMAS DE QUADRADOS CORRIGIDAS
% MATRIZ DE DESVIOS
a=Y-(ones(m)*Y).*(1/m)
a =
  -45.7046   -9.4554  -90.1050    0.0008   -0.0358
  -21.2246   -5.0854  -42.7650    0.0008   -0.0258
  -23.9946   -6.2654  -49.5350   -0.0192   -0.0558
  -18.6846   -3.5154  -44.3750    0.0308   -0.0258
  -41.7846   -9.3454  -85.5950    0.0308   -0.0158
  -36.2846   -7.2954  -68.5850    0.0108   -0.0358
  -27.8846   -5.0054  -54.0750   -0.0292   -0.0658
  -29.4846   -6.8354  -65.2950    0.0108   -0.0458
  -17.0846   -0.4154  -20.6450   -0.0392    0.0042
```

```
  -18.3146    3.0546  -25.0050   -0.0492   -0.0258
   -3.8746    4.3046   33.3450   -0.0492   -0.0258
  -23.1446    2.8246  -39.1350   -0.0392    0.0142
  -15.1246   -0.1054   52.5850   -0.0492   -0.0158
    0.3454    0.9246   58.7250   -0.0192    0.0142
  -20.0346    0.7046  -44.0850    0.0208    0.0242
   -1.1946    3.2546  -17.1550    0.0208    0.0342
   20.8954    1.5546   14.2650    0.0208    0.0542
   70.1854   15.6846  115.0750    0.0208    0.0442
   37.9054   -0.5754   59.0650    0.0408    0.0342
   22.3354   -3.4954   37.3550    0.0208    0.0142
   53.9454    6.3646   93.2250    0.0008    0.0342
   48.2454    9.3346   72.6650    0.0208    0.0542
   26.2354    2.4246   26.9550    0.0208    0.0442
   63.7254    6.9646   83.0950    0.0208    0.0042

% MATRIZ DE SOMAS DE QUADRADOS TOTAIS CORRIGIDA
SQTOTCORR=a'*a
SQTOTCORR =
   1.0e+04 *
    2.7269    0.3940    4.4243    0.0008    0.0021
    0.3940    0.0870    0.7058   -0.0000    0.0004
    4.4243    0.7058    8.5667    0.0004    0.0033
    0.0008   -0.0000    0.0004    0.0000    0.0000
    0.0021    0.0004    0.0033    0.0000    0.0000

% MATRIZ DE SOMAS DE QUADRADOS DOS TRATAMENTOS CORRIGIDA
% SQTRATCORR=SQTOTCORR-SQRES
SQTRATCORR =
   1.0e+04 *
    2.3464    0.3081    3.6882    0.0009    0.0021
    0.3081    0.0558    0.5699    0.0000    0.0003
    3.6882    0.5699    6.2754    0.0008    0.0037
    0.0009    0.0000    0.0008    0.0000    0.0000
    0.0021    0.0003    0.0037    0.0000    0.0000

% CÁLCULO DE LAMBDA DE WILKS
L=det(SQRES)/det(SQTRATCORR+SQRES)
L =
    0.0110

% CÁLCULO DO VALOR DE LAMDA DE WILKS CORRIGIDO
LCORRIGIDO = -(m-1-((p+g)/2))*log(L)

LCORRIGIDO =
   85.7637

% VALOR CRÍTICO DE CHI-QUADRADO
QUI2=chi2inv(0.99,(p*(g-1)))
QUI2 =
   23.2093

# SOLUÇÃO DO EXEMPLO 6.6 NO JULIA
#Matriz das variáveis respostas
Y=[
    43.48   8.42    73.70   0.56    0.49
    67.96   12.79   121.04  0.56    0.50
    65.19   11.61   114.27  0.54    0.47
```

```
    70.50   14.36   119.43  0.59    0.50
    47.40   8.53    78.21   0.59    0.51
    52.90   10.58   95.22   0.57    0.49
    61.30   12.87   109.73  0.53    0.46
    59.70   11.04   98.51   0.57    0.48
    72.10   17.46   143.16  0.52    0.53
    70.87   20.93   138.80  0.51    0.50
    85.31   22.18   197.15  0.51    0.50
    66.04   20.70   124.67  0.52    0.54
    74.06   17.77   216.39  0.51    0.51
    89.53   18.80   222.53  0.54    0.54
    69.15   18.58   119.72  0.58    0.55
    87.99   21.13   146.65  0.58    0.56
    110.08  19.43   178.07  0.58    0.58
    159.37  33.56   278.88  0.58    0.57
    127.09  17.30   222.87  0.60    0.56
    111.52  14.38   201.16  0.58    0.54
    143.13  24.24   257.03  0.56    0.56
    137.43  27.21   236.47  0.58    0.58
    115.42  20.30   190.76  0.58    0.57
    152.91  24.84   246.90  0.58    0.53]

#Matriz de variáveis categóricas para representar os grupos (tratamentos)
X=[
    1   1   0   0
    1   1   0   0
    1   1   0   0
    1   1   0   0
    1   1   0   0
    1   1   0   0
    1   1   0   0
    1   1   0   0
    1   0   1   0
    1   0   1   0
    1   0   1   0
    1   0   1   0
    1   0   1   0
    1   0   1   0
    1   0   1   0
    1   0   0   1
    1   0   0   1
    1   0   0   1
    1   0   0   1
    1   0   0   1
    1   0   0   1
    1   0   0   1
    1   0   0   1]

#Determinação do tamanho da matriz X
(m,n)=size(X)
```

```
#Somas de quadrados e produtos não corrigidas
S=X'X

#Matriz dos coeficientes / inversa de X'X obtida pelo método de moore penrose
B=pinv(X'X)X'Y

#Matriz de somas de quadrados dos resíduos corrigida
SQRes=Y'Y-B'X'Y

#Matriz de somas de quadrados dos tratamentos não corrigida
SQTrat_NC=B'X'Y

#Matriz de soma de quadrados totais não corrigida
SQTot_NC=Y'Y

#Somas de quadrados corrigidas
#Matriz de desvios
a=Y-((ones(m)*ones(m)')*Y).*(1/m)

#Matriz de somas de quadrados totais corrigida
SQTot_Corr=a'*a

#Matriz de somas de quadrados dos tratamentos corrigida
SQTrat_Corr=SQTot_Corr-SQRes

#Cálculo de Lambda de Wilks
L=det(SQRes)/det(SQTrat_Corr+SQRes)

#Número de variáveis
p=size(Y,2)

#Número de grupos
g=n-1

#Cálculo do valor de Lambda de Wilks corrigido
Lcorrigido=-(m-1-((p+g)/2))*log(L)

#Valor crítico de Qui-Quadrado
QUIQUADRAINV = quantile(Chisq(p * (g - 1)), 0.99)
```

6. 10 SOLUÇÃO DO EXEMPLO 6.6, MANOVA PELO MÉTODO DOS DESVIOS, NO MATLAB E JULIA

```
% SOLUÇÃO NO MATLAB
% MATRIZ DE DADOS GERAL (VARIÁVEIS RESPOSTAS)

X =
   43.4800    8.4200   73.7000    0.5600    0.4900
   67.9600   12.7900  121.0400    0.5600    0.5000
   65.1900   11.6100  114.2700    0.5400    0.4700
   70.5000   14.3600  119.4300    0.5900    0.5000
```

```
   47.4000    8.5300   78.2100    0.5900    0.5100
   52.9000   10.5800   95.2200    0.5700    0.4900
   61.3000   12.8700  109.7300    0.5300    0.4600
   59.7000   11.0400   98.5100    0.5700    0.4800
   72.1000   17.4600  143.1600    0.5200    0.5300
   70.8700   20.9300  138.8000    0.5100    0.5000
   85.3100   22.1800  197.1500    0.5100    0.5000
   66.0400   20.7000  124.6700    0.5200    0.5400
   74.0600   17.7700  216.3900    0.5100    0.5100
   89.5300   18.8000  222.5300    0.5400    0.5400
   69.1500   18.5800  119.7200    0.5800    0.5500
   87.9900   21.1300  146.6500    0.5800    0.5600
  110.0800   19.4300  178.0700    0.5800    0.5800
  159.3700   33.5600  278.8800    0.5800    0.5700
  127.0900   17.3000  222.8700    0.6000    0.5600
  111.5200   14.3800  201.1600    0.5800    0.5400
  143.1300   24.2400  257.0300    0.5600    0.5600
  137.4300   27.2100  236.4700    0.5800    0.5800
  115.4200   20.3000  190.7600    0.5800    0.5700
  152.9100   24.8400  246.9000    0.5800    0.5300

% MATRIZ DE DADOS PARA O GRUPO 1 (TRATAMENTO 1)
X1=X(1:8,:)
X1 =
   43.4800    8.4200   73.7000    0.5600    0.4900
   67.9600   12.7900  121.0400    0.5600    0.5000
   65.1900   11.6100  114.2700    0.5400    0.4700
   70.5000   14.3600  119.4300    0.5900    0.5000
   47.4000    8.5300   78.2100    0.5900    0.5100
   52.9000   10.5800   95.2200    0.5700    0.4900
   61.3000   12.8700  109.7300    0.5300    0.4600
   59.7000   11.0400   98.5100    0.5700    0.4800

% MATRIZ DE DADOS PARA O GRUPO 2 (TRATAMENTO 2)
X2=X(9:16,:)
X2 =
   72.1000   17.4600  143.1600    0.5200    0.5300
   70.8700   20.9300  138.8000    0.5100    0.5000
   85.3100   22.1800  197.1500    0.5100    0.5000
   66.0400   20.7000  124.6700    0.5200    0.5400
   74.0600   17.7700  216.3900    0.5100    0.5100
   89.5300   18.8000  222.5300    0.5400    0.5400
   69.1500   18.5800  119.7200    0.5800    0.5500
   87.9900   21.1300  146.6500    0.5800    0.5600

% MATRIZ DE DADOS PARA O GRUPO 3 (TRATAMENTO 3)
X3=X(17:24,:)
X3 =
  110.0800   19.4300  178.0700    0.5800    0.5800
  159.3700   33.5600  278.8800    0.5800    0.5700
  127.0900   17.3000  222.8700    0.6000    0.5600
  111.5200   14.3800  201.1600    0.5800    0.5400
  143.1300   24.2400  257.0300    0.5600    0.5600
  137.4300   27.2100  236.4700    0.5800    0.5800
  115.4200   20.3000  190.7600    0.5800    0.5700
  152.9100   24.8400  246.9000    0.5800    0.5300
```

```
% TAMANHO DA AMOSTRA DO GRUPO 1 (n1)E NÚMERO DE VARIÁVEIS (p)
[n1,p]=size(X1)
n1 =
     8
p =
     5

% TAMANHO DA AMOSTRA DO GRUPO 2 (n2)E NÚMERO DE VARIÁVEIS (p)
[n2,p]=size(X2)
n2 =
     8
p =
     5

% TAMANHO DA AMOSTRA DO GRUPO 3 (n3)E NÚMERO DE VARIÁVEIS (p)
[n3,p]=size(X3)
n3 =
     8
p =
     5

% TAMANHO TOTAL DAS AMOTRAS
n=n1+n2+n3
n =
    24

% VETOR DO TAMANHO DAS AMOSTRAS DOS GRUPOS
ng=[n1;n2;n3]
ng =
     8
     8
     8

% NÚMERO DE GRUPOS
g=size(ng,1)
g =
     3

% VETOR DE MÉDIAS PARA O GRUPO 1
MEDIAX1=mean(X1)
MEDIAX1 =
   58.5537   11.2750  101.2638    0.5637    0.4875

% VETOR DE MÉDIAS PARA O GRUPO 2
MEDIAX2=mean(X2)
MEDIAX2 =
   76.8812   19.6937  163.6338    0.5338    0.5288

% VETOR DE MÉDIAS PARA O GRUPO 3
MEDIAX3=mean(X3)
MEDIAX3 =
  132.1187   22.6575  226.5175    0.5800    0.5613
% VETOR DE MÉDIAS DOS GRUPOS
MEDIAS=[MEDIAX1;MEDIAX2;MEDIAX3]
MEDIAS =
   58.5537   11.2750  101.2638    0.5637    0.4875
   76.8812   19.6937  163.6338    0.5338    0.5288
  132.1187   22.6575  226.5175    0.5800    0.5613
```

```
% VETOR DE MÉDIAS PARA GERAL
MEDIAGERAL=mean(X)
MEDIAGERAL =
   89.1846   17.8754  163.8050    0.5592    0.5258

% MATRIZ DE DIFERENÇAS DAS MÉDIAS DOS GRUPOS E A MÉDIA GERAL
DIFMEDIA=MEDIAS-MEDIAGERAL
DIFMEDIA =
  -30.6308   -6.6004  -62.5413    0.0046   -0.0383
  -12.3033    1.8183   -0.1712   -0.0254    0.0029
   42.9342    4.7821   62.7125    0.0208    0.0354

% MATRIZ DE DIFERENÇAS DAS MÉDIAS DOS GRUPOS E A MÉDIA GERAL
CORRIGIDAS PELO TAMANHO DOS GRUPOS
DIFMEDIACOR=DIFMEDIA.*ng
DIFMEDIACOR =
 -245.0467  -52.8033 -500.3300    0.0367   -0.3067
  -98.4267   14.5467   -1.3700   -0.2033    0.0233
  343.4733   38.2567  501.7000    0.1667    0.2833

% MATRIZ DE SOMA DOS QUADRADOS DOS GRUPOS
B=DIFMEDIACOR' * DIFMEDIA
B =
   1.0e+04 *
    2.3464    0.3081    3.6882    0.0009    0.0021
    0.3081    0.0558    0.5699    0.0000    0.0003
    3.6882    0.5699    6.2754    0.0008    0.0037
    0.0009    0.0000    0.0008    0.0000    0.0000
    0.0021    0.0003    0.0037    0.0000    0.0000

% MATRIZ DE DIFERENÇAS DOS VALORES DE CADA GRUPO E A MÉDIA DE SEU
RESPECTIVO GRUPO
DIFX1=X1-MEDIAX1
DIFX1 =
  -15.0737   -2.8550  -27.5637   -0.0037    0.0025
    9.4063    1.5150   19.7763   -0.0037    0.0125
    6.6363    0.3350   13.0062   -0.0237   -0.0175
   11.9463    3.0850   18.1663    0.0262    0.0125
  -11.1537   -2.7450  -23.0538    0.0262    0.0225
   -5.6537   -0.6950   -6.0438    0.0062    0.0025
    2.7463    1.5950    8.4663   -0.0337   -0.0275
    1.1463   -0.2350   -2.7537    0.0062   -0.0075
DIFX2=X2-MEDIAX2
DIFX2 =
   -4.7813   -2.2337  -20.4738   -0.0138    0.0012
   -6.0112    1.2363  -24.8338   -0.0238   -0.0288
    8.4288    2.4863   33.5162   -0.0238   -0.0288
  -10.8412    1.0063  -38.9638   -0.0138    0.0112
   -2.8212   -1.9237   52.7562   -0.0238   -0.0188
   12.6488   -0.8937   58.8962    0.0062    0.0112
   -7.7312   -1.1137  -43.9138    0.0462    0.0212
   11.1088    1.4363  -16.9838    0.0462    0.0313
DIFX3=X3-MEDIAX3
DIFX3 =
  -22.0387   -3.2275  -48.4475         0    0.0187
   27.2513   10.9025   52.3625         0    0.0087
   -5.0287   -5.3575   -3.6475    0.0200   -0.0012
  -20.5987   -8.2775  -25.3575         0   -0.0212
```

```
   11.0113    1.5825   30.5125   -0.0200   -0.0012
    5.3113    4.5525    9.9525         0    0.0187
  -16.6987   -2.3575  -35.7575         0    0.0087
   20.7913    2.1825   20.3825         0   -0.0313
DIF=[DIFX1;DIFX2;DIFX3]
DIF =
  -15.0737   -2.8550  -27.5637   -0.0037    0.0025
    9.4063    1.5150   19.7763   -0.0037    0.0125
    6.6363    0.3350   13.0062   -0.0237   -0.0175
   11.9463    3.0850   18.1663    0.0262    0.0125
  -11.1537   -2.7450  -23.0538    0.0262    0.0225
   -5.6537   -0.6950   -6.0438    0.0062    0.0025
    2.7463    1.5950    8.4663   -0.0337   -0.0275
    1.1463   -0.2350   -2.7537    0.0062   -0.0075
   -4.7813   -2.2337  -20.4738   -0.0138    0.0012
   -6.0112    1.2363  -24.8338   -0.0238   -0.0288
    8.4288    2.4863   33.5162   -0.0238   -0.0288
  -10.8412    1.0063  -38.9638   -0.0138    0.0112
   -2.8212   -1.9237   52.7562   -0.0238   -0.0188
   12.6488   -0.8937   58.8962    0.0062    0.0112
   -7.7312   -1.1137  -43.9138    0.0462    0.0212
   11.1088    1.4363  -16.9838    0.0462    0.0313
  -22.0387   -3.2275  -48.4475         0    0.0187
   27.2513   10.9025   52.3625         0    0.0087
   -5.0287   -5.3575   -3.6475    0.0200   -0.0012
  -20.5987   -8.2775  -25.3575         0   -0.0212
   11.0113    1.5825   30.5125   -0.0200   -0.0012
    5.3113    4.5525    9.9525         0    0.0187
  -16.6987   -2.3575  -35.7575         0    0.0087
   20.7913    2.1825   20.3825         0   -0.0313

% MATRIZ DE SOMAS DE QUADRADOS DOS RESÍDUOS
W=DIF' * DIF
W =
   1.0e+04 *
    0.3805    0.0859    0.7360   -0.0000   -0.0000
    0.0859    0.0312    0.1359   -0.0000    0.0000
    0.7360    0.1359    2.2913   -0.0005   -0.0004
   -0.0000   -0.0000   -0.0005    0.0000    0.0000
   -0.0000    0.0000   -0.0004    0.0000    0.0000

% CÁLCULO DO VALOR DE LAMBDA DE WILKS
L=det(W)/ det(B+W)
L =
    0.0110

% CÁLCULO DO VALOR DE LAMBDA DE WILKS CORRIGIDO
LCORRIGIDO = -(n-1-((p+g)/2))* log(L)
LCORRIGIDO =
   85.7637

% VALOR CRÍTICO DE QUI-QUADRADO
QUI2=chi2inv(0.99,(p*(g-1)))
QUI2 =
   23.2093
```

```
# SOLUÇÃO NO JULIA
#Matriz de dados geral (variáveis respostas)
X=[43.48    8.42    73.70   0.56    0.49
67.96   12.79   121.04  0.56    0.50
65.19   11.61   114.27  0.54    0.47
70.50   14.36   119.43  0.59    0.50
47.40   8.53    78.21   0.59    0.51
52.90   10.58   95.22   0.57    0.49
61.30   12.87   109.73  0.53    0.46
59.70   11.04   98.51   0.57    0.48
72.10   17.46   143.16  0.52    0.53
70.87   20.93   138.80  0.51    0.50
85.31   22.18   197.15  0.51    0.50
66.04   20.70   124.67  0.52    0.54
74.06   17.77   216.39  0.51    0.51
89.53   18.80   222.53  0.54    0.54
69.15   18.58   119.72  0.58    0.55
87.99   21.13   146.65  0.58    0.56
110.08  19.43   178.07  0.58    0.58
159.37  33.56   278.88  0.58    0.57
127.09  17.30   222.87  0.60    0.56
111.52  14.38   201.16  0.58    0.54
143.13  24.24   257.03  0.56    0.56
137.43  27.21   236.47  0.58    0.58
115.42  20.30   190.76  0.58    0.57
152.91  24.84   246.90  0.58    0.53]

#Matriz de dados para o grupo 1 (Tratamento 1)
X1=X[1:8,:]

#Matriz de dados para o grupo 2 (Tratamento 2)
X2=X[9:16,:]

#Matriz de dados para o grupo 3 (Tratamento 3)
X3=X[17:24,:]

#Tamanho da amostra do grupo 1 (n1) e número de variáveis (p)
(n1,p)=size(X1)
#Tamanho da amostra do grupo 2 (n2) e número de variáveis (p)
(n2,p)=size(X2)

#Tamanho da amostra do grupo 3 (n3) e número de variáveis (p)
(n3,p)=size(X3)

#Tamanho total das amostras
n=n1+n2+n3

#Vetor do tamanho das amostras dos grupos
ng=[n1;n2;n3]
```

```
#Número de grupos
g=size(ng,1)

#Vetor de médias para o grupo 1
MEDIAX1=mean(X1,dims=1)

#Vetor de médias para o grupo 2
MEDIAX2=mean(X2,dims=1)

#Vetor de médias para o grupo 3
MEDIAX3=mean(X3,dims=1)

#Vetor de médias dos grupos
MEDIAS=[MEDIAX1;MEDIAX2;MEDIAX3]

#Vetor de médias geral
MEDIAGERAL=mean(X,dims=1)

#Matriz de diferenças das médias dos grupos e a média geral
DIFMEDIA=MEDIAS.-MEDIAGERAL

#Matriz de diferenças das médias dos grupos e a média geral corrigida pelo
tamanho dos grupos
DIFMEDIACOR=DIFMEDIA.*ng

#Matriz de somas de quadrados dos grupos
B=DIFMEDIACOR'*DIFMEDIA

#Matriz de diferenças dos valores de cada grupo e a média de seu respectivo
grupo
DIFX1=X1.-MEDIAX1
DIFX2=X2.-MEDIAX2
DIFX3=X3.-MEDIAX3
DIF=[DIFX1;DIFX2;DIFX3]

#Matriz de somas de quadrados dos resíduos
W=DIF'*DIF

#Cálculo do valor de Lambda de Wilks
L=det(W)/det(B+W)

#Cálculo do valor de Lambda de Wilks corrigido
Lcorrigido=-(n-1-((p+g)/2))*log(L)

#Valor crítico de Qui-Quadrado
QUIQUADRAINV= quantile(Chisq(p * (g - 1)), 0.99)
```

7 ANÁLISE DE DISCRIMINANTES

7.1 INTRODUÇÃO

Análise discriminante, principalmente em problemas taxonômicos, começou com Fisher (1936) que afirmou: "quando duas ou mais populações foram medidas em várias características, $x_1, x_2, \ldots, x_s$, existe especial interesse em obter funções lineares das medições por meio das quais as populações são mais bem discriminadas". Sneath e Sokal (1973), ressaltaram que o propósito de uma função discriminante é minimizar a probabilidade de uma classificação errada de indivíduos desconhecidos em seu respectivo grupo.

O principal objetivo da análise discriminante é classificar um objeto em uma de $k \geq 2$ populações com base em um vetor de observações multivariada x desse objeto. O termo grupo é usado para representar uma população ou uma amostra da população. Os dois maiores objetivos da separação em grupos são a descrição da separação dos grupos, nas quais funções lineares das variáveis (funções discriminantes) são usadas para descrever a diferença entre dois ou mais grupos, identificando a contribuição relativa das p variáveis na separação dos grupos. O segundo objetivo é a predição ou alocação de observações aos grupos nas quais funções lineares ou quadráticas das variáveis (funções de classificação) são empregadas para alocar uma unidade de amostra individual a um dos grupos.

O principal problema é que os métodos tradicionais de análise discriminante assumem que as matrizes de dispersão dos grupos sejam homogêneas (isto é, que todos os conglomerados possuam o mesmo tamanho, forma e orientação no espaço fenético), e que os conglomerados ou grupos sigam distribuições normais multivariadas. Se o grupo das matrizes de dispersão forem iguais, então procedimentos de classificação lineares devem ser usados, caso contrário, procedimentos quadráticos devem ser empregados (Sneath e Sokal, 1973).

Em análise de discriminante o objetivo é examinar o quanto é possível distinguir entre membros de vários grupos (ex.: tipos florestais, variedades de plantas, qualidade de um produto etc.) baseando-se em observações feitas sobre eles. Análise de discriminante é então uma extensão de observações multivariadas de uma análise de variância comum dentro e entre grupos. A obtenção das variáveis discriminantes é selecionada formando-se uma ou mais combinações lineares na seguinte forma:

$$D_i = d_{i1}z_1 + d_{i2}z_2 + \cdots + d_{ip}z_p$$

Em que: D_i é o valor (score) da função discriminante i; $d's$ são os pesos dos coeficientes; e $z's$ são os valores padronizados das p variáveis discriminantes usadas na análise.

A possível ordem n da função discriminante (subespaço) depende dos tamanhos relativos de g, do número de grupos, e p, o número de variáveis discriminantes. Assim, o número máximo possível de funções é uma a menos do que o número de grupos ou igual ao número de variáveis discriminantes, se houver mais grupos do que variáveis.

A análise de discriminantes algumas vezes falha, pois o resultado da função discriminante não separa muito bem os grupos, devido a dois fatores, quais sejam, os grupos se sobrepõem no espaço de medição escolhido, ou os grupos não podem ser separados por meio da função da forma adaptada para a análise.

Se a análise de discriminantes fornece resultados insatisfatórios, devido ao conjunto de dados conterem menos grupos identificáveis do que o conhecimento anterior sugere, então a análise de conglomerados (cluster) deverá estar em primeiro lugar das ferramentas de diagnóstico.

A análise discriminante pode ser usada para dois ou mais grupos. Exemplos de cada situação estão apresentados a seguir.

7.2 ANÁLISE DE DISCRIMINANTE PARA DOIS GRUPOS

Partindo do pressuposto que as duas populações a serem comparadas tem a mesma matriz variância-covariância ($\sum$), mas distintos vetores de médias μ_{1y} e μ_{2y}, em que cada vetor y_{ij} consiste em medidas em p variáveis, a função discriminante é a combinação linear dessas p variáveis que maximizam a distância entre os dois grupos (transformados) de vetores de médias. A combinação linear transforma cada vetor de observação em um escalar.

O objetivo, então, é desenvolver as combinações lineares dos $\underline{X}'s$ (observações multivariadas originais), criando os $Y's$ (observações univariadas transformadas). Seja considerada μ_{1y} a média dos $Y's$ obtidos dos $\underline{X}'s$ pertencentes à população π_1 e μ_{2y} a média dos $Y's$ obtidos dos $\underline{X}'s$ pertencentes à população π_2, então Fisher selecionou a combinação linear que maximiza a distância quadrática entre μ_{1y} e μ_{2y} relativamente à variabilidade dos $Y's$.

Proposta por Fisher, a ideia é transformar as observações multivariadas $\underline{X}$ em observações univariadas Y, tal que os $Y's$ nas populações π_1 e π_2 sejam separadas tanto quanto possível. Não é necessário assumir normalidade das populações, mas considera-se que elas possuem mesma matriz de variância-covariância. A melhor combinação linear vem da razão entre o quadrado da distância entre as médias e a variância de Y, ou seja:

$$\frac{\left(\mu_{1y}-\mu_{2y}\right)^2}{\sigma_Y^2}=\frac{\left(\underline{C}'\underline{\mu}_1-\underline{C}'\underline{\mu}_2\right)^2}{\underline{C}'\sum\underline{C}}$$

Em que: μ_{1y} é a média univariada da população 1, obtida a partir da combinação linear; μ_{2y} é a média univariada da população 2, obtida a partir da combinação linear; σ_Y^2 é a variância univariada, obtida a partir da combinação linear; $\underline{C}'$ é o vetor da combinação linear; $\underline{\mu}_1$ é o vetor de médias das variáveis originais da população 1; $\underline{\mu}_2$ é o vetor de médias das variáveis originais da população 2; $\sum$ é a matriz variância-covariância das variáveis originais.

A função discriminante linear de Fisher será dada por:

$$Y=\underline{C}'\underline{\mathrm{X}}=\left(\underline{\mu}_1-\underline{\mu}_2\right)'\sum^{-1}\underline{\mathrm{X}}$$

Em que: Y é a função discriminante linear de Fisher; $\underline{\mathrm{X}}$ é a matriz das varáveis originais.

A F.D.L. de Fisher transforma populações multivariadas π_1 e π_2 em populações univariadas, tais que as médias das populações univariadas sejam separadas, tanto o quanto possível, relativamente à variância populacional considerada comum.

Para o caso de amostras, a F.D.L. de Fisher é dada por:

$$\hat{Y}=\hat{\underline{C}}'\underline{\mathrm{X}}=\left(\underline{\bar{X}}_1-\underline{\bar{X}}_2\right)'{S_p}^{-1}\underline{\mathrm{X}}$$

Em que: $\underline{\bar{X}}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}X_{i1}$ (estima $\underline{\mu}_1$); $\underline{\bar{X}}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}X_{i2}$ (estima $\underline{\mu}_2$); $S_p=\frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$ (estima $\sum$); Médias: $\underline{\bar{X}}_1\,y=\hat{\underline{C}}'\underline{\bar{X}}_1$ e $\underline{\bar{X}}_2\,y=\hat{\underline{C}}'\underline{\bar{X}}_2$; Variância: $S_y^2=\hat{\underline{C}}'\underline{\sum}\hat{\underline{C}}$; Ponto médio: $\hat{m}=\frac{1}{2}\left(\underline{\bar{X}}_1+\underline{\bar{X}}_2\right)'{S_p}^{-1}\left(\underline{\bar{X}}_1+\underline{\bar{X}}_2\right)=\frac{\bar{Y}_1+\bar{Y}_2}{2}$.

Dada uma nova observação multivariada ($\underline{X}_0$), esta pode ser classificada obtendo-se:

$$\hat{Y}_0=\hat{\underline{C}}'\underline{\mathrm{X}}=\left(\underline{\bar{X}}_1-\underline{\bar{X}}_2\right)'{S_p}^{-1}\underline{X}_0$$

Em que: $\hat{Y}_0$ é um escalar obtido a partir da F.D.L.

Com o valor da F.D.L. de Fisher para uma nova observação $\underline{X}_0$, considerando o ponto médio ($\hat{m}$) entre as duas populações univariadas, tem-se que:

$$\hat{m} = \frac{1}{2}\left(\underline{\bar{X}}_1 + \underline{\bar{X}}_2\right)' S_p^{-1}\left(\underline{\bar{X}}_1 + \underline{\bar{X}}_2\right) = \frac{\overline{Y}_1 + \overline{Y}_2}{2}$$

As regras de alocação para classificação de uma nova observação são:

$$Y_0 \geq m \text{ alocar } \underline{X}_0 \text{ em } \pi_1$$

$$Y_0 < m \text{ alocar } \underline{X}_0 \text{ em } \pi_2$$

Na Figura 7.1 está ilustradra a separação de dois grupos bivariados normais com duas variáveis ($p = 2$) em uma dimensão simples representada pela função discriminante (Y). Nesta ilustração as matrizes de variância-covariância são iguais. A combinação linear projeta os pontos multivariados das observações das duas variáveis originais para a linha ótima de separação univariada dos dois grupos. Como as duas variáveis originais são consideradas bivariadas normais, a combinação linear será uma normal univariada.

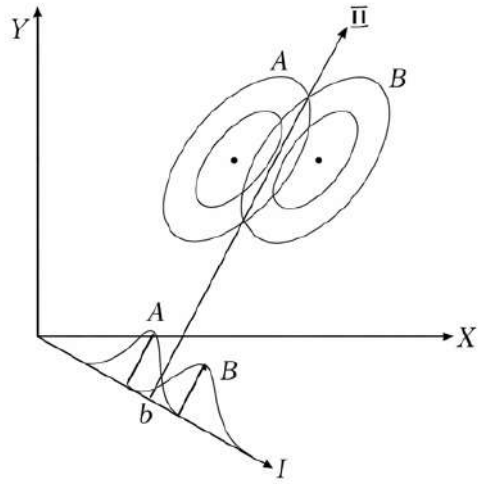


Figura 7. 1: Análise discriminante para dois grupos e duas variáveis originais. Dois conjuntos de elipses concêntricas representando a sobreposição bivariada para os dois grupos.

Exemplo 7.1

A seguir está apresentado um exemplo de análise de discriminante onde se comparou dois componentes da biomassa de árvores de acácia-negra (grupos), denominados de biomassa do fuste (madeira e casca) e copa (galhos e folhas). Cinco variáveis energéticas, usualmente utilizadas em análises de qualidade da biomassa para energia, foram avaliadas. O experimento foi conduzido na região de Piratini (RS), cujos dados apresentados na Tabela 7.1.

Uma nova observação $\underline{X}_0' = [0{,}45 \quad 68{,}3 \quad 26{,}4 \quad 3{,}2 \quad 5100{,}41]$ será classificada no exemplo se pertencente a amostra da copa ou do fuste.

A análise discriminante linear de Fisher foi aplicada com o objetivo de avaliar a separação entre duas populações multivariadas descritas por cinco variáveis quantitativas. Inicialmente, estimaram-se os vetores de médias das populações, obtendo-se para a População 1 (copa) o vetor médio $\underline{\bar{X}}_1 = [0{,}50 \quad 69{,}33 \quad 27{,}43 \quad 3{,}24 \quad 5113{,}53]$ e, para a População 2 (fuste), $\underline{\bar{X}}_2 = [0{,}45 \quad 75{,}98 \quad 22{,}80 \quad 1{,}22 \quad 4499{,}58]$. Esses valores evidenciam diferenças claras entre os grupos, sobretudo nas variáveis 2, 3, 4 e 5, indicando perfis médios distintos.

Assumindo-se igualdade das matrizes de variância-covariância, foi calculada a matriz de variância-covariância (S_p) comum e, a partir dela, construiu-se a função discriminante linear de Fisher. A função obtida foi dada pelo vetor de coeficientes

$\hat{Y} = [2{,}16 \quad -21{,}25 \quad -21{,}16 \quad -20{,}30 \quad 0{,}00206]$, revelando que as variáveis 2, 3 e 4 exercem maior influência na discriminação entre as populações, todas com pesos negativos e de magnitude relativa.

A projeção dos vetores médios das populações nessa dimensão discriminante resultou em médias univariadas iguais a $\overline{Y}_1 = -2108{,}01$ para a População 1 e $\overline{Y}_2 = -2111{,}79$ para a População 2. A variância comum associada a essa projeção foi $S_y^2 = 3{,}78$, indicando uma dispersão considerável em torno das médias projetadas. O ponto de corte ótimo para classificação, definido como o ponto médio entre as médias discriminantes, foi $\hat{m} = -2109{,}90$.

Tabela 7.1: Resultados obtidos para a biomassa do fuste e copa e cinco variáveis energéticas, com 21 repetições cada.

n	COMPONENTE COPA					COMPONENTE FUSTE				
	Y1	Y2	Y3	Y4	Y5	Y1	Y2	Y3	Y4	Y5
1	0,47	68,62	28,10	3,29	5433,00	0,40	75,98	23,08	0,94	4563,00
2	0,47	69,02	27,85	3,14	5235,00	0,44	77,52	21,45	1,04	4601,00
3	0,48	69,10	27,67	3,23	5092,24	0,40	75,94	22,69	1,37	4568,00
4	0,51	69,10	27,67	3,23	5092,24	0,43	75,21	23,40	1,40	4518,57
5	0,53	69,13	28,23	2,64	5312,00	0,45	75,97	23,14	0,89	4659,00
6	0,55	69,10	27,67	3,23	5092,24	0,45	75,85	22,89	1,25	4552,00
7	0,45	69,10	27,67	3,23	5092,24	0,46	75,42	23,44	1,14	4508,00
8	0,50	69,10	27,67	3,23	5092,24	0,48	77,08	21,93	0,99	4481,00
9	0,54	71,19	25,52	3,29	4991,00	0,44	77,53	21,37	1,10	4522,00
10	0,57	70,68	27,19	2,13	5018,00	0,41	77,07	22,05	0,88	4412,00
11	0,50	69,17	27,44	3,39	5157,00	0,42	76,05	22,47	1,47	4449,00
12	0,51	69,10	27,67	3,23	5092,24	0,43	75,21	23,40	1,40	4518,57
13	0,66	69,10	27,67	3,23	5092,24	0,44	75,07	23,59	1,35	4549,00
14	0,50	67,48	29,47	3,05	5121,00	0,57	75,21	23,40	1,40	4620,00
15	0,44	69,10	27,67	3,23	5092,24	0,32	73,64	24,73	1,63	4541,00
16	0,48	68,63	27,51	3,86	5168,00	0,35	75,02	23,58	1,40	4469,00
17	0,58	69,10	27,67	3,23	5092,24	0,44	75,66	22,92	1,42	4411,00
18	0,40	69,75	26,33	3,92	5112,00	0,39	75,64	23,38	0,98	4487,00
19	0,41	70,67	25,99	3,34	4945,00	0,34	75,20	23,32	1,49	4406,00
20	0,42	70,30	26,14	3,57	5059,00	0,47	77,08	21,88	1,03	4386,00
21	0,56	69,41	27,19	3,40	5003,00	0,40	78,15	20,71	1,14	4270,00

Em que: $Y1 =$ densidade básica (kg/m^3), $Y2 =$ teor de materiais voláteis (%), $Y3 =$ teor de carbono fixo (%), $Y4 =$ teor de cinzas (%) e $Y5 =$ poder calorífico superior (kcal/g).

Com base nesses valores, estimou-se a probabilidade de classificação incorreta a partir da distribuição normal padrão, resultando em uma probabilidade total de erro de aproximadamente 33,11%. Esse valor relativamente elevado indica a sobreposição das distribuições univariadas das duas populações no eixo discriminante, o que foi confirmado pela análise gráfica das densidades normais associadas a cada grupo (Figura 7.2).

A função discriminante foi utilizada para a classificação de uma nova observação $\underline{X}_0'$. O valor discriminante obtido foi $\hat{Y}_0 = -2063{,}63$, o qual se mostrou maior no ponto

de corte $\hat{m} = -2109{,}90$. Dessa forma, a nova observação foi classificada como pertencente à População 1 (copa). Esse resultado é consistente com o perfil multivariado da observação, que apresenta maior proximidade com as médias dessa população. A análise discriminante linear evidenciou diferenças multivariadas relevantes entre os grupos e forneceu um critério objetivo de classificação, ainda que com uma taxa de erro moderada decorrente da sobreposição entre as populações.

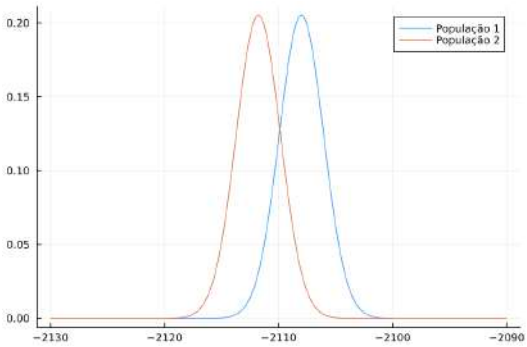


Figura 7.2: Densidades normais associadas a cada grupo (Copa – população 1 e Fuste – população 2).

7.3 ANÁLISE DE DISCRIMINANTE PARA VÁRIOS GRUPOS

Na análise discriminante para vários grupos (tratamentos) o objetivo é encontrar combinações lineares das variáveis que melhor separam os grupos com observações multivariadas. A análise de discriminante para vários grupos pode servir a vários propósitos, quais sejam examinar a separação de grupos em um gráfico bidimensional; encontrar um subconjunto das variáveis originais que separam os grupos tão bem quanto o conjunto original de dados; ordenar as variáveis em termos de sua relativa contribuição na separação dos grupos; interpretar as novas dimensões representadas pelas funções discriminantes; classificar uma nova observação ou complementar uma MANOVA dos efeitos dos grupos.

O método de discriminação anteriormente discutido pode ser estendido para diversas populações. Para vários grupos com n observações em cada um deles, cada vetor de observação multivariado é transformado para se obter um novo vetor e suas respectivas médias transformadas. Nesse caso, a generalização multi-grupal no caso de duas populações pode ser obtida por:

$$\frac{\underline{C}' \left[\sum_{i=1}^{g} \left(\underline{\mu}_i - \overline{\underline{\mu}}\right)\left(\underline{\mu}_i - \overline{\underline{\mu}}\right)'\right] \underline{C}}{\underline{C}' \sum \underline{C}} = \frac{\underline{C}' B_0 \underline{C}}{\underline{C}' \sum \underline{C}}$$

Em que: $\underline{C}$ é o vetor da combinação linear, que correspondem aos autovetores; B_0 é a matriz variância-covariância entre os grupos (sem considerar o número de observações); $\sum$ é a matriz variância-covariância dentro dos grupos.

Para vários grupos é utilizada a razão matriz B_0 e W, que correspondem as matrizes de matriz variância-covariância entre e dentro dos grupos. Esta razão mede a variabilidade entre os grupos de valores Y concernente à variabilidade comum dentro dos grupos. Este método não necessita da suposição de que as diversas populações sejam normais, entretanto, é assumido que as matrizes de covariâncias populacionais $\sum$ sejam iguais, ou seja, $\sum_1 = \sum_2 = \cdots = \sum_g$.

O número de funções (s) será dado por $s \leq mín(g-1, p)$, em que g é o número de grupos e p é o número de variáveis. A partir da matriz $\sum^{-1} B_0$ são extraídos os autovalores $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_s \geq 0$. Os autovetores $\underline{e}_1$, $\underline{e}_2$, $\ldots$, $\underline{e}_s$ que correspondem aos

autovalores, são escalonados, tal que $\underline{e}'\sum\underline{e} = 1$, gerando as combinações lineares de Y (funções discriminantes).

A primeira combinação linear $\underline{e_1'}\underline{X}$ é denominada primeira discriminante. O vetor $\underline{\mathrm{e}}_2$ que maximiza a razão $\frac{\underline{C}'B_0\underline{C}}{\underline{C}'\sum\underline{C}}$, com a condição de que $cov\left(\underline{C_1'}\underline{X}, \underline{C_2'}\underline{X}\right) = 0$, é a combinação linear $\underline{e_2'}\underline{X}$, que gera a segunda discriminante e assim sucessivamente, obtém-se as demais funções. Essas funções discriminantes não são correlacionadas e, também, não são ortogonais, porque $\sum^{-1}B_0$ não é uma matriz simétrica. Importante salientar que as funções discriminantes são correspondentes aos autovalores e não aos grupos.

As estimativas de $\sum$ e $\underline{\mu_i}$ são obtidas a partir de amostras aleatórias de tamanhos n_i das populações $\pi_i(i = 1,2, \dots, g)$, sendo que:

$$\underline{\bar{X}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \text{ estima } \underline{\mu_i}$$

$$\underline{\overline{X}} = \frac{\sum_{i=1}^{g} n_i\underline{\overline{X}_i}}{\sum_{i=1}^{g} n_i} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\underline{X_{ij}}}{\sum_{i=1}^{g} n_i} \text{ estima } \underline{\overline{\mu}}, \text{ média geral}$$

A matriz soma de produtos cruzados (variação entre grupos amostrais), sem considerar o tamanho das amostras, é dada por:

$$B_0 = \sum_{i=1}^{g}\left(\underline{\overline{X}_i} - \underline{\overline{X}}\right)\left(\underline{\overline{X}_i} - \underline{\overline{X}}\right)'$$

O estimador da matriz $\sum$ é obtido com base na matriz W (variação dentro dos grupos amostrais):

$$W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\underline{X_{ij}} - \underline{\overline{X}_i}\right)\left(\underline{X_{ij}} - \underline{\overline{X}_i}\right)'$$

Consequentemente:

$$\frac{W}{n_1 + n_2 + \cdots + n_g - g} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + \left(n_g - 1\right)S_g}{n_1 + n_2 + \cdots + n_g - g} = S_p$$

$$W = \left(n_1 + n_2 + \cdots + n_g - g\right)S_p$$

Uma vez obtida a matriz $W^{-1}B_0$, nos passos seguintes são obtidos os autovalores e autovetores. Para se obter as funções discriminantes deve-se escalonar os autovetores, multiplicando-se por uma constante k:

$$\underline{\mathrm{e}}'\mathrm{S_p}\underline{\mathrm{e}} = 1$$
$$\left(\mathrm{k}\underline{\mathrm{e}}'\right)\mathrm{S_p}\left(\mathrm{k}\underline{\mathrm{e}}\right) = 1$$
$$\mathrm{k}^2\left(\underline{\mathrm{e}}'\mathrm{S_p}\underline{\mathrm{e}}\right) = 1$$
$$\mathrm{k} = \frac{1}{\sqrt{\underline{\mathrm{e}}'\mathrm{S_p}\underline{\mathrm{e}}}}$$

O novo autovetor escalonado é dado por:

$$\underline{e_{esc}} = k\,\underline{e}$$

Os vetores escalonados obtidos correspondem aos coeficientes das funções discriminantes não padronizadas.

No procedimento anterior (duas populações) a contribuição relativa da função discriminante na separação dos grupos era avaliada pela comparação dos coeficientes na função discriminante. Tais comparações são meramente informativas, quando as

variáveis foram medidas na mesma escala. Se esse não for o caso, deve-se obter coeficientes que são aplicáveis a variáveis padronizadas.

Importante salientar que os autovalores e autovetores da matriz $W^{-1}B_0$ podem ser obtidos a partir dos dados originais padronizados. A função discriminante padronizada é dada por:

$$e^{*}_{esc} = \left(diagonal\ S_p\right)^{0,5}\ e_{esc}$$

Para a obtenção da importância relativa de cada função discriminante deve-se calcular novamente a matriz $W^{-1}B$, só que, neste caso, B será multiplicado pelos tamanhos das amostras n_i, o que corresponde ao mesmo procedimento utilizado na aplicação da MANOVA. A importância relativa de cada função (IR_i) pode ser obtida considerando cada autovalor como parte do total, ou seja:

$$IR_i = \frac{\lambda_i}{\sum_{i=1}^{s}\lambda_i}100$$

Por esse critério, duas ou três funções discriminantes serão suficientes para descrever a diferença entre os grupos. As funções discriminantes associadas a pequenos autovalores podem ser desconsideradas. Testes de significância e correlações canônicas para cada função discriminante podem ser obtidos, para auxiliar na escolha das funções a serem utilizadas.

A pergunta que se apresenta agora é: quão bem as variáveis separam os grupos? Anteriormente na MANOVA foi calculado o teste de Roy. A raiz quadrada da estatística de Roy serve como uma medida de associação. Essa estatística é a correlação máxima entre a primeira função discriminante e a melhor combinação das variáveis medidas, podendo ser estendida para as demais funções. A estatística de Roy corresponde à razão $\frac{\lambda_1}{1+\lambda_1}$. Então, a correlação canônica para cada função discriminante é dada por:

$$r_i = \left(\frac{\lambda_i}{1+\lambda_i}\right)^{0,5}$$

Outra questão que surge é o cálculo da significância das funções, que pode ser obtida por meio do teste de Wilks lambda. O critério de discriminação $L = \frac{|W|}{|S|}$ é maximizado por λ_1, o maior autovalor de $W^{-1}B$, sendo que os autovalores remanescentes correspondem às demais dimensões discriminantes. Estes autovalores são os mesmos que os obtidos quando da ralização da MANOVA, e permitem avaliar as diferenças entre os vetores de médias dos grupos.

Para cada função discriminante, o cálculo de Wilks lambda é dado por:

$L_1 = \prod_{i=1}^{s}\frac{1}{1+\lambda_i}$, para a primeira função.

$L_2 = \prod_{i=2}^{s}\frac{1}{1+\lambda_i}$, para a segunda função, e assim sucessivamente.

Após o cálculo de Wilks – L_i, é efetuado o teste de significância por meio do teste χ^2 (Qui-quadrado) aproximado de Bartlett, baseado no critério de Wilk's, para cada função.

$$\chi_i^2 = -\left[n - 1 - \frac{1}{2}\ (p+q)\right] lnL_i$$

Em que: n é o número total de observações; p é o número de variáveis; q é o número de grupos -1; $\ln L_i$ é o logaritmo neperiano de L_i.

O Qui-quadrado calculado é então comparado com a distribuição de χ^2 a um determinado nível de probabilidade estabelecido. Os graus de liberdade para cada função são dados por:

$$gl_i = (p - m + 1)\,(q - m)$$

Em que: m corresponde a ordem da aplicação do teste de χ^2 para cada função discriminante (função 1 $m = 1$, função 2 $m = 2$, e sucessivamente).

Outra questão refere-se à avaliação da contribuição de cada variável na separação dos grupos, sem considerar a dependência conjunta das variáveis. Para tanto, a matriz de estruturas (me) pode ser utilizada. Esta matriz pode ser calculada como:

$$R_p = \left[\left(diagonal\ S_p\right)^{0,5}\right]^{-1} S_p \left[\left(diagonal\ S_p\right)^{0,5}\right]^{-1}$$

$$c = e_{esc}\left(diagonal\ S_p\right)^{0,5}$$

$$me = R_p\, c'$$

Como etapa final, pode-se calcular os escores das funções padronizadas, que são obtidos pela multiplicação dos coeficientes pela matriz de dados padronizada. Com esses valores é possível obter gráficos que representam os grupos com seus respectivos centróides em um espaço bidimensional.

Uma síntese da análise discriminante está descrita a seguir:

- Organização e definição dos dados
- Padronização das variáveis
- Vetores médios dos grupos e média geral
- Matrizes de variância–covariância
- Matrizes de variação entre e dentro dos grupos
- Número de funções discriminantes
- Autovalores das funções discriminantes
- Importância das funções discriminantes
  - Variância explicada
  - Correlação canônica
  - Teste de significância – Lambda de Wilks
- Coeficientes das funções discriminantes
  - Autovetores escalonados
  - Coeficientes Não padronizados
  - Coeficientes padronizados
- Matriz de estruturas
- Escores discriminantes e centróides, representação gráfica

Exemplo 7.2

A seguir está apresentado um exemplo de análise de discriminante, no qual se comparou três idades de acácia-negra (grupos/tratamentos), cada idade com quatro variáveis energéticas usualmente utilizadas em análises de qualidade da biomassa da copa para energia. O experimento foi conduzido na região de Piratini (RS), sendo os dados apresentados na Tabela 7.2.

A análise discriminante linear foi conduzida com o objetivo de avaliar a capacidade de separação entre três grupos, cada um composto por dez observações, considerando quatro variáveis explicativas. O número máximo de funções discriminantes possíveis é dado pelo menor valor entre o número de variáveis e o número de grupos menos um. Assim, foram extraídas duas funções discriminantes canônicas.

Inicialmente, os dados foram padronizados, de modo que cada variável apresentasse média zero e desvio-padrão unitário no conjunto total de observações. Esse procedimento garante que todas as variáveis contribuam de forma equivalente para a análise, evitando que as diferenças de escala influenciem os resultados.

Em seguida, foram calculados os vetores médios padronizados de cada grupo, os quais representam os centróides no espaço multivariado. A análise desses centróides

revelou padrões claros de diferenciação entre os grupos, indicando que o primeiro grupo se associa a valores relativamente mais elevados de determinadas variáveis, enquanto o terceiro grupo apresenta comportamento oposto, com o segundo grupo ocupando uma posição intermediária. Esses resultados iniciais já sugerem a existência de separação multivariada consistente entre os grupos.

Tabela 7.2: Resultados obtidos para três idades (anos) e quatro variáveis energéticas, com 10 repetições cada.

Grupo	Rep	Y1	Y2	Y3	Y4
1	1	0,47	68,62	28,10	5433,00
1	2	0,47	69,02	27,85	5235,00
1	3	0,48	69,10	27,67	5092,24
1	4	0,51	69,10	27,67	5092,24
1	5	0,53	69,13	28,23	5312,00
1	6	0,55	69,10	27,67	5092,24
1	7	0,45	69,10	27,67	5092,24
1	8	0,50	69,10	27,67	5092,24
1	9	0,54	71,19	25,52	4991,00
1	10	0,57	70,68	27,19	5018,00
2	1	0,49	73,44	24,92	4694,00
2	2	0,51	74,27	23,93	4759,00
2	3	0,49	74,46	23,53	4870,00
2	4	0,57	72,28	24,97	4691,00
2	5	0,56	72,37	25,12	4776,00
2	6	0,55	70,23	26,77	4878,00
2	7	0,52	73,49	23,83	4724,00
2	8	0,56	72,39	24,89	4815,00
2	9	0,54	73,16	24,37	4705,00
2	10	0,51	72,62	23,95	4855,00
3	1	0,56	80,99	18,50	4434,00
3	2	0,58	74,21	23,69	4442,00
3	3	0,51	73,77	24,34	4705,00
3	4	0,64	76,84	21,61	4613,00
3	5	0,54	74,01	23,99	4682,00
3	6	0,52	72,19	25,79	4708,00
3	7	0,34	71,55	25,49	4683,00
3	8	0,59	73,32	24,41	4653,83
3	9	0,52	77,53	21,59	4607,00
3	10	0,58	74,73	23,30	4607,00

Em que: $Y1 =$ densidade básica (kg/m^3), $Y2 =$ teor de materiais voláteis (%), $Y3 =$ teor de carbono fixo (%) e $Y4 =$ poder calorífico superior (kcal/g)

A variabilidade interna dos grupos foi sintetizada por meio da matriz de covariância combinada, calculada a partir das matrizes de covariância individuais de cada grupo. Na sequência foi calculada a matriz de dispersão entre grupos, que quantifica o grau de afastamento entre os centróides dos grupos no espaço padronizado. A comparação

entre essas duas fontes de variabilidade indicou que a separação entre grupos é expressiva quando comparada à variabilidade dentro dos grupos.

A partir da matriz $W^{-1}B$ foram obtidos quatro autovalores, dos quais apenas dois foram utilizados. O primeiro autovalor (5,12) foi maior que o segundo (0,10), indicando que a maior parte da discriminação entre os grupos é explicada por uma única função discriminante, pois a primeira função foi responsável por aproximadamente 98% da variância discriminante total, enquanto a segunda explicou 2% do total.

A correlação canônica associada à primeira função discriminante apresentou valor de 0,92 e a segunda função apresentou correlação canônica de 0,31. O teste Lambda de Wilks indicou significância estatística apenas para a primeira função discriminante.

Na etapa seguinte, foram obtidos os coeficientes das funções discriminantes, tanto em sua forma não padronizada quanto padronizada. Os coeficientes não padronizados são úteis para fins de classificação de novas observações, ao passo que os coeficientes padronizados permitem comparar a importância relativa das variáveis na formação das funções discriminantes.

A interpretação das funções discriminantes foi efetuada com base na matriz de estrutura, que expressa as correlações entre as variáveis originais e as funções discriminantes. Essa matriz revelou que a primeira função discriminante absorve todas as variáveis, em especial a variável Y4 (poder calorífico superior). A segunda função, por sua vez, apresentou correlações menores, absorvendo variações secundárias entre os grupos, sem relevância estatística significativa.

Os escores discriminantes foram obtidos a partir das funções padronizadas e utilizados na construção do gráfico. Este gráfico revelou uma separação visual clara entre os grupos ao longo da primeira função discriminante, com centróides bem afastados e mínima sobreposição. Considerando que a variável Y4 (poder calorífico superior) foi a de maior importância, cumpre salientar que o grupo 1 foi o que apresentou a maior média para esta variável (5.145,1 kcal/g).

De forma geral, os resultados indicam que a análise discriminante linear foi eficaz na separação dos grupos avaliados. A discriminação entre os grupos é efetuada por uma única função discriminante, estatisticamente significante e responsável pela quase totalidade da variância explicada.

7.4 CLASSIFICAÇÃO DE UMA NOVA OBSERVAÇÃO E AVALIAÇÃO DAS TAXAS DE ACERTO

Os aspectos descritivos na análise discriminante envolvem a separação entre grupos por meio das funções discriminantes. A alocação de observações aos grupos, que representa o aspecto preditivo da análise discriminante é chamada de análise de classificação, para distinguir dos aspectos descritivos. Alguns autores usam o termo análise de classificação para descrever análise de agrupamentos (Cluster), na qual as observações são agrupadas de acordo com os valores das variáveis, ao invés de em grupos pré-definidos.

A classificação de uma nova observação ($\underline{X}_0$) pode ser feita por meio da avaliação dos valores discriminantes gerados pelas funções obtidas, que propiciam a menor distância euclidiana entre os centroides de cada grupo. Para tanto, calcula-se as distâncias euclidianas entre todos os centroides dos grupos e a menor distância será a que define o grupo ao qual a nova observação pertence. As distâncias podem ser obtidas por:

$$d\left(\underline{x},\underline{y}\right) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

Em que: $\underline{x}$ é o vetor da observação a ser classificada; $\underline{y}$ é o vetor do centróide de cada grupo.

A efetividade da avaliação das funções discriminantes na separação em grupos é avaliada pelo uso de testes de significância ou pela importância de cada autovalor. Para julgar a habilidade dos procedimentos de classificação para predizer a classificação em grupos normalmente é usado a probabilidade de classificação errada, que é conhecida como taxa de erro. Também é possível utilizar seu complemento, a taxa de acerto.

Uma simples estimativa da taxa de erro pode ser obtida aplicando-se o procedimento de classificação nos mesmos dados que foram utilizados para calcular as funções. Cada observação é submetida às funções de classificação e alocadas a um grupo. As classificações corretas e incorretas são contadas. A proporção de classificação erradas é chamada de taxa aparente de erro. Os resultados são convenientemente mostrados em uma tabela de classificação ou matriz de confusão.

Supondo que existem dois grupos ($G1$ e $G2$), dentro das n_1 observações do grupo 1 ($G1$), n_{11} é o número de classificações corretas em $G1$ e n_{12} é o número de classificações erradas em $G2$, em que $n_1 = n_{11} + n_{12}$. De modo similar, dentro das n_2 observações de $G2$, n_{21} é o número de classificações erradas em $G1$ e n_{22} é o número de classificações corretas em $G2$, em que $n_2 = n_{21} + n_{22}$. Desta forma tem-se a taxa aparente de erro (TE), dada por:

$$TE = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

$$TE = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$$

Os resultados das classificações podem ser sumarizados em uma tabela, denominada de Matriz de Confusão (Tabela 7.3).

Tabela 7.3: Matriz de confusão para dois grupos.

Grupo original	Número de observações	Grupo predito	
		Grupo 1	Grupo 2
Grupo 1	n_1	n_{11}	n_{12}
Grupo 2	n_2	n_{21}	n_{22}

Este procedimento pode ser estendido para mais do que dois grupos.

Exemplo 7.3

Para os dados apresentados na Tabela 7.2 a matriz de confusão é a seguinte:

Grupo original	Número de observações	Grupo predito		
		Grupo 1	Grupo 2	Grupo 3
Grupo 1	10	10	0	0
Grupo 2	10	0	8	2
Grupo 3	10	0	2	8

A matriz de confusão resultou em uma taxa aparente de erro de 13,33%.

A classificação de uma nova observação com os valores das variáveis $\underline{X_0} = [0{,}62 \quad 76{,}78 \quad 21{,}62 \quad 4614{,}11]$ resultou como sendo do grupo 2.

7.5 ANÁLISE DISCRIMINANTE PARA DUAS POPULAÇÕES NO MATLAB

```
#Solução do exemplo 7.1
% MATRIZ DE DADOS DA POPULAÇÂO 1
X1 =
          0.47         68.62          28.1          3.29          5433
          0.47         69.02         27.85          3.14          5235
          0.48          69.1         27.67          3.23        5092.2
          0.51          69.1         27.67          3.23        5092.2
          0.53         69.13         28.23          2.64          5312
          0.55          69.1         27.67          3.23        5092.2
          0.45          69.1         27.67          3.23        5092.2
           0.5          69.1         27.67          3.23        5092.2
          0.54         71.19         25.52          3.29          4991
          0.57         70.68         27.19          2.13          5018
           0.5         69.17         27.44          3.39          5157
          0.51          69.1         27.67          3.23        5092.2
          0.66          69.1         27.67          3.23        5092.2
           0.5         67.48         29.47          3.05          5121
          0.44          69.1         27.67          3.23        5092.2
          0.48         68.63         27.51          3.86          5168
          0.58          69.1         27.67          3.23        5092.2
           0.4         69.75         26.33          3.92          5112
          0.41         70.67         25.99          3.34          4945
          0.42          70.3         26.14          3.57          5059
          0.56         69.41         27.19           3.4          5003

% MATRIZ DE DADOS DA POPULAÇÂO 2
X2 =
           0.4         75.98         23.08          0.94          4563
          0.44         77.52         21.45          1.04          4601
           0.4         75.94         22.69          1.37          4568
          0.43         75.21          23.4           1.4        4518.6
          0.45         75.97         23.14          0.89          4659
          0.45         75.85         22.89          1.25          4552
          0.46         75.42         23.44          1.14          4508
          0.48         77.08         21.93          0.99          4481
          0.44         77.53         21.37           1.1          4522
          0.41         77.07         22.05          0.88          4412
          0.42         76.05         22.47          1.47          4449
          0.43         75.21          23.4           1.4        4518.6
          0.44         75.07         23.59          1.35          4549
          0.57         75.21          23.4           1.4          4620
          0.32         73.64         24.73          1.63          4541
          0.35         75.02         23.58           1.4          4469
          0.44         75.66         22.92          1.42          4411
          0.39         75.64         23.38          0.98          4487
          0.34          75.2         23.32          1.49          4406
          0.47         77.08         21.88          1.03          4386
           0.4         78.15         20.71          1.14          4270
```

```
% MATRIZ DE DADOS GERAL
X=[X1;X2]
X =
          0.47         68.62          28.1          3.29          5433
          0.47         69.02         27.85          3.14          5235
          0.48          69.1         27.67          3.23        5092.2
          0.51          69.1         27.67          3.23        5092.2
          0.53         69.13         28.23          2.64          5312
          0.55          69.1         27.67          3.23        5092.2
          0.45          69.1         27.67          3.23        5092.2
           0.5          69.1         27.67          3.23        5092.2
          0.54         71.19         25.52          3.29          4991
          0.57         70.68         27.19          2.13          5018
           0.5         69.17         27.44          3.39          5157
          0.51          69.1         27.67          3.23        5092.2
          0.66          69.1         27.67          3.23        5092.2
           0.5         67.48         29.47          3.05          5121
          0.44          69.1         27.67          3.23        5092.2
          0.48         68.63         27.51          3.86          5168
          0.58          69.1         27.67          3.23        5092.2
           0.4         69.75         26.33          3.92          5112
          0.41         70.67         25.99          3.34          4945
          0.42          70.3         26.14          3.57          5059
          0.56         69.41         27.19           3.4          5003
           0.4         75.98         23.08          0.94          4563
          0.44         77.52         21.45          1.04          4601
           0.4         75.94         22.69          1.37          4568
          0.43         75.21          23.4           1.4        4518.6
          0.45         75.97         23.14          0.89          4659
          0.45         75.85         22.89          1.25          4552
          0.46         75.42         23.44          1.14          4508
          0.48         77.08         21.93          0.99          4481
          0.44         77.53         21.37           1.1          4522
          0.41         77.07         22.05          0.88          4412
          0.42         76.05         22.47          1.47          4449
          0.43         75.21          23.4           1.4        4518.6
          0.44         75.07         23.59          1.35          4549
          0.57         75.21          23.4           1.4          4620
          0.32         73.64         24.73          1.63          4541
          0.35         75.02         23.58           1.4          4469
          0.44         75.66         22.92          1.42          4411
          0.39         75.64         23.38          0.98          4487
          0.34          75.2         23.32          1.49          4406
          0.47         77.08         21.88          1.03          4386
           0.4         78.15         20.71          1.14          4270

% MÉDIAS DAS POPULAÇÕES
MI1=mean(X1)
MI1 =
       0.50143        69.331        27.428        3.2424        5113.5
MI2=mean(X2)
MI2 =
       0.42524        75.976        22.801        1.2243        4499.6

% MATRIZ VARIÂNCIA COVARIÂNCIA
S=cov(X)
```

```
S =
    0.0047935     -0.12351     0.090585     0.032903       12.418
     -0.12351        12.2       -8.671      -3.5295      -1085.3
     0.090585       -8.671       6.2748       2.3969        769.2
     0.032903      -3.5295       2.3969       1.1328        316.32
       12.418      -1085.3        769.2        316.32    1.0614e+05

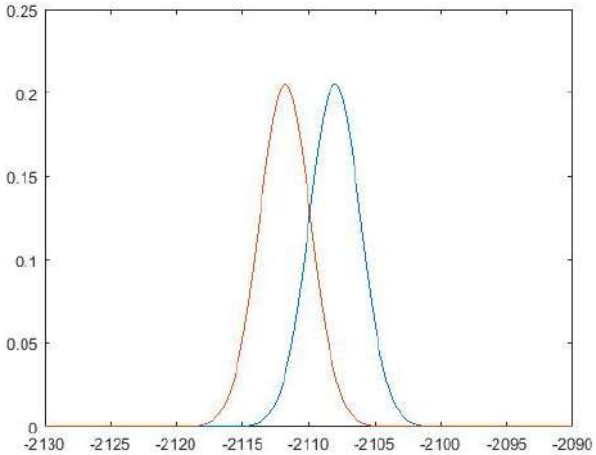
% FUNÇÃO DISCRIMINANTE LINEAR DE FISHER
Y=(MI1-MI2)*inv(S)
Y =
       2.1621      -21.252      -21.161      -20.297    0.0020578

% MÉDIA DAS POPULAÇÕES UNIVARIADAS
MI1Y=Y*MI1'
MI1Y =
        -2108
MI2Y=Y*MI2'
MI2Y =
      -2111.8

% MATRIZ VARIÂNCIA COVARIÂNCIA DAS POPULAÇÕES UNIVARIADAS
VY=Y*S*Y'
VY =
       3.7783

% PONTO MÉDIO ENTRE AS POPULAÇÕES UNIVARIADAS
m=(MI1Y+MI2Y)/2
m =
      -2109.9

% GRÁFICO DAS FUNÇÕES UNIVARIADAS
x = -2130 : 0.1 : -2090;
y1 = (1/((sqrt(VY))*(sqrt(2*pi))))*exp(-0.5*(((x-
MI1Y)/(sqrt(VY))).^2));
y2 = (1/((sqrt(VY))*(sqrt(2*pi))))*exp(-0.5*(((x-
MI2Y)/(sqrt(VY))).^2));
plot(x,y1, x,y2)
```




```
% PROBABILIDADE DE CLASSIFICAÇÃO ERRADA
z=(m- MI1Y)/sqrt(VY)
z =
     -0.97189
```

```
AREA=normcdf(z,0,1)
AREA =
      0.16555
p=AREA*2
p =
       0.3311

% ALOCAÇÃO DE UMA NOVA OBSERVAÇÃO
X0=[0.45 68.3 26.4 3.2 5100.41]
X0 =
         0.45         68.3         26.4          3.2       5100.4
Y0=Y*X0'
Y0 =
      -2063.8
```

7.6 ANÁLISE DISCRIMINANTE PARA VÁRIAS POPULAÇÕES NO MATLAB

```
% VALORES DAS VARIÁVEIS PARA CADA GRUPO
X1 =
         0.47        68.62         28.1         5433
         0.47        69.02        27.85         5235
         0.48         69.1        27.67       5092.2
         0.51         69.1        27.67       5092.2
         0.53        69.13        28.23         5312
         0.55         69.1        27.67       5092.2
         0.45         69.1        27.67       5092.2
          0.5         69.1        27.67       5092.2
         0.54        71.19        25.52         4991
         0.57        70.68        27.19         5018
X2 =
         0.49        73.44        24.92         4694
         0.51        74.27        23.93         4759
         0.49        74.46        23.53         4870
         0.57        72.28        24.97         4691
         0.56        72.37        25.12         4776
         0.55        70.23        26.77         4878
         0.52        73.49        23.83         4724
         0.56        72.39        24.89         4815
         0.54        73.16        24.37         4705
         0.51        72.62        23.95         4855
X3 =
         0.56        80.99         18.5         4434
         0.58        74.21        23.69         4442
         0.51        73.77        24.34         4705
         0.64        76.84        21.61         4613
         0.54        74.01        23.99         4682
         0.52        72.19        25.79         4708
         0.34        71.55        25.49         4683
         0.59        73.32        24.41       4653.8
         0.52        77.53        21.59         4607
         0.58        74.73         23.3         4607

% MATRIZ DE DADOS GERAL
X = [X1; X2; X3]
X =
         0.47        68.62         28.1         5433
         0.47        69.02        27.85         5235
         0.48         69.1        27.67       5092.2
```

```
          0.51          69.1         27.67        5092.2
          0.53         69.13         28.23          5312
          0.55          69.1         27.67        5092.2
          0.45          69.1         27.67        5092.2
           0.5          69.1         27.67        5092.2
          0.54         71.19         25.52          4991
          0.57         70.68         27.19          5018
          0.49         73.44         24.92          4694
          0.51         74.27         23.93          4759
          0.49         74.46         23.53          4870
          0.57         72.28         24.97          4691
          0.56         72.37         25.12          4776
          0.55         70.23         26.77          4878
          0.52         73.49         23.83          4724
          0.56         72.39         24.89          4815
          0.54         73.16         24.37          4705
          0.51         72.62         23.95          4855
          0.56         80.99          18.5          4434
          0.58         74.21         23.69          4442
          0.51         73.77         24.34          4705
          0.64         76.84         21.61          4613
          0.54         74.01         23.99          4682
          0.52         72.19         25.79          4708
          0.34         71.55         25.49          4683
          0.59         73.32         24.41        4653.8
          0.52         77.53         21.59          4607
          0.58         74.73          23.3          4607

% NÚMERO DE GRUPOS
g=3;

% MATRIZ DE DADOS PADRONIZADOS
mediaX = mean(X)
mediaX =
         0.525          72.4        25.141        4845.1
desvX  = std(X)
desvX =
      0.054187         2.908        2.2763        248.76
Z1 = (X1 - mediaX) ./ desvX

Z1 =
        -1.015       -1.2997        1.2999        2.3634
        -1.015       -1.1622        1.1901        1.5675
      -0.83046       -1.1347         1.111        0.9936
      -0.27682       -1.1347         1.111        0.9936
      0.092273       -1.1244         1.357         1.877
       0.46137       -1.1347         1.111        0.9936
       -1.3841       -1.1347         1.111        0.9936
      -0.46137       -1.1347         1.111        0.9936
       0.27682      -0.41597        0.1665       0.58663
       0.83046      -0.59135       0.90014       0.69517
Z2 = (X2 - mediaX) ./ desvX
Z2 =
      -0.64591        0.35774     -0.097087      -0.60727
      -0.27682        0.64316        -0.532      -0.34598
      -0.64591         0.7085      -0.70773       0.10023
       0.83046      -0.04115     -0.075122      -0.61933
       0.64591     -0.010202    -0.0092255      -0.27764
```

```
      0.46137     -0.74609      0.71563      0.13238
    -0.092273      0.37494     -0.57593     -0.48668
      0.64591   -0.0033241     -0.11027     -0.12087
      0.27682      0.26146     -0.33871     -0.56306
     -0.27682     0.075767     -0.52322     0.039927
Z3 = (X3 - mediaX) ./ desvX
Z3 =
      0.64591        2.954      -2.9175      -1.6524
        1.015      0.62253     -0.63744      -1.6203
     -0.27682      0.47122     -0.35189     -0.56306
       2.1223       1.5269      -1.5512     -0.93288
      0.27682      0.55375     -0.50564     -0.65551
    -0.092273    -0.072099      0.28511       -0.551
      -3.4141     -0.29218      0.15332     -0.65149
       1.1996      0.31648     -0.32113     -0.76875
    -0.092273       1.7642        -1.56       -0.957
        1.015      0.80134     -0.80877       -0.957

% VETORES MÉDIOS DOS GRUPOS
MIX1 = mean(Z1)
MIX1 =
     -0.33218      -1.0267       1.0469       1.2058
MIX2 = mean(Z2)
MIX2 =
     0.092273      0.16208     -0.22537     -0.27483
MIX3 = mean(Z3)
MIX3 =
      0.23991      0.86461     -0.82151     -0.93094

% TAMANHO DAS AMOSTRAS DE CADA GRUPO
[n1,p] = size(X1)
n1 =
    10
p =
     4
[n2,p] = size(X2)
n2 =
    10
p =
     4
[n3,p] = size(X3)
n3 =
    10
p =
     4

% VETOR MÉDIO GERAL
MI = (n1*MIX1 + n2*MIX2 + n3*MIX3) / (n1+n2+n3)
MI =
  -2.6645e-16  -3.6119e-15  -2.1316e-15  -6.5133e-16

% MATRIZES DE VARIÂNCIA-COVARIÂNCIA
S1 = cov(Z1)
S1 =
      0.53773      0.13392    -0.099882     -0.17189
      0.13392      0.08036    -0.083883     -0.10183
    -0.099882    -0.083883      0.11088      0.12276
     -0.17189     -0.10183      0.12276      0.31007
```

```
S2 = cov(Z2)
S2 =
      0.30273     -0.15407      0.12999     -0.01731
     -0.15407      0.17267     -0.15417    -0.035559
      0.12999     -0.15417      0.17063     0.017776
     -0.01731    -0.035559     0.017776     0.087944

S3 = cov(Z3)
S3 =
       2.1857      0.64038      -0.6119     -0.22378
      0.64038      0.93816     -0.92152     -0.26246
      -0.6119     -0.92152      0.91671      0.27105
     -0.22378     -0.26246      0.27105       0.1619

% MATRIZ VARIÂNCIA COVARIÂNCIA COMBINADA
Sp = ((n1-1)*S1 + (n2-1)*S2 + (n3-1)*S3) / (n1+n2+n3-g)
Sp =
       1.0087      0.20674     -0.19393     -0.13766
      0.20674      0.39706     -0.38652     -0.13328
     -0.19393     -0.38652       0.3994       0.1372
     -0.13766     -0.13328       0.1372      0.18664

% MATRIZ B REPRESENTA A VARIAÇÃO ENTRE GRUPOS
B = n1*(MIX1-MI)'*(MIX1-MI) + n2*(MIX2-MI)'*(MIX2-MI) + n3*(MIX3-
MI)'*(MIX3-MI)
B
       1.7642       5.6344      -5.6564      -6.4924
       5.6344       18.279      -18.216      -20.874
      -5.6564      -18.216       18.216        20.89
      -6.4924      -20.874        20.89       23.961

% MATRIZ W REPRESENTA A VARIAÇÃO ENTRE GRUPOS
W = (n1+n2+n3-g) * Sp
W =
       27.236        5.582      -5.2362      -3.7169
        5.582       10.721      -10.436      -3.5986
      -5.2362      -10.436       10.784       3.7043
      -3.7169      -3.5986       3.7043       5.0392

% AUTOVALORES DE W^-1B
WINV  = inv(W)
WINV =
     0.042916     -0.03508    -0.020574     0.021727
     -0.03508       1.6388        1.573    -0.011871
    -0.020574        1.573       1.6379    -0.095917
     0.021727    -0.011871    -0.095917       0.2765

WINVB = WINV * B
WINVB =
     -0.14663     -0.47817      0.47537      0.54443
      0.35133        1.352      -1.2487      -1.4052
      0.18436      0.80192     -0.70438     -0.78253
      -1.2812       -4.119       4.1222       4.7281

% NÚMERO DE FUNÇÕES DISCRIMINANTES
s = min(g-1,p)
s =
     2
```

```
% AUTOVALORES E AUTOVETORES DE W^-1 B
[Vec,Val] = eig(WINVB)
Vec =
     -0.10825     0.018901     -0.77176     -0.10275
      0.28587      -0.7264     -0.28233      0.12018
      0.16188     -0.68636     -0.56837      0.78106
     -0.93827    -0.029965     0.040459     -0.60411

Val =
       5.1241            0            0            0
            0      0.10509            0            0
            0            0  -8.4499e-19            0
            0            0            0   4.0293e-15

% AUTOVALORES ORDENADOS
l =diag(Val)
l =
       5.1241
      0.10509
  -8.4499e-19
   4.0293e-15
l = l(1:s)
l =
       5.1241
      0.10509

% AUTOVALORES ACUMULADOS
AVA = cumsum(l)
AVA =
       5.1241
       5.2292

% VARIÂNCIA EXPLICADA ACUMULADA PELAS FUNÇÕES DISCRIMINANTES
VE  = AVA / AVA(end) * 100
VE =
        97.99
          100

% CORRELAÇÃO CANÔNICA
r = sqrt(abs(l) ./ (1 + abs(l)))
r =
      0.91472
      0.30838

% TESTE LAMBDA DE WILKS PARA CADA FUNÇÃO DISCRIMINANTE

% WILKS PARA A PRIMEIRA FUNÇAO DISCRIMINANTE
Wilks1 = (1/(1+l(1))) * (1/(1+l(2)))
Wilks1 =
      0.14776

% WILKS PARA AS DEMAIS FUNÇÕES DISCRIMINANTES
Wilks  = 1 ./ (1 + l(2:end))
Wilks =
       0.9049
```

```
% NÚMERO DE GRUPOS
q = g
q =
     3

% NÚMERO DE OBSERVAÇÕES TOTAL
n = n1+n2+n3
n =
    30

% VALORES DE QUI-QUADRADO PARA CADA FUNÇÃO DISCRIMINANTE
Qui2W1 = -(n - 1 - 0.5*(p+q)) * log(Wilks1)
Qui2W1 =
        48.76
Qui2W2 = -(n - 1 - 0.5*(p+q)) * log(Wilks)
Qui2W2 =
       2.5481

% GRAUS DE LIBERDADE PARA CADA FUNÇÃO DISCRIMINANTE
% m CORRESPONDE A ORDEM DO TESTE DE QUI-QUADRADO
m = 1
m =
     1
gl1 = (p-m+1)*(q-m)
gl1 =
     8
m = 2
m =
     2
gl2 = (p-m+1)*(q-m)
gl2 =
     3

% VALORES DE ALFA CORRESPOMDENTE AO QUI-QUADRADO DE CADA FUNÇÃO
DISCRIMINANTE
Alpha1 = 1 - chi2cdf(Qui2W1, gl1)
Alpha1 =
   7.0683e-08
Alpha2 = 1 - chi2cdf(Qui2W2, gl2)
Alpha2 =
      0.46666

% OBTENÇÃO DOS COEFICIENTES DAS FUNÇÕES DISCRIMINANTES
% CÁLCULO DE B0 (SEM n)
B0 = (MIX1-MI)'*(MIX1-MI) + (MIX2-MI)'*(MIX2-MI) + (MIX3-
MI)'*(MIX3-MI)
B0 =

     0.057557      0.20743     -0.19709     -0.22334
      0.20743      0.74756     -0.71029     -0.80491
     -0.19709     -0.71029      0.67487      0.76478
     -0.22334     -0.80491      0.76478      0.86666

% MATRIZ INVERSA DE W MULTIPLIACADA POR B0
WINVB0 = WINV * B0
```

```
WINVB0 =
    -0.014663    -0.047817     0.047537     0.054443
     0.035133       0.1352     -0.12487     -0.14052
     0.018436     0.080192    -0.070438    -0.078253
     -0.12812      -0.4119      0.41222      0.47281

% AUTOVALORES E AUTOVETORES DE WINVB0
[eigVec,eigVal] = eig(WINVB0)
eigVec =

     -0.10825     0.018901      0.85135      0.20852
      0.28587      -0.7264      0.17526    -0.089098
      0.16188     -0.68636     -0.11255     -0.74448
     -0.93827    -0.029965      0.48148      0.62795
eigVal =
      0.51241            0            0            0
            0     0.010509            0            0
            0            0   2.7337e-17            0
            0            0            0    5.363e-16

% AUTO VALORES ORDENADOS
l = diag(eigVal)
l =
      0.51241
     0.010509
   2.7337e-17
    5.363e-16

% AUTOVETORES CORRESPONDENTES AOS AUTOVALORES
e = eigVec
e =
     -0.10825     0.018901      0.85135      0.20852
      0.28587      -0.7264      0.17526    -0.089098
      0.16188     -0.68636     -0.11255     -0.74448
     -0.93827    -0.029965      0.48148      0.62795

% FATOR DE ESCALONAMENTO E AUTOVETORES ESCALONADOS
e1 = e(:,1)
e1 =
     -0.10825
      0.28587
      0.16188
     -0.93827
e2 = e(:,2)
e2 =
     0.018901
      -0.7264
     -0.68636
    -0.029965
k1 = 1/sqrt(e1' * Sp * e1)
k1 =
       2.3628
k2 = 1/sqrt(e2' * Sp * e2)
k2 =
       9.0802
e1escal = k1 * e1
```

```
e1escal =
     -0.25577
      0.67543
      0.38248
      -2.2169
e2escal = k2 * e2
e2escal =
      0.17162
      -6.5959
      -6.2322
     -0.27208
ee = [e1escal e2escal];

% COEFICIENTES DAS FUNÇÕES DISCRIMINANTES NÃO PADRONIZADAS
ee

     -0.25577      0.17162
      0.67543      -6.5959
      0.38248      -6.2322
      -2.2169     -0.27208

% COEFICIENTES DAS FUNÇÕES DISCRIMINANTES PADRONIZADAS
D  = diag(Sp)
D =
       1.0087
      0.39706
       0.3994
      0.18664
DR = sqrt(D)
DR =
       1.0044
      0.63013
      0.63198
      0.43202

eep = DR .* ee
eep =
     -0.25689      0.17237
      0.42561      -4.1562
      0.24172      -3.9387
     -0.95774     -0.11754

% MATRIZ DE ESTRUTURAS
Spd = diag(DR)
Spd =

       1.0044            0            0            0
            0      0.63013            0            0
            0            0      0.63198            0
            0            0            0      0.43202

Rp  = inv(Spd) * Sp * inv(Spd)
Rp =
            1      0.32667     -0.30553     -0.31727
      0.32667            1      -0.9706      -0.4896
     -0.30553      -0.9706            1       0.5025
     -0.31727      -0.4896       0.5025            1
```

```
% MATRIZ DE ESTRUTURA PARA A PRIMEIRA FUNÇÃO DISCRIMINANTE
c1 = e1escal' * Spd
c1 =
     -0.25689      0.42561      0.24172     -0.95774
me1 = Rp * c1'
me1 =
      0.11215
      0.57599
     -0.57415
     -0.96315

% MATRIZ DE ESTRUTURA PARA A SEGUNDA FUNÇÃO DISCRIMINANTE
c2 = e2escal' * Spd
c2 =
      0.17237      -4.1562      -3.9387     -0.11754
me2 = Rp * c2'
me2 =
     0.055335
     -0.21951
     -0.01636
     -0.11652
me = [me1 me2]

% MATRIZ DE ESTRUTURA FINAL
me =
      0.11215     0.055335
      0.57599     -0.21951
     -0.57415     -0.01636
     -0.96315     -0.11652

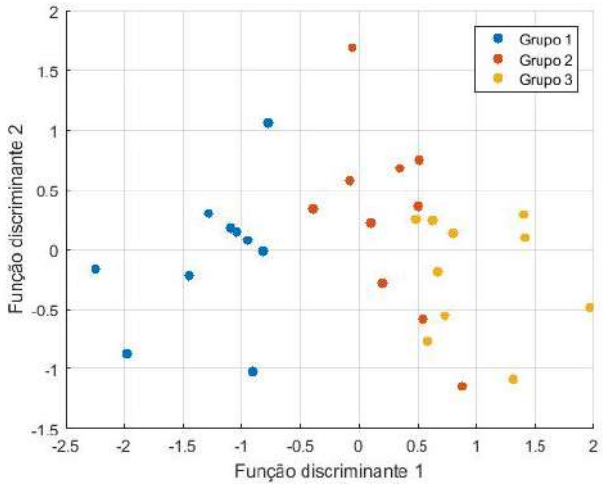
% GRÁFICO DAS FUNÇÕES DISCRIMINANTES PADRONIZADAS

% GRUPO 1
G1y1 = Z1 * eep(:,1)
G1y1 =
      -2.2418
      -1.4475
     -0.95265
      -1.0949
      -1.9719
      -1.2845
     -0.81043
      -1.0475
     -0.76975
     -0.91322
G1y2 = Z1 * eep(:,2)
G1y2 =
     -0.17072
     -0.21627
     0.080108
      0.17554
     -0.87652
      0.30278
    -0.015323
      0.14373
       1.0519
      -1.0262
```

```
% GRUPO 2
G2y1 = Z2 * eep(:,1)
G2y1 =
      0.87633
      0.54761
      0.20041
      0.34415
     0.093411
     -0.38987
      0.51018
    -0.078236
      0.49756
    -0.061354
G2y2 = Z2 * eep(:,2)
G2y2 =
      -1.1444
     -0.58479
     -0.28029
      0.68286
      0.22271
      0.34625
      0.75139
      0.57366
      0.36127
       1.6935

% GRUPO 3
G3y1 = Z3 * eep(:,1)
G3y1 =
       1.9687
       1.4019
      0.72587
      0.62318
      0.67016
      0.58965
       1.4137
      0.48519
        1.314
      0.80138
G3y2 = Z3 * eep(:,2)
G3y2 =
     -0.48105
       0.2887
     -0.55408
      0.23891
     -0.18519
     -0.77444
      0.09858
      0.24661
      -1.0915
      0.14235
scatter(G1y1, G1y2, 'filled')
scatter(G2y1, G2y2, 'filled')
scatter(G3y1, G3y2, 'filled')
figure; hold on
scatter(G1y1, G1y2, 'filled')
scatter(G2y1, G2y2, 'filled')
scatter(G3y1, G3y2, 'filled')
```

```
legend('Grupo 1','Grupo 2','Grupo 3')
xlabel('Função discriminante 1')
ylabel('Função discriminante 2')
grid on
```




```
% CENTRÓIDES DE CADA GRUPO
CG1 = [mean(G1y1) mean(G1y2)]
CG1 =
      -1.2534    -0.055097
CG2 = [mean(G2y1) mean(G2y2)]
CG2 =
      0.25402      0.26221
CG3 = [mean(G3y1) mean(G3y2)]
CG3 =
      0.99938     -0.20711
```

```
                    % SOLUÇÃO DO EXEMPLO 7.3

% OBTENÇÃO DAS INFORMAÇÕES PARA A MATRIZ DE CONFUSÃO

% DISTÂNCIA EUCLIDIANA DO GRUPO 1 AO CENTRÓIDE DO GRUPO 2
dG1C1 = sqrt( (G1y1 - CG1(1)).^2 + (G1y2 - CG1(2)).^2 )
dG1C1 =
      0.99517
      0.25233
       0.3298
      0.27989
       1.0914
      0.35921
      0.44482
      0.28629
        1.208
       1.0289

% DISTÂNCIA EUCLIDIANA DO GRUPO 1 AO CENTRÓIDE DO GRUPO 2
dG1C2 = sqrt( (G1y1 - CG2(1)).^2 + (G1y2 - CG2(2)).^2 )
dG1C2 =
       2.5331
       1.7676
       1.2203
       1.3516
       2.5004
```

```
        1.539
          1.1
       1.3068
        1.293
       1.7386

% DISTÂNCIA EUCLIDIANA DO GRUPO 1 AO CENTRÓIDE DO GRUPO 3
dG1C3 = sqrt( (G1y1 - CG3(1)).^2 + (G1y2 - CG3(2)).^2 )
dG1C3 =
       3.2415
        2.447
        1.973
       2.1289
       3.0459
       2.3401
       1.8199
       2.0767
       2.1714
       2.0807

% MATRIZ DE DISTÂNCIAS DO GRUPO 1 AOS CENTRÓIDES
DG1 = [dG1C1 dG1C2 dG1C3]
DG1 =
      0.99517       2.5331       3.2415
      0.25233       1.7676        2.447
       0.3298       1.2203        1.973
      0.27989       1.3516       2.1289
       1.0914       2.5004       3.0459
      0.35921        1.539       2.3401
      0.44482          1.1       1.8199
      0.28629       1.3068       2.0767
        1.208        1.293       2.1714
       1.0289       1.7386       2.0807

% ÍNDICE DO CENTRÓIDE MAIS PRÓXIMO (CLASSIFICAÇÃO)
[DG1, GrupoClassificado1] = min(DG1, [], 2)
DG1 =
      0.99517
      0.25233
       0.3298
      0.27989
       1.0914
      0.35921
      0.44482
      0.28629
        1.208
       1.0289
GrupoClassificado1 =
     1
     1
     1
     1
     1
     1
     1
     1
     1
     1
```

```
% CONTAGEM DE OCORRÊNCIAS
contagem1 = histcounts(GrupoClassificado1, 1:4)
contagem1 =
    10     0     0

% DISTÂNCIAS EUCLIDIANAS DO GRUPO 2 EM RELAÇÃO AOS CENTROÍDES DE
CADA GRUPO
dG2C1 = sqrt( (G2y1 - CG1(1)).^2 + (G2y2 - CG1(2)).^2 )
dG2C1 =
       2.3922
       1.8773
       1.4712
       1.7597
       1.3751
      0.95216
       1.9393
       1.3328
       1.7998
       2.1163
dG2C2 = sqrt( (G2y1 - CG2(1)).^2 + (G2y2 - CG2(2)).^2 )
dG2C2 =
       1.5382
      0.89642
      0.54506
      0.43015
      0.16541
       0.6494
      0.55222
      0.45541
      0.26291
       1.4657
dG2C3 = sqrt( (G2y1 - CG3(1)).^2 + (G2y2 - CG3(2)).^2 )
dG2C3 =
      0.94544
      0.58886
      0.80232
       1.1052
       1.0028
       1.4955
       1.0761
       1.3308
      0.75823
       2.1766

% MATRIZ DE DISTÂNCIAS
DG2 = [dG2C1 dG2C2 dG2C3]
DG2 =
       2.3922       1.5382      0.94544
       1.8773      0.89642      0.58886
       1.4712      0.54506      0.80232
       1.7597      0.43015       1.1052
       1.3751      0.16541       1.0028
      0.95216       0.6494       1.4955
       1.9393      0.55222       1.0761
       1.3328      0.45541       1.3308
       1.7998      0.26291      0.75823
       2.1163       1.4657       2.1766
```

```
% CLASSIFICAÇÃO
[DG2, GrupoClassificado2] = min(DG2, [], 2)
DG2 =
       0.94544
       0.58886
       0.54506
       0.43015
       0.16541
        0.6494
       0.55222
       0.45541
       0.26291
        1.4657
GrupoClassificado2 =
     3
     3
     2
     2
     2
     2
     2
     2
     2
     2

% CONTAGEM
contagem2 = histcounts(GrupoClassificado2, 1:4)
contagem2 =
     0     8     2

% DISTÂNCIAS EUCLIDIANAS DO GRUPO 3 EM RELAÇÃO AOS CENTROIDES DE
CADA GRUPO
dG3C1 = sqrt( (G3y1 - CG1(1)).^2 + (G3y2 - CG1(2)).^2 )
dG3C1 =
        3.2503
        2.6775
        2.0412
        1.8996
         1.928
        1.9784
        2.6715
        1.7647
        2.7688
        2.0643
dG3C2 = sqrt( (G3y1 - CG2(1)).^2 + (G3y2 - CG2(2)).^2 )
dG3C2 =
         1.869
        1.1482
       0.94287
          0.37
       0.61103
        1.0897
        1.1711
       0.23181
        1.7194
       0.56037
```

```
dG3C3 = sqrt( (G3y1 - CG3(1)).^2 + (G3y2 - CG3(2)).^2 )
dG3C3 =
       1.0074
      0.63858
      0.44186
      0.58354
      0.32999
      0.69996
      0.51473
      0.68569
      0.93867
      0.40167

% MATRIZ DE DISTÂNCIAS
DG3 = [dG3C1 dG3C2 dG3C3]
DG3 =
       3.2503        1.869       1.0074
       2.6775       1.1482      0.63858
       2.0412      0.94287      0.44186
       1.8996         0.37      0.58354
        1.928      0.61103      0.32999
       1.9784       1.0897      0.69996
       2.6715       1.1711      0.51473
       1.7647      0.23181      0.68569
       2.7688       1.7194      0.93867
       2.0643      0.56037      0.40167

% Classificação
[DG3, GrupoClassificado3] = min(DG3, [], 2)
DG3 =
       1.0074
      0.63858
      0.44186
         0.37
      0.32999
      0.69996
      0.51473
      0.23181
      0.93867
      0.40167
GrupoClassificado3 =
     3
     3
     3
     2
     3
     3
     3
     2
     3
     3

% CONTAGEM
contagem3 = histcounts(GrupoClassificado3, 1:4)
contagem3 =
     0     2     8
```

```
% NOVA OBSERVAÇÃO
% CLASSIFICAÇÃO DE UMA NOVA OBSERVAÇÃO
Xnovo = [0.62 76.78 21.62 4614.11]
Xnovo =
         0.62        76.78        21.62       4614.1

% DADOS PADRONIZADOS DA NOVA OBSERVAÇÃO
Znovo = (Xnovo - mediaX) ./ desvX
Znovo =
       1.7532       1.5063      -1.5468     -0.92841

% ESCORES DAS FUNÇÕES DISCRIMINANTES PADRONIZADAS
Ny1 = Znovo * eep(:,1)
Ny1 =
     -0.70609
Ny2 = Znovo * eep(:,2)
Ny2 =
      0.24335

% DISTÂNCIAS DA NOVA OBSERVAÇÃO EM RELAÇÃO AOS CENTRÓIDES DE CADA
GRUPO
dNovoC1 = sqrt( (Ny1 - CG1(1)).^2 + (Ny2 - CG1(2)).^2 )
dNovoC1 =
       1.9821
dNovoC2 = sqrt( (Ny1 - CG2(1)).^2 + (Ny2 - CG2(2)).^2 )
dNovoC2 =
      0.45249
dNovoC3 = sqrt( (Ny1 - CG3(1)).^2 + (Ny2 - CG3(2)).^2 )
dNovoC3 =
      0.53752

% MATRIZ DE DISTÂNCIAS
DNovo = [dNovoC1 dNovoC2 dNovoC3]
DNovo =
       1.9821      0.45249      0.53752

% CLASSIFICAÇÃO DA NOVA OBSERVAÇÃO
[DNovo, GrupoClassificadonovo] = min(DNovo, [], 2)
DNovo =
      0.45249
GrupoClassificadonovo =
     2
```

7.7 ANÁLISE DISCRIMINANTE PARA DUAS POPULAÇÕES NO JULIA

```
#Método 1. Análise discriminante para duas populações
#Solução do exemplo 7.1

using DataFrames, CSV, Statistics, LinearAlgebra, Distributions, Plots;

#Matriz de dados da população 1
X1=[0.47    68.62   28.10   3.29    5433.00
0.47    69.02   27.85   3.14    5235.00
0.48    69.10   27.67   3.23    5092.24
0.51    69.10   27.67   3.23    5092.24
0.53    69.13   28.23   2.64    5312.00
```

```
0.55    69.10   27.67   3.23    5092.24
0.45    69.10   27.67   3.23    5092.24
0.50    69.10   27.67   3.23    5092.24
0.54    71.19   25.52   3.29    4991.00
0.57    70.68   27.19   2.13    5018.00
0.50    69.17   27.44   3.39    5157.00
0.51    69.10   27.67   3.23    5092.24
0.66    69.10   27.67   3.23    5092.24
0.50    67.48   29.47   3.05    5121.00
0.44    69.10   27.67   3.23    5092.24
0.48    68.63   27.51   3.86    5168.00
0.58    69.10   27.67   3.23    5092.24
0.40    69.75   26.33   3.92    5112.00
0.41    70.67   25.99   3.34    4945.00
0.42    70.30   26.14   3.57    5059.00
0.56    69.41   27.19   3.40    5003.00]

#Matriz de dados da população 2
X2=[0.40    75.98   23.08   0.94    4563.00
0.44    77.52   21.45   1.04    4601.00
0.40    75.94   22.69   1.37    4568.00
0.43    75.21   23.40   1.40    4518.57
0.45    75.97   23.14   0.89    4659.00
0.45    75.85   22.89   1.25    4552.00
0.46    75.42   23.44   1.14    4508.00
0.48    77.08   21.93   0.99    4481.00
0.44    77.53   21.37   1.10    4522.00
0.41    77.07   22.05   0.88    4412.00
0.42    76.05   22.47   1.47    4449.00
0.43    75.21   23.40   1.40    4518.57
0.44    75.07   23.59   1.35    4549.00
0.57    75.21   23.40   1.40    4620.00
0.32    73.64   24.73   1.63    4541.00
0.35    75.02   23.58   1.40    4469.00
0.44    75.66   22.92   1.42    4411.00
0.39    75.64   23.38   0.98    4487.00
0.34    75.20   23.32   1.49    4406.00
0.47    77.08   21.88   1.03    4386.00
0.40    78.15   20.71   1.14    4270.00]

#Matriz de dados geral
X=[X1
X2]

# Médias das populações
MI1 = mean(X1,dims=1)
MI2 = mean(X2,dims=1)

# Matriz variância-covariância
S = cov(X)
```

```
# Função Discriminante Linear de Fisher
Y = (MI1 - MI2) * inv(S)

# Médias das populações univariadas
MI1Y = Y * MI1'
MI2Y = Y * MI2'

# Variância das populações univariadas
VY = Y * S * Y'

# Ponto médio entre as populações univariadas
m = (MI1Y + MI2Y) / 2

#Gráfico das funções univariadas
x = -2130:0.01:-2090
a=MI1Y[1,1]
b=MI2Y[1,1]
v=VY[1,1]

y1 = (1 / (sqrt(v) * sqrt(2*pi))) .*
     exp.(-0.5 .* (((x .- (a)) ./ sqrt(v)).^2))

y2 = (1 / (sqrt(v) * sqrt(2*pi))) .*
     exp.(-0.5 .* (((x .- b) ./ sqrt(v)).^2))

plot(x, y1, label="População 1")
plot!(x, y2, label="População 2")

#Probabilidade da classificação errada
z=(m- MI1Y)/sqrt(v)
AREA=cdf(Normal(0,1),z)
p=AREA*2

#Alocação de uma nova observação
X0=[0.45    68.3    26.4    3.2 5100.41]
Y0=Y*X0'
```

7.8 ANÁLISE DISCRIMINANTE PARA VÁRIAS POPULAÇÕES NO JULIA

```
#Método 2. Análise discriminate para várias populações
#Solução do exemplo 7.2
using DataFrames, CSV, Statistics, LinearAlgebra, Distributions, Plots,
StatsBase;
#Valores das variáveis para cada grupo
X1=[0.47    68.62   28.10   5433.00
0.47    69.02   27.85   5235.00
0.48    69.10   27.67   5092.24
0.51    69.10   27.67   5092.24
0.53    69.13   28.23   5312.00
0.55    69.10   27.67   5092.24
0.45    69.10   27.67   5092.24
```

```
0.50    69.10   27.67   5092.24
0.54    71.19   25.52   4991.00
0.57    70.68   27.19   5018.00]

X2=[0.49    73.44   24.92   4694.00
0.51    74.27   23.93   4759.00
0.49    74.46   23.53   4870.00
0.57    72.28   24.97   4691.00
0.56    72.37   25.12   4776.00
0.55    70.23   26.77   4878.00
0.52    73.49   23.83   4724.00
0.56    72.39   24.89   4815.00
0.54    73.16   24.37   4705.00
0.51    72.62   23.95   4855.00]

X3=[0.56    80.99   18.50   4434.00
0.58    74.21   23.69   4442.00
0.51    73.77   24.34   4705.00
0.64    76.84   21.61   4613.00
0.54    74.01   23.99   4682.00
0.52    72.19   25.79   4708.00
0.34    71.55   25.49   4683.00
0.59    73.32   24.41   4653.83
0.52    77.53   21.59   4607.00
0.58    74.73   23.30   4607.00]

#Matriz de dados geral
X=[
    X1
    X2
    X3
]

#Número de grupos
g=3

#Matriz de dados padronizados
Z1 = (X1 .- mean(X, dims=1)) ./ std(X, dims=1)
Z2 = (X2 .- mean(X, dims=1)) ./ std(X, dims=1)
Z3 = (X3 .- mean(X, dims=1)) ./ std(X, dims=1)

#Vetores médios dos grupos
MIX1=mean(Z1, dims=1)
MIX2=mean(Z2, dims=1)
MIX3=mean(Z3,dims=1)

#Tamanho das amostras de cada grupo
(n1,p)=size(X1)
(n2,p)=size(X2)
(n3,p)=size(X3)
```

```
#Vetor médio geral
MI=(n1*MIX1+n2*MIX2+n3*MIX3)/(n1+n2+n3)

#Matrizes de variância-covariância
S1=cov(Z1)
S2=cov(Z2)
S3=cov(Z3)

#Matriz variância-covariância combinada
Sp=((n1-1)*S1+(n2-1)*S2+(n3-1)*S3)/(n1+n2+n3-g)

#Matriz B, representa a variação entre grupos
B=(n1*((MIX1-MI)'*(MIX1-MI)))+(n2*((MIX2-MI)'*(MIX2-MI)))+(n3*((MIX3-
MI)'*(MIX3-MI)))

#Matriz W, representa a variação dentro dos grupos
W=(n1+n2+n3-g)*Sp

#Autovalores e autovetores de WinvB
#Matriz inversa de W e WinvB
WINV=inv(W)
WINVB=WINV*B

#Número de funções discriminantes
s=min(g-1,p)

#Autovalores e autovetores de WINVB
AV = eigen(WINVB)

#Autovalores ordenados
l = reverse(AV.values, dims=1)
l=l[1:s]

#Autovalores acumulados
AVA = cumsum(l)

#Variância explicada acumulada pelas funções discriminantes
VE = AVA ./ AVA[end]*100

#Correlação canônica para cada função discriminante
r=(l./([1].+l)).^0.5

#Teste Lambda de Wilks para cada função discriminante
#Wilks para a primeira função discriminante
Wilks1=(1/(1+l[1,1]))*(1/(1+l[2,1]))

#Wilks para as demais funções discriminantes
Wilks=[1]./([1].+l[2,end])

#Número de grupos
q=g
```

```
#Número de observações total
n=n1+n2+n3

#Valores de Qui-quadrado para cada função discriminante
Qui2W1=-(n-1-0.5*(p+q))*log(Wilks1)
Qui2W2=-(n-1-0.5*(p+q))*log.(Wilks)

#Graus de liberdade de cada função discriminante
#m corresponde a ordem do teste de Qui-quadrado
m=1
gl1=(p-m+1)*(q-m)
m=2
gl2=(p-m+1)*(q-m)

#Valores de alfa correspondentes ao Qui-quadrado de cada função discriminante
Alpha1 = ccdf(Chisq(gl1), Qui2W1)
Alpha2 = ccdf(Chisq(gl2), Qui2W2)

#Obtenção dos coeficientes das funções discriminantes

#Matriz B0, sem n
B0=(((MIX1-MI)'*(MIX1-MI)))+(((MIX2-MI)'*(MIX2-MI)))+(((MIX3-MI)'*(MIX3-MI)))

#Matriz inversa de W multiplicada por B0
WINVB0=WINV*B0

#Autovalores e autovetores de WINVB0
AV = eigen(WINVB0)

#Autovalores ordenados
l = reverse(AV.values, dims=1)

#Autovetores correspondentes aos autovalores
e = reverse(AV.vectors, dims=2)

#Fator de escalonamento e autovetores escalonados
e1=e[:,1]
e2=e[:,2]
k1=1/sqrt((e1)'*Sp*e1)
e1escalonado=k1*e1
k2=1/sqrt((e2)'*Sp*e2)
e2escalonado=k2*e2

#Coeficientes das funções discriminantes não padronizadas
ee=[e1escalonado e2escalonado]

#Coeficientes das funções discriminantes padronizados
Sp
ee
D=diag(Sp)
DR=sqrt.(D)
```

```
eep=DR.*ee

#Matriz de estruturas
Spd=diagm(DR)
Rp=inv(Spd)*Sp*inv(Spd) #Matriz de correlação

#Matriz de estruturas para a primeira função discriminante
ee1=ee[:,1]
c1=ee1'*Spd
me1=Rp*c1'

#Matriz de estruturas para a segunda função discriminante
ee2=ee[:,2]
c2=ee2'*Spd
me2=Rp*c2'

#Matriz de estruturas final
me=[me1 me2]

#Gráfico das funções discriminantes padronizadas
#Grupo 1
G1y1= Z1 * eep[:,1]
G1y2= Z1 * eep[:,2]
scatter(G1y1, G1y2, label="Grupo 1")

#Grupo 2
G2y1= Z2 * eep[:,1]
G2y2= Z2 * eep[:,2]
scatter!(G2y1, G2y2, label="Grupo 2")

#Grupo 3
G3y1= Z3 * eep[:,1]
G3y2= Z3 * eep[:,2]
scatter!(G3y1, G3y2, label="Grupo 3")

#Centróide grupo 1
CG1y1=mean(G1y1)
CG1y2=mean(G1y2)
CG1=[CG1y1 CG1y2]

#Centróide grupo 2
CG2y1=mean(G2y1)
CG2y2=mean(G2y2)
CG2=[CG2y1 CG2y2]

#Centróide grupo 3
CG3y1=mean(G3y1)
CG3y2=mean(G3y2)
CG3=[CG3y1 CG3y2]
```

```
#Solução do exemplo 7.3
#Obtencão das informações para a matriz de confusão
#Distância euclidiana do grupo 1 em relação ao centróide do grupo 1
dG1C1=sqrt.((G1y1.-CG1y1).^2+(G1y2.-CG1y2).^2)
#Distância euclidiana do grupo 1 em relação ao centróide do grupo 2
dG1C2=sqrt.((G1y1.-CG2y1).^2+(G1y2.-CG2y2).^2)
#Distância euclidiana do grupo 1 em relação ao centróide do grupo 3
dG1C3=sqrt.((G1y1.-CG3y1).^2+(G1y2.-CG3y2).^2)

#matriz de distância do grupo 1 em relação aos centróides
DG1=[dG1C1 dG1C2 dG1C3]
# Índice do centróide mais próximo
col_min_G1 = mapslices(DG1 -> argmin(DG1), DG1; dims = 2)
# Converter para vetor
GrupoClassificado1 = vec(col_min_G1)
# Contagem de ocorrências para o grupo 1
contagem1 = countmap(col_min_G1)

#Distância euclidiana do grupo 2 em relação ao centróide do grupo 1
dG2C1=sqrt.((G2y1.-CG1y1).^2+(G2y2.-CG1y2).^2)
#Distância euclidiana do grupo 2 em relação ao centróide do grupo 2
dG2C2=sqrt.((G2y1.-CG2y1).^2+(G2y2.-CG2y2).^2)
#Distância euclidiana do grupo 2 em relação ao centróide do grupo 3
dG2C3=sqrt.((G2y1.-CG3y1).^2+(G2y2.-CG3y2).^2)

#matriz de distância do grupo 2 em relação aos centróides
DG2=[dG2C1 dG2C2 dG2C3]
#Índice do centróide mais próximo
col_min_G2 = mapslices(DG2 -> argmin(DG2), DG2; dims = 2)
# Converter para vetor
GrupoClassificado2 = vec(col_min_G2)
# Contagem de ocorrências para o grupo 2
contagem2 = countmap(col_min_G2)

#Distância euclidiana do grupo 3 em relação ao centróide do grupo 1
dG3C1=sqrt.((G3y1.-CG1y1).^2+(G3y2.-CG1y2).^2)
#Distância euclidiana do grupo 3 em relação ao centróide do grupo 2
dG3C2=sqrt.((G3y1.-CG2y1).^2+(G3y2.-CG2y2).^2)
#Distância euclidiana do grupo 3 em relação ao centróide do grupo 3
dG3C3=sqrt.((G3y1.-CG3y1).^2+(G3y2.-CG3y2).^2)

#matriz de distância do grupo 3 em relação aos centróides
DG3=[dG3C1 dG3C2 dG3C3]
#Índice do centróide mais próximo
col_min_G3 = mapslices(DG3 -> argmin(DG3), DG3; dims = 2)
# Converter para vetor
GrupoClassificado3 = vec(col_min_G3)
# Contagem de ocorrências para o grupo 3
contagem3 = countmap(col_min_G3)
```

```
#Classificação de uma nova observação
Xnovo=[0.62 76.78 21.62 4614.11]
#Dados padronizados da nova observação
Znovo = (Xnovo .- mean(X, dims=1)) ./ std(X, dims=1)

#Escores das funções discriminantes padronizadas
Ny1= Znovo * eep[:,1]
Ny2= Znovo * eep[:,2]

#Distância euclidiana da nova observação em relação ao centróide do grupo 1
dNovoC1=sqrt.((Ny1.-CG1y1).^2+(Ny2.-CG1y2).^2)
#Distância euclidiana da nova observação em relação ao centróide do grupo 2
dNovoC2=sqrt.((Ny1.-CG2y1).^2+(Ny2.-CG2y2).^2)
#Distância euclidiana da nova observação em relação ao centróide do grupo 3
dNovoC3=sqrt.((Ny1.-CG3y1).^2+(Ny2.-CG3y2).^2)

#matriz de distância da nova observação em relação aos centróides
DNovo=[dNovoC1 dNovoC2 dNovoC3]
# Índice do centróide mais próximo
col_min_novo = mapslices(DNovo -> argmin(DNovo), DNovo; dims = 2)
# Converter para vetor
GrupoClassificadonovo = vec(col_min_novo)
```

8 ANÁLISE DE CONGLOMERADOS

8.1 INTRODUÇÃO

A análise discriminante é uma técnica utilizada quando o número de grupos é conhecido, tendo como principal objetivo alocar corretamente uma nova observação a um desses grupos e determinar quais variáveis são importantes na separação dos grupos. A análise de agrupamentos é aplicada quando os grupos são desconhecidos, sendo uma técnica voltada à identificação de grupos com base em medidas de similaridade ou distância entre as observações. Esse método busca identificar a estrutura natural existente nos dados, permitindo a definição de grupos homogêneos. A análise de agrupamentos reúne indivíduos ou objetos de modo a maximizar a homogeneidade dentro dos grupos e a heterogeneidade entre eles, sendo fundamentado nas características que esses indivíduos ou objetos possuem.

A análise de conglomerados (*cluster analysis*) envolve diversas técnicas com o objetivo de descobrir a estrutura dos dados (objetos, pessoas, casos), sendo cada observação descrita por variáveis selecionadas representadas por números (atributos, características, medições). O objetivo é agrupar tanto os dados quanto as variáveis em conglomerados (grupos ou clusters), tal que os elementos dentro de um grupo tenham muita associação natural entre eles, enquanto os grupos são distintos entre si. Os procedimentos para a solução do problema e os resultados alcançados dependem de como o pesquisador escolhe o significado operacional para os conceitos associação natural e relativamente distintos.

Quais seriam então as principais maneiras para a realização de uma análise de conglomerados? Pelo menos nove principais passos existem em agrupar um conjunto de dados, sendo que todos podem mudar o resultado profundamente. Esses passos têm como objetivo principal definir a estrutura dos dados, alocando-os em grupos mais semelhantes entre si. A formação dos grupos é efetuada com base em medidas de similaridade ou dissimilaridade entre as observações, permitindo identificar padrões e relações naturais existentes nos dados.

8.1.1 Escolha das unidades dos dados

Os objetos das análises podem ser pessoas, espécies florestais, comodities, opiniões ou outras entidades. As análises são realizadas em um conjunto de amostras destas entidades, cada membro sendo denominado como uma unidade de dados, objetos, observações, casos, elementos ou eventos.

Na escolha das unidades de dados existem duas situações diferentes de interesse. No primeiro caso a amostra é o objeto completo da análise, na qual nenhuma unidade de dados será omitida. No segundo caso, a amostra é parte de uma população maior e processos estatísticos são necessários, porém, em análise de conglomerados os testes de hipóteses não são envolvidos, mas os seus princípios devem ser considerados.

8.1.2 Escolha das variáveis

As unidades de dados não podem ser dispostas de qualquer maneira. Elas precisam ser consistentemente descritas em termos de suas características ou atributos a que classe pertence e constituem as variáveis do problema. A escolha das variáveis tem influência no resultado da análise de conglomerados.

Quando uma variável muito importante não é considerada nas análises, alguns conglomerados podem se juntar de forma errônea. Por outro lado, a inclusão de variáveis muito discriminantes, mas não relevantes ao objetivo proposto, poderá gerar resultados satisfatórios da análise, mas sem utilidade prática.

8.1.3 Agrupamento de observações e/ou variáveis

Até este ponto o objetivo é agrupar unidades de dados (observações) por meio de variáveis. Também é razoável agrupar variáveis de acordo com o seu mútuo comportamento nas unidades de dados. Existem métodos (Fischer, 1936; Hartigan, 1972) para simultaneamente agrupar (conglomerar) unidades de dados e variáveis. Ainda, é possível considerar um procedimento iterativo, em que o agrupamento é alternado entre variáveis e unidades de dados, até que classificações harmônicas mutuamente sejam alcançadas para ambos.

Agrupar variáveis tem muito em comum com análise fatorial (*factor analysis*). As duas metodologias objetivam explorar as relações entre variáveis, apesar de usar técnicas completamente diferentes. Devido a estas diferenças, espera-se encontrar resultados divergentes que complementam as duas análises e propiciem uma excelente visão quando usadas juntamente.

8.1.4 Homogeneização das variáveis

Um dos problemas encontrados entre as variáveis é que elas não são homogêneas. Isto acontece devido ao fato de que as unidades se dispõem em diferentes tipos de escalas na maioria das vezes e, isto, representa um sério problema para a análise de agrupamentos. Se uma variável está na sua forma linear e quadrática, a variável na forma quadrática assumirá um peso maior. A solução para o efeito da escala é a padronização dos valores de todas as variáveis. A padronização e normalização são dois procedimentos matemáticos diferentes.

- Normalização

Vetores são considerados normalizados se a soma dos elementos ao quadrado é igual a 1, incluindo linhas ou colunas de vetores individuais das matrizes. Outro ponto a ser ressaltado é que o termo normalização não tem relação com o termo distribuição normal.

A normalização de um vetor é dito ser de comprimento unitário, em que o termo comprimento se refere à representação geométrica do vetor e não ao número dos seus elementos. Algebricamente, o comprimento de um vetor é a raiz quadrada da soma de seus elementos ao quadrado. Se A é o comprimento do vetor X, que contêm n elementos, então:

$$A = (X_n' X_n)^{\frac{1}{2}} \text{ (em termos matriciais)}$$

Para normalizar o vetor de escores originais (X), cada um deles será dividido por A. O processo completo que produz um vetor normalizado Y é descrito pela seguinte fórmula:

$$Y_n = (X_n'\, X_n)^{-\frac{1}{2}}\, X_n = A^{-1}\, X_n$$

$$\begin{matrix} X_1' & X_1 & X_1'X_1 \\ [6 & 8] & \begin{bmatrix}6\\8\end{bmatrix} & = 100 \end{matrix} \qquad \begin{matrix} (X_1'\, X_1)^{-\frac{1}{2}} & X_1 & Y_1 \\ [1/10] & \begin{bmatrix}6\\8\end{bmatrix} & = \begin{bmatrix}0{,}6\\0{,}8\end{bmatrix} \end{matrix}$$

$$\begin{matrix} Y_1' & Y_1 & Y_1 \\ [0{,}6 & 0{,}8] & \begin{bmatrix}0{,}6\\0{,}8\end{bmatrix} & = 1 \end{matrix}$$

A sequência da normalização geral para todas as linhas ou todas as colunas de uma matriz simultaneamente é a seguinte:

- para normalizar as colunas de uma matriz:

$$Y_{nk} = X_{nk}(X'_{kn}\, X_{nk})^{-1/2}_{\Delta k}$$

- para normalizar as linhas de uma matriz:

$$Y_{nk} = (X_{nk}X'_{kn}\,)^{-1/2}_{\Delta n} X_{nk}$$

- Padronização

Padronização é a conversão dos escores para a distribuição z pela fórmula:

$$z = \frac{(x-\mu)}{\sigma}$$

Em que: μ é a média dos escores e σ é o seu desvio padrão.

A conversão de um vetor de escores originais para a forma de escores, z (padronizados), muda o comprimento do vetor para a raiz quadrada do número de seus elementos $\left(N^{1/2}\right)$ e, também, muda a soma dos elementos para zero. Assim, um conjunto de dados padronizados é normalmente distribuído com média zero e variância igual a 1 $(N{\sim}0{,}1)$. A aplicação tanto da normalização como a padronização serão utilizadas em quase todas as técnicas multivariadas.

8.1.5 Medidas de similaridade

A maioria dos métodos de análise de agrupamentos exige uma medida de similaridade, definida para cada par de combinações das observações a serem agrupadas. Quando se agrupa unidades de dados, a proximidade dos indivíduos é usualmente obtida, utilizando a distância Euclidiana.

Agrupar variáveis geralmente envolve o uso da correlação ou outras medidas de associação. As medidas interagem com os critérios da análise de agrupamento, sendo que algumas medidas produzem resultados idênticos com um critério e resultados completamente diferentes com outros. A escolha combinada de variáveis, transformações e medidas de similaridade resulta na obtenção de significado operacional ao termo associação natural.

As medidas de similaridade e dissimilaridade são fundamentais na análise de agrupamentos. No agrupamento de itens, como indivíduos ou objetos, a proximidade entre eles é geralmente indicada por uma medida de distância. Já no agrupamento de variáveis, estas são reunidas com base em medidas de correlação ou associação entre si.

A similaridade indica o quanto os elementos são similares, sendo que valores maiores representam maior semelhança, como ocorre no coeficiente de correlação. Por outro lado, a dissimilaridade expressa o quanto os elementos são distintos, de modo que valores mais elevados indicam maior diferença entre eles, sendo a distância euclidiana um exemplo amplamente utilizado.

Embora a correlação possa representar uma forma de similaridade, ela não é a medida mais comumente empregada na análise de agrupamentos. De modo geral, as medidas de dissimilaridade baseadas em distâncias são as mais utilizadas, pois valores de distância maiores indicam menor similaridade entre os elementos. Além disso, a distância pode ser convertida em uma medida de similaridade por meio do uso de sua relação inversa.

Existem vários tipos de distâncias. Estas são sensíveis à escala ou à magnitude das variáveis, o que torna necessária a padronização dos dados antes da aplicação da análise de agrupamentos. A padronização consiste em converter cada variável em escores z, garantindo que todas contribuam de forma equilibrada para o cálculo das distâncias.

Distância euclidiana

Uma função de distância comum é a distância euclidiana entre dois vetores $x = [\, x_1 \quad x_2 \quad \cdots \quad x_p \,]'$ e $y = [\, y_1 \quad y_2 \quad \cdots \quad y_p \,]'$ definida como:

$$d\left(\underline{x},\underline{y}\right) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

A maioria das técnicas de análise de conglomerados começa com o cálculo da matriz de similaridades ou distâncias entre as entidades e, consequentemente, são necessárias considerações sobre as possíveis maneiras de se definir estas quantidades. Por exemplo, um coeficiente de similaridade comum é o coeficiente de correlação e, talvez, a medida de distância mais comum é a distância Euclidiana.

Em termos mais simples, a distância entre dois pontos, dada as suas coordenadas, é uma aplicação do teorema de Pitágoras, que expressa: o comprimento da hipotenusa de um triângulo retângulo é igual a raiz quadrada da soma dos quadrados dos outros dois lados do triângulo. Então, se as coordenadas dos dois pontos são (x_1, y_1) e (x_2, y_2) (Figura 8.1), a distância entre eles é:

$$d_{i=1,j=2} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$$
$$d_{i=1,j=2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

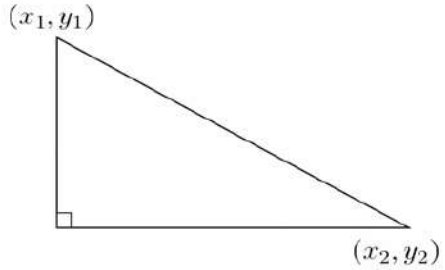


Figura 8.1: A distância entre dois pontos.

A distância Euclidiana usada nos dados originais pode ser insatisfatória, pois é fortemente afetada pela mudança de escala da variável. Por exemplo, suponha três entidades A, B e C, que foram cada uma medida em duas variáveis, sendo peso (em quilos) e altura (em metros) com os seguintes resultados:

Entidade	Peso em quilos	Altura em metros	Altura em centímetro
A	60	3,0	36
B	65	3,5	42
C	63	4,0	48

As distâncias Euclidianas são:

$$D_{AB}^2 = 25{,}25$$
$$D_{AC}^2 = 10{,}00$$
$$D_{BC}^2 = 4{,}25$$

Contudo, se a altura é medida em centímetros estas distâncias vem a ser:

$$D_{AB}^2 = 61$$
$$D_{AC}^2 = 153$$
$$D_{BC}^2 = 40$$

A entidade A está agora mais próxima da entidade B do que da entidade C. Desta forma, mesmo quando todas as variáveis são exclusivamente determinadas, exceto para mudanças de escala, a distância Euclidiana não preservará a ordem das distâncias. Devido

a isto, as variáveis são frequentemente padronizadas antes de se empregar a distância Euclidiana em tomando-se $z_{ik} = \frac{x_{ik}}{\sigma_k}$, em que σ_k é o desvio padrão da *kth* variável. A distância Euclidiana calculada para as variáveis padronizadas preservará as distâncias relativas.

Isto é facilmente demonstrado calculando-se a média e os desvios das variáveis, como segue:

Entidade	Peso em quilos	Altura em metros	Altura em centímetro
$\bar{x}$	62,67	3,50	42
s	2,52	0,50	6,0

A seguir, padroniza-se as variáveis utilizando-se a fórmula $z_{ik} = \frac{x_{ik}}{\sigma_k}$, em que:

Entidade	Entidades Padronizadas		
	Peso em quilos	Altura em metros	Altura em centímetro
A	23,81	6,0	6,0
B	25,79	7,0	7,0
C	25,00	8,0	8,0

As distâncias Euclidianas, como se observa, não se alterarão, pois a altura em metros e em centímetros na forma padronizada não se alteram e preservam as distâncias na forma original (dados não padronizados). Desta forma, a mudança de escala na forma padronizada não altera o valor relativo da distância Euclidiana.

Um outro ponto a ser considerado, refere-se ao uso de distâncias como medida de dissimilaridade.

Se coordenadas num diagrama $p-$ dimensional representam os valores de p variáveis medidas em N indivíduos, é óbvio, em termos gerais, que os pontos que estão bem próximos uns dos outros representam indivíduos que são similares nestes caracteres, enquanto pontos que estão distantes representam indivíduos que são dissimilares ou diferentes. Então, distância num diagrama desta natureza é uma medida de dissimilaridade.

A distância Euclidiana, como mostrado acima, é usualmente usada como uma medida de dissimilaridade, somente quando todas as médias são do mesmo tipo, ou quando elas foram padronizadas de alguma maneira.

Marriot (1974) define a medida de similaridade como sendo uma quantidade não-negativa que aumenta com a similaridade entre dois indivíduos. A medida de similaridade entre A e B pode ser escrita $s(AB)$. Do mesmo modo, a medida de dissimilaridade $d(AB)$ é uma medida não negativa que diminui quando $s(AB)$ aumenta. Então, $\frac{1}{s(AB)}$ e $C - s(AB)$, em que C não é menor do que qualquer valor de $s(AB)$, são possíveis medidas de dissimilaridade.

Se uma medida de dissimilaridade for considerada como distância, então precisa ter propriedades métricas, $d(AB) \leq d(AC) + d(BC)$, para quaisquer itens A, B e C.

Distância de Mahalanobis

Para ajustar variâncias e covariâncias diferentes entre as p variáveis, quando os dados não são padronizados, pode se utilizar a distância de Mahalanobis:

$$d\left(\underline{x}, \underline{y}\right) = \sqrt{\left(\underline{x} - \underline{y}\right)' S^{-1} \left(\underline{x} - \underline{y}\right)}$$

Em que: S é a matriz de variância-covariância.

A distância Mahalanobis em sua forma quadrática (D^2) apresenta vantagem sobre a distância Euclidiana, pois permite obter as correlações entre as variáveis. Quando estas são zero, é equivalente à distância Euclidiana medida usando variáveis padronizadas.

Para finalizar, deve-se traçar um paralelo, concernente às diferenças entre as distâncias de similaridade e distâncias mensuráveis.

A diferença mais óbvia entre estes dois tipos de medias é que, enquanto as similaridades resultam em valores entre 0 e 1, as mensuráveis podem resultar em qualquer valor positivo. Contudo, pode-se facilmente transformar o conjunto de valores da função distância para um conjunto correspondente de valores da função similaridade, aplicando-se a seguinte transformação:

$$s_{ij} = \frac{1}{\left(1 + d_{ij}\right)}$$

O processo reverso é muito mais difícil, devido a necessidade de as distâncias atenderem à desigualdade do triângulo, e nisto reside a diferença essencial entre os dois tipos de medidas. Contudo, métodos foram propostos para obtenção da representação métrica das estruturas de similaridade.

Um dos métodos, proposto por Gower (1966, 1967), mostra que:

$$d_{ij} = \sqrt{2\left(1 - s_{ij}\right)}$$

pode funcionar como distância Euclidiana, desde que a matriz de coeficientes de similaridade, s, seja positiva e semi-definida. As coordenadas dos pontos com estas distâncias são elementos dos vetores latentes de s, escalonados, sendo que suas somas de quadrados se igualam às raízes latentes. Gower também demonstra que é possível usar o método dos componentes principais nestas coordenadas para encontrar o melhor ajuste em menos dimensões. Como é usualmente observado nos componentes principais, uma boa representação do conjunto de pontos pode ser obtida em um número reduzido de dimensões, quando algumas das raízes latentes são pequenas. Este método de análise é conhecido como coordenadas principais.

Quadrado da distância euclidiana

Algumas vezes a distância euclidiana é usada ao quadrado, resultando numa matriz simétrica com elementos na diagonal iguais a zero. A distância é calculada como:

$$d\left(\underline{x}, \underline{y}\right) = \sum_{i=1}^{p} (x_i - y_i)^2$$

Métrica de Minkowski

Outra medida de distância sugerida é, por exemplo, a métrica de Minkowski:

$$d\left(\underline{x}, \underline{y}\right) = \sqrt[n]{\sum_{i=1}^{p} |x_i - y_i|^n}$$

Quando $n = 2$ essa distância torna-se a distância euclidiana.

Distância city-blok (Manhattan)

Quando $p = 2$ e $n = 1$, a distância métrica de Minkowski torna-se a distância *city-blok* entre duas observações:

$$d\left(\underline{x},\underline{y}\right)=\sum_{i=1}^{p}|x_i-y_i|$$

8.1.6 Critério de agrupamento

Para encontrar grupos em dados reais com significado é necessário ser claro e objetivo. A escolha de um critério de agrupamento é importante para definir um conglomerado.

Muitos critérios de agrupamento foram propostos e usados. Em algumas situações existe uma escolha natural, enquanto em outros vários critérios podem ser candidatos. Não é necessário escolher somente um critério, pois conglomerar um conjunto de dados várias vezes com diferentes critérios é uma boa estratégia para revelar as várias faces da estrutura.

8.1.7 Algoritmos e implementação

Tendo sido feita a escolha das unidades de dados, variáveis, o que agrupar, transformações, medida de similaridade e um critério para tanto, ainda permanece o problema de se gerar o conjunto de conglomerados. Mesmo que estes critérios tenham sido definidos antecipadamente, existem vários algoritmos que podem fornecer os mesmos resultados.

A sequência das escolhas aparenta mostrar alguma lógica. A escolha do algoritmo que implementará os agrupamentos é realizada a seguir e representa somente um passo na condução da análise de agrupamentos.

8.1.8 Número de conglomerados

Um problema prático de grande importância, quando se realiza análise de agrupamentos, é a decisão quanto ao número de grupos a serem gerados. Métodos de agrupamentos hierárquicos fornecem uma configuração para cada número de conglomerados de um (todo o conjunto de dados) até o número de entidades (cada grupo possui somente um membro). Outros algoritmos fornecem a melhor estrutura de agrupamento para um número de grupos previamente definidos, que podem ser repetidos para diferentes números de conglomerados, resultando em uma variedade de alternativas.

Alguns algoritmos iniciam com a escolha do número de grupos e, então, variam este número como indicado por meio critérios, com o objetivo de simultaneamente determinar tanto o número de conglomerados como as suas configurações.

8.1.9 Interpretação dos resultados

Uma vez que o conjunto de conglomerados é encontrado, apresenta-se o momento de se fazer uso dos resultados. Em nível menos sofisticado os conglomerados são o resumo estatístico descritivo, semelhante à média e a variância.

Para uma dada medida de similaridade, critério e algoritmo, a obtenção dos conglomerados pode ser um procedimento mecânico direto. O foco principal dos algoritmos e critérios de agrupamentos é fornecer ao conjunto de unidades uma diferenciação entre elas. Resultados significantes deverão ocorrer com regularidade.

8.2 MÉTODOS DE AGRUPAMENTOS

8.2.1 Métodos de agrupamentos hierárquicos

As medidas de associação são ou podem ser usadas para construir a matriz de similaridade, que descreve a associação de todos os pares dentro de um conjunto de dados, sejam variáveis ou unidades de dados.

Os métodos de análise de agrupamentos hierárquicos usam a matriz de similaridade para elaborar um dendrograma (árvore), que apresenta as relações específicas entre as entidades. O tipo de algoritmo empregado não usa a estrutura interna dos grupos formados em cada estágio, mas simplesmente trata o conjunto de grupos como um novo conjunto de objetos (Kirchner, 1980).

Os algoritmos de agrupamento hierárquico envolvem um processo sequencial. Em cada passo deste procedimento uma observação ou grupo de observações é unido em outro grupo. Neste processo o número de grupos diminui e os próprios clusters tornam-se maiores. O início se dá a partir de n itens individuais e termina com um grupo contendo todo o conjunto de dados. Uma abordagem alternativa, chamada de método divisivo, começa com um único grupo contendo os n itens e o particiona em dois grupos a cada passo. O resultado desta abordagem é a obtenção de n grupos de um item cada.

O método de agrupamento hierárquico inicia com tantos grupos quantos os itens existentes, sejam indivíduos ou objetos, formando inicialmente grupos unitários. A partir desse ponto, os grupos são estabelecidos com base na similaridade entre os itens, e o processo segue até que todos os elementos estejam reunidos em um único grupo.

No procedimento inicial, considera-se a existência de n grupos, cada um formado por um único item. Em seguida, calcula-se a matriz simétrica de distâncias $n\,x\,n$, representada por $D = \left(d_{ij}\right)$, na qual os elementos da diagonal principal são iguais a zero. Cada elemento d_{ij} corresponde à distância ou similaridade entre o item i e o item j.

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Em que: $d_{11} = d_{22} = \cdots = d_{nn} = 0$; d_{ij} é a distância ou similaridade entre o item i e objeto j.

A partir da matriz D, identifica-se o par de grupos mais próximos, isto é, aquele que apresenta a menor distância, e eles são então reunidos. O novo grupo formado é denominado (AB), quando os grupos originais são A e B. Em seguida, constrói-se uma nova matriz de distâncias, eliminando-se as linhas e colunas correspondentes aos grupos A e B e adicionando-se uma nova linha e coluna, que representam as distâncias entre o grupo (AB) e os grupos remanescentes. Os passos de identificação do par mais próximo e atualização da matriz de distâncias são repetidos sucessivamente $(n - 1)$ vezes, registrando-se as identidades dos grupos formados ao longo do processo, até que todos os itens estejam reunidos em um único grupo. Após isso, as informações são resumidas em um dendrograma (Figura 8.2).

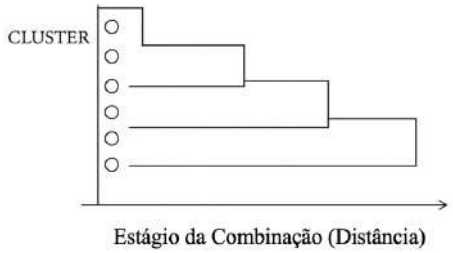


Figura 8.2: Exemplo de dendrograma.

Entre os algoritmos aglomerativos para análise hierárquica de conglomerados destacam-se as apresentadas na sequência

- Ligações simples ou vizinho mais próximo (Single-Linkage)

Agrupa os dois objetos mais próximos para formar o primeiro grupo e em seguida os objetos mais próximos a este grupo para formar um novo conglomerado, e assim sucessivamente. Se os dois objetos a serem unidos caem em conglomerados diferentes obtidos nos passos anteriores, os dois conglomerados são então unidos. O termo união-simples (*single-linkage*) é usado, porque dois conglomerados são unidos se qualquer das distâncias entre os objetos nos diferentes conglomerados é suficientemente pequena, isto é, se houver uma simples união entre os conglomerados.

No método de ligação simples a distância entre dois grupos A e B é definida como a distância entre um ponto de A e um ponto de B:

$$d_{(A,B)} = mínimo\{d(y_i, y_j), \text{ para } y_i \text{ em A e } y_j \text{ em B}\}$$

Em que: $d(y_i, y_j)$ é a distância euclidiana ou alguma outra distância entre os vetores y_i e y_j.

- Ligação completa ou vizinho mais distante (Complete-Linkage)

Este algoritmo é muito parecido com a União-Simples, na qual cada conglomerado é caracterizado pela união mais separada necessária para unir o membro de um conglomerado para outro membro.

No método de ligação completa a distância entre dois grupos A e B é definida como a máxima distância entre um ponto em A e um ponto em B:

$$d_{(A,B)} = máximo\{d(y_i, y_j), \text{ para } y_i \text{ em A e } y_j \text{ em B}\}$$

Destaca-se que uma vez obtidas as distâncias para cada par de grupos, os dois grupos com a menor distância são unidos.

- Ligação média (Average-Linkage) – Dentro

Na abordagem pelo método de ligação média, a distância entre dois grupos A e B é definida como a média das distâncias $n_A n_B$ entre os pontos n_A em A e n_B pontos em B:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (y_i, y_j)$$

Em que: A soma é em todos os valores de y_i em A e todos os valores de y_j em B. A cada passo são unidos os dois grupos com a menor distância $D(A,B)$.

- Centróide

Em análises estatísticas a média é geralmente a estatística básica para um conjunto de dados. O conhecido teste *t* e a técnica da análise de variância podem ser usados para identificar diferenças entre grupos, testando-se as diferenças entre suas médias. Isto, então, deverá ser um atrativo para unir hierarquicamente ao se agrupar em cada estágio aqueles dois conglomerados com os vetores médias mais similares ou centróides.

No método centróide a distância entre dois grupos A e B é definida como a distância euclidiana entre os vetores médios (centróides) dos dois grupos:

$$D(A,B) = d(\overline{y}_A, \overline{y}_B)$$

Em que: $\overline{y}_A$ e $\overline{y}_B$ são os vetores de médias das observações em A e B.

Os dois grupos com a menor distância são unidos. Depois que os grupos são unidos o centróide do novo grupo AB é dado pela média ponderada.

- *Método de Ward*

No método (algoritmo) de Ward as uniões em cada estágio são escolhidas, visando maximizar uma função objetiva baseada na soma de quadrados do erro com o propósito de encontrar em cada estágio aqueles dois conglomerados cuja união resulta no mínimo de aumento na soma de quadrados do erro no grupo total dentro.

O método de Ward usa as distâncias ao quadrado dentro dos grupos e as distâncias ao quadrado entre os grupos. Se AB é um grupo obtido pela combinação dos grupos A e B, então a soma das distâncias dentro são:

$$SQW_A = \sum_{i=1}^{n_A} (y_i - \overline{y}_A)'(y_i - \overline{y}_A)$$

$$SQW_B = \sum_{i=1}^{n_B} (y_i - \overline{y}_B)'(y_i - \overline{y}_B)$$

$$SQW_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \overline{y}_{AB})'(y_i - \overline{y}_{AB})$$

Em que: $\overline{y}_{AB}$ é o vetor médio ponderado entre A e B; n_A, n_B, n_{AB} são o número de pontos em A, B e AB.

O método de Ward une os dois grupos A e B, que minimizam o aumento na soma de quadrados das distâncias dentro, definido como:

$$I_{AB} = SQW_{AB} - (SQW_A + SQW_B)$$

8.2.2 Determinação do número de grupos e avaliação dos agrupamentos

A escolha do número de grupos nos métodos hierárquicos pode ser feita determinando um ponto da escala de distâncias e cortando os ramos do dendrograma. O ponto de corte determinará o número de grupo e a qualidade do agrupamento.

Um método usual e frequente que pode ser utilizado é determinar o número de grupos por meio do corte na metade da distância obtida no dendrograma. Outro critério é buscar a maior amplitude nas distâncias nas junções dos grupos. Pode-se, ainda, determinar o número de grupos desejados à priori dos agrupamentos.

A validação da análise de agrupamentos visa certificar que os grupos obtidos realmente diferem entre si. Para tanto, pode-se utilizar uma MANOVA para verificar se os grupos diferem ou análise de discriminante com o mesmo intuito e, também, identificar as variáveis mais importantes no agrupamento.

A avaliação da formação dos agrupamentos pelo método hierárquico pode ser efetuada por meio da comparação entre a informação gerada pela função de ligação e os dados originais de distância. Um agrupamento é considerado válido quando a ligação dos objetos no agrupamento apresenta forte correlação com as distâncias entre os objetos no vetor de distâncias. Essa avaliação é feita por meio da função cofenética, que compara esses dois conjuntos de valores e calcula a correlação entre eles. A melhor solução de agrupamento ocorre quando a correlação cofenética é igual a 1, indicando perfeita concordância entre as distâncias originais e as distâncias representadas pelo agrupamento. Quando o resultado obtido não é satisfatório, pode-se recorrer a outro método de agrupamento, como os métodos não hierárquicos.

8.2.3 Agrupamento não-hierárquico

Os métodos de agrupamento não hierárquicos não envolvem a construção de dendrogramas. Existem três técnicas de agrupamento não hierárquico: particionamento,

mistura de distribuições e estimação de densidades. Entre eles o particionamento é o mais comumente utilizado.

Os métodos não-hierárquicos são destinados a agrupar unidades de dados em classificações individuais de q grupos, em que q é tanto especificado a priori ou é determinado como parte do método. Importante ressaltar que os métodos não-hierárquicos não são adequados para variáveis.

Na abordagem de particionamento, as observações são separadas em q grupos sem a utilização da abordagem hierárquica baseada na matriz de distância ou similaridade entre todos os pares de pontos. Estes métodos são algumas vezes chamados de métodos de otimização ao invés de particionamento.

Uma estratégia seria examinar todos os possíveis caminhos para particionar n itens em q grupos e encontrar o agrupamento ótimo de acordo com algum critério. Entretanto, o número de possíveis grupos é proibitivamente grande, mesmo para valores pequenos de n e q.

Considera-se como abordagem para particionar a que é usualmente chamada de método de k-médias. O método permite que itens sejam movidos de um grupo para outro, realocação esta que não é disponível nos métodos hierárquicos.

O procedimento inicia-se com uma partição arbitrária dos itens em q grupos iniciais (sementes). Estas sementes serão substituídas mais tarde pelos centróides dos grupos. Depois que as sementes são escolhidas, cada ponto remanescente do conjunto de dados é alocado a grupo com a semente mais próxima baseada na distância euclidiana. Tão logo o grupo tenha mais que um membro, a semente é substituída pelo centróide.

Em seguida, cada item é realocado para o grupo cujo centroide, isto é, a média do grupo, esteja mais próximo, sendo a distância euclidiana a medida mais comumente empregada. Após cada realocação, o centroide é recalculado tanto para o grupo que recebeu o novo item quanto para o grupo que perdeu algum item. Este processo é repetido até que não ocorram mais realocações, indicando a estabilização dos grupos.

O método de k-médias é de alguma forma sensível à escolha das sementes iniciais. É aconselhável executar o procedimento novamente com sementes diferentes. Se as escolhas das sementes iniciais produzem diferentes grupos no final, ou se existe uma convergência extremamente lenta, pode ser que não existam grupos naturais nos dados. O método de particionamento de k-médias pode ser usado como uma melhoria das técnicas hierárquicas. Neste caso, os itens são agrupados com um método hierárquico e os centróides desses grupos são utilizados como sementes na abordagem com k-médias, permitindo que todos os pontos sejam realocados de um cluster para outro.

Exemplo 8.1

Os dados utilizados no exemplo de análise de agrupamento são os mesmos da análise de discriminante, com exceção da identificação do grupo. Importante mencionar que o conjunto de dados originais é composto de três grupos com 10 observações cada. O que se espera como resultado da análise de agrupamentos seriam três grupos, com o grupo 1 com as unidades de um a dez, o grupo 2 com as unidades de 11 a 20 e o grupo 3 com as unidades de 21 a 30.

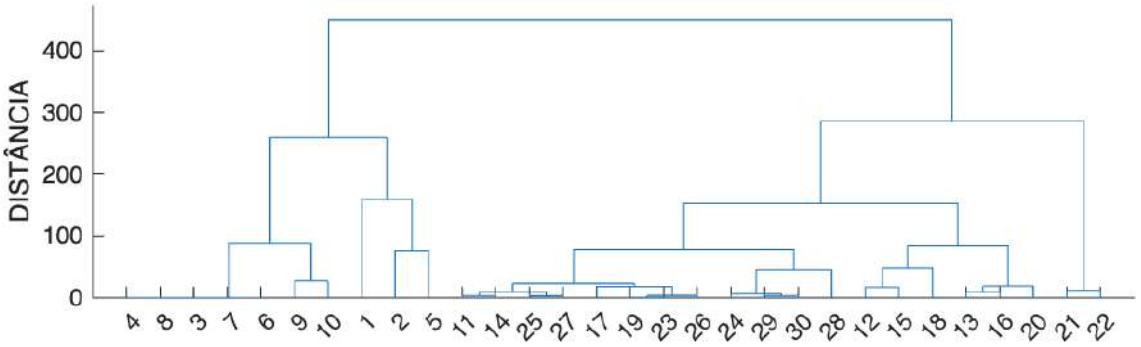
- Análise de agrupamento hierárquico com distância euclidiana e ligação simples

Quanto ao número de grupos, se aplicado um critério de metade da distância total (1,40), encontrou-se cinco grupos como mostra a linha do dendrograma. Os resultados encontrados no agrupamento hierárquico com distância euclidiana e ligação simples não revelaram a constituição de três grupos como se esperava.

A correlação cofenética de 0,6886 indica que o dendrograma obtido pelo método de ligação simples representa apenas de forma moderada as distâncias euclidianas originais entre as observações, preservando parcialmente a estrutura dos dados, mas com perda considerável de informação métrica.

- *Avaliação da análise de agrupamento hierárquico com testes de distâncias e tipos de ligação*

Com base na análise de correlação cofenética entre diferentes combinações de métricas de distância e métodos de ligação, a melhor combinação encontrada foi para o quadrado da distância euclidiana e método de ligação centróide, resultando em uma correlação cofenética de 0,8111, com o seguinte dendrograma:



Destaca-se que as unidades de 1 a 10 foram bem agrupadas, resultando em um único grupo. As demais unidades, 11 a 30 foram misturadas, o que confirma o resultado já percebido quando da análise de discriminante, onde estas observações são oriundas de grupos com menores diferenças entre eles.

- *Análise de agrupamento não hierárquico*

A análise de agrupamento não hierárquico pelo método k-médias, com k = 3, resultou na formação de três grupos distintos e internamente homogêneos, caracterizados por centroides bem definidos. O procedimento resultou em nove unidades no grupo 1; 18 no grupo 2 e 3 no grupo 3, indicando que somente para o grupo 1 as unidades foram bem agrupadas, enquanto para os grupos 2 e 3 houve transferência do grupo 3 para o grupo 2. Isto confirma o agrupamento dos grupos 2 e 3 já percebido, quando da análise de discriminante (Figura 8.3), em que esses dois grupos são bem próximos e, também, quando da análise de agrupamentos hierárquicos com quadrado da distância euclidiana e método de ligação centróide.

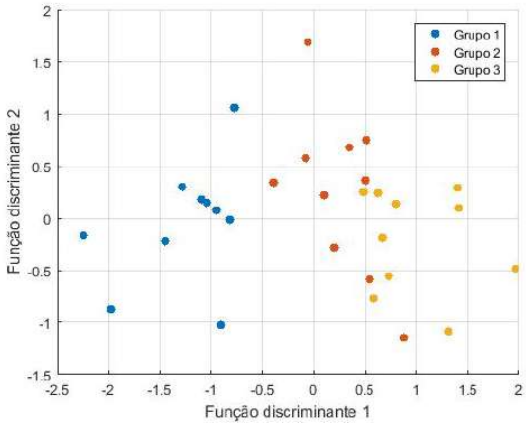


Figura 8.3: Grupos formados pela análise discriminante.

8.3 EXEMPLO DE ANÁLISE DE CONGLOMERADOS (CLUSTER) NO MATLAB

```
% ANÁLISE DE AGRUPAMENTO COM DISTÂNCIA EUCLIDIANA E LIGAÇÃO SIMPLES

% DADOS DE ENTRADA
X = [
   -1.0200   -1.3000    1.3000    2.3600
   -0.9500   -1.1600    1.1900    1.5700
   -0.8900   -1.1300    1.1100    0.9900
   -0.3100   -1.1300    1.1100    0.9900
    0.1300   -1.1200    1.3600    1.8800
    0.5300   -1.1300    1.1100    0.9900
   -1.3500   -1.1300    1.1100    0.9900
   -0.4700   -1.1300    1.1100    0.9900
    0.2700   -0.4200    0.1700    0.5900
    0.8200   -0.5900    0.9000    0.7000
   -0.7300    0.3600   -0.1000   -0.6100
   -0.2000    0.6400   -0.5300   -0.3500
   -0.6100    0.7100   -0.7100    0.1000
    0.7600   -0.0400   -0.0700   -0.6200
    0.6100   -0.0100   -0.0100   -0.2800
    0.4600   -0.7500    0.7200    0.1300
   -0.1700    0.3800   -0.5800   -0.4900
    0.7500         0   -0.1100   -0.1200
    0.3600    0.2600   -0.3400   -0.5600
   -0.2100    0.0800   -0.5200    0.0400
    0.7000    2.9500   -2.9200   -1.6500
    0.9500    0.6200   -0.6400   -1.6200
   -0.2900    0.4700   -0.3500   -0.5600
    2.1900    1.5300   -1.5500   -0.9300
    0.2400    0.5500   -0.5100   -0.6600
   -0.1400   -0.0700    0.2800   -0.5500
   -3.3900   -0.2900    0.1500   -0.6500
    1.1800    0.3200   -0.3200   -0.7700
   -0.1700    1.7600   -1.5600   -0.9600
    0.9700    0.8000   -0.8100   -0.9600
];

% NÚMERO DE CASOS
numero_casos = size(X,1);

% MATRIZ DE DISTÂNCIAS EUCLIDIANAS
matriz_distancias = zeros(n);
for i = 1:numero_casos
    for j = i+1:numero_casos
        matriz_distancias(i,j) = sqrt(sum((X(i,:) - X(j,:)).^2));
        matriz_distancias(j,i) = matriz_distancias(i,j);
    end
end

% INICIALIZAÇÃO DOS CLUSTERS
conglomerados = num2cell(1:numero_casos);
distancias_atuais = matriz_distancias;
matriz_ligacao = [];
matriz_cofenetica = zeros(numero_casos);
```

```
% IDENTIFICAÇÃO DOS CONGLOMERADOS
identificador_conglomerado = numero_casos;

% ALGORITMO HIERÁRQUICO – LIGAÇÃO SIMPLES
while numel(conglomerados) > 1
    matriz_temporaria = distancias_atuais + ...
        diag(inf(size(distancias_atuais,1),1));
    [distancia_minima, indice] = min(matriz_temporaria(:));
    [i,j] = ind2sub(size(matriz_temporaria), indice);
    conglomerado_A = conglomerados{i};
    conglomerado_B = conglomerados{j};

    % ATUALIZAÇÃO DA MATRIZ COFENÉTICA
    for p = conglomerado_A
        for q = conglomerado_B
            if matriz_cofenetica(p,q) == 0
                matriz_cofenetica(p,q) = distancia_minima;
                matriz_cofenetica(q,p) = distancia_minima;
            end
        end
    end

    % ATUALIZAÇÃO DA MATRIZ DE LIGAÇÃO
    identificador_conglomerado = identificador_conglomerado + 1;
    matriz_ligacao = [matriz_ligacao; ...
        conglomerado_A(1), conglomerado_B(1), distancia_minima];
    novo_conglomerado = [conglomerado_A, conglomerado_B];

    % REMOÇÃO DOS CONGLOMERADOS UNIFICADOS
    indices_remocao = sort([i j], 'descend');
    conglomerados(indices_remocao) = [];
    distancias_atuais(indices_remocao,:) = [];
    distancias_atuais(:,indices_remocao) = [];

    % DISTÂNCIAS DO NOVO CONGLOMERADO
    numero_conglomerados = numel(conglomerados);
    novas_distancias = zeros(1,numero_conglomerados);
    for k = 1:numero_conglomerados
        lista_distancias = [];
        for a = novo_conglomerado
            for b = conglomerados{k}
                lista_distancias(end+1) = matriz_distancias(a,b);
            end
        end
        novas_distancias(k) = min(lista_distancias);
    end
    distancias_atuais = [distancias_atuais, novas_distancias';
                         novas_distancias, 0];
    conglomerados{end+1} = novo_conglomerado;
end

% CÁLCULO DA CORRELAÇÃO COFENÉTICA

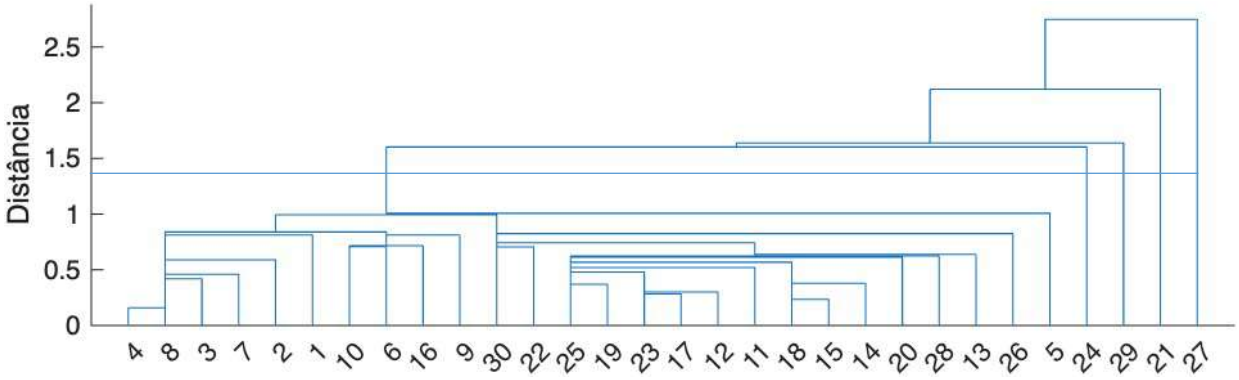
% VETOR DAS DISTÂNCIAS ORIGINAIS
distancias_originais = ...
    matriz_distancias(triu(true(numero_casos),1));
```

```
% VETOR DAS DISTÂNCIAS COFENÉTICAS
distancias_cofeneticas = ...
    matriz_cofenetica(triu(true(numero_casos),1));

% CORRELAÇÃO COFENÉTICA DE PEARSON
correlacao_cofenetica = ...
    corr(distancias_originais, distancias_cofeneticas);

% RESULTADOS
disp('Correlação cofenética:');
disp(correlacao_cofenetica)
Correlação cofenética:
    0.6886

% DENDROGRAMA
figure;
dendrogram(matriz_ligacao);
title('Dendrograma - Ligação Simples');
```




```
% MAIOR DISTÂNCIA NO DENDROGRAMA
maior_distancia = max(matriz_ligacao(:,3))
maior_distancia =
    2.7499

% FUNÇÕES PARA TESTE DE DISTÂNCIAS E LIGAÇÕES

% MATRIZ DE DADOS

% TIPOS DE DISTÂNCIA
dist_types = { ...
    'euclidean', ...
    'seuclidean', ...
    'cityblock', ...
    'mahalanobis', ...
    'minkowski'};

% TIPOS DE LIGAÇÃO
link_types = { ...
    'single', ...     % vizinho mais próximo
    'complete', ...   % vizinho mais distante
    'average', ...    % média
    'centroid', ...   % centróide
    'ward'};          % Ward

% MATRIZ PARA ARMAZENAR RESULTADOS
R = zeros(length(dist_types), length(link_types));
```

```
% LOOP PRINCIPAL PARA COMBINAR DISTÂNCIAS E MÉTODOS
for i = 1:length(dist_types)

    % Vetor de distâncias
    Y = pdist(X, dist_types{i});
    for j = 1:length(link_types)
        % Ward só é matematicamente consistente com distância
euclidiana
        if strcmp(link_types{j}, 'ward') && ...
           ~strcmp(dist_types{i}, 'euclidean') && ...
           ~strcmp(dist_types{i}, 'seuclidean')
            R(i,j) = NaN;
            continue
        end

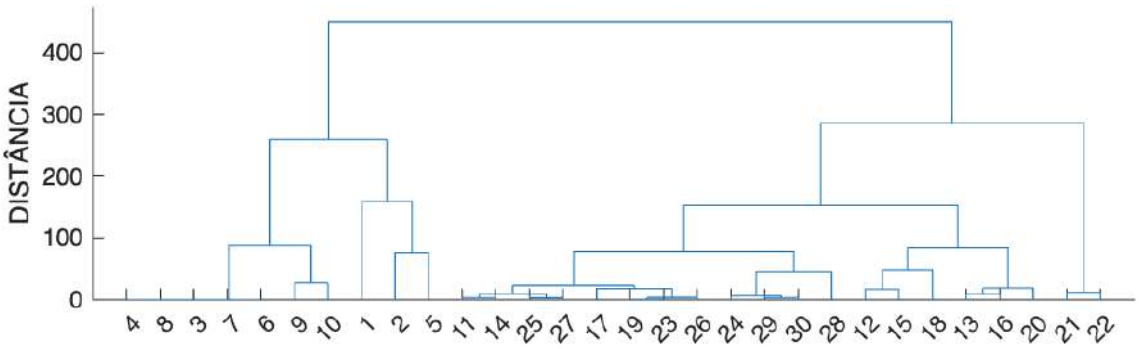
        % AGRUPAMENTO HIERÁRQUICO
        Z = linkage(Y, link_types{j});
        % Correlação cofenética
        R(i,j) = cophenet(Z, Y);
    end
end

% RESULTADOS DA CORRELAÇÃO COFENÉTICA PARA COMBINAÇÕES DE TIPOS DE
DISTÂNCIAS E MÉTODOS DE LIGAÇÃO
T = array2table(R, ...
    'RowNames', dist_types, ...
    'VariableNames', link_types);
disp(T)
disp(T)
```

	Simples	Completa	Média	Centróide	Ward
Euclidiana	0.6886	0.68272	0.8072	0.81183	0.61977
Eucldiana²	0.6886	0.68273	0.8072	0.81185	0.61979
Cityblock	0.69932	0.76579	0.79895	0.79629	NaN
Mahalanobis	0.73581	0.51376	0.7603	0.77955	NaN
Minkowski	0.6886	0.68272	0.8072	0.81183	NaN

```
% MELHOR COMBINAÇÃO – DISTÂNCIA: QUADRADO DA DISTÂNCIA EUCLIDIANA,
COM LIGAÇÃO CENTRÓIDE, R=0,81

% DENDROGRAMA
>> Z = linkage(Y,'centroid');
dendrogram(Z,0)
```




```
%RESULTADOS PARA VARIÁVEIS ORIGINAIS (NÃO-PADRONIZADAS)
disp(T)
```

	Simples	Completa	Média	Centróide	Ward
Euclidiana	0.76845	0.79692	0.80457	0.80457	0.78525
Euclidiana²	0.66238	0.67923	0.78413	0.81108	0.63198
Cityblock	0.76809	0.79707	0.80465	0.80465	NaN
Mahalanobis	0.74122	0.52106	0.78411	0.77888	NaN
Minkowski	0.76845	0.79692	0.80457	0.80457	NaN

8.4 ANÁLISE DE AGRUPAMENTO NÃO HIERÁRQUICO UTILZIANDO COM K-MEANS NO MATLAB

```
% DADOS (30 x 4)
X = [
   -1.0200   -1.3000    1.3000    2.3600
   -0.9500   -1.1600    1.1900    1.5700
   -0.8900   -1.1300    1.1100    0.9900
   -0.3100   -1.1300    1.1100    0.9900
    0.1300   -1.1200    1.3600    1.8800
    0.5300   -1.1300    1.1100    0.9900
   -1.3500   -1.1300    1.1100    0.9900
   -0.4700   -1.1300    1.1100    0.9900
    0.2700   -0.4200    0.1700    0.5900
    0.8200   -0.5900    0.9000    0.7000
   -0.7300    0.3600   -0.1000   -0.6100
   -0.2000    0.6400   -0.5300   -0.3500
   -0.6100    0.7100   -0.7100    0.1000
    0.7600   -0.0400   -0.0700   -0.6200
    0.6100   -0.0100   -0.0100   -0.2800
    0.4600   -0.7500    0.7200    0.1300
   -0.1700    0.3800   -0.5800   -0.4900
    0.7500         0   -0.1100   -0.1200
    0.3600    0.2600   -0.3400   -0.5600
   -0.2100    0.0800   -0.5200    0.0400
    0.7000    2.9500   -2.9200   -1.6500
    0.9500    0.6200   -0.6400   -1.6200
   -0.2900    0.4700   -0.3500   -0.5600
    2.1900    1.5300   -1.5500   -0.9300
    0.2400    0.5500   -0.5100   -0.6600
   -0.1400   -0.0700    0.2800   -0.5500
   -3.3900   -0.2900    0.1500   -0.6500
    1.1800    0.3200   -0.3200   -0.7700
   -0.1700    1.7600   -1.5600   -0.9600
    0.9700    0.8000   -0.8100   -0.9600
];
[n, p] = size(X);
k = 3;

% Inicialização dos centroides
% Escolha manual de 3 observações iniciais
centroides = X([1 11 21], :);   % 3 x 4

% Parâmetros do algoritmo
maxiter = 100;
tol = 1e-6;
grupos = zeros(n,1);
centroides_antigos = centroides;
% Algoritmo k-means (iterativo)
for iter = 1:maxiter
```

```
    % PASSO 1: Atribuição aos clusters
    for i = 1:n
        distancias = zeros(k,1);
        for g = 1:k
            distancias(g)=sqrt(sum((X(i,:) - centroides(g,:)).^2));
        end
        [~, grupos(i)] = min(distancias);
    end

% PASSO 2: Atualização dos centroides
    for g = 1:k
        idx = find(grupos == g);
        centroides(g,:) = mean(X(idx,:), 1);
    end

    % PASSO 3: Critério de parada
    deslocamento = max(max(abs(centroides - centroides_antigos)));

    if deslocamento < tol
        fprintf('Convergência atingida na iteração %d\n', iter);
        break
    end
    centroides_antigos = centroides;
end

%Resultados finais
disp('Grupo atribuído a cada observação:');
disp(grupos);
disp('Centroides finais:');
disp(centroides);

%Resultados finais
disp('Grupo atribuído a cada observação:');
disp(grupos);
disp('Centroides finais:');
disp(centroides);
Convergência atingida na iteração 3
Grupo atribuído a cada observação:
     1
     1
     1
     1
     1
     1
     1
     1
     2
     1
     2
     2
     2
     2
     2
     2
     2
     2
     2
     2
```

```
     3
     2
     2
     3
     2
     2
     2
     2
     3
     2
Centroides finais:
   -0.3900   -1.0911    1.1444    1.2733
    0.0450    0.2006   -0.2378   -0.4411
    0.9067    2.0800   -2.0100   -1.1800

% CONTAGEM DE OBSERVAÇÕES POR GRUPO
disp('Número de observações em cada grupo:');
for g = 1:k
    fprintf('Grupo %d: %d observações\n', g, sum(grupos == g));
end
Número de observações em cada grupo:
Grupo 1: 9 observações
Grupo 2: 18 observações
Grupo 3: 3 observações
```

8.5 EXEMPLO DE ANÁLISE DE CONGLOMERADOS (CLUSTER) NO JULIA

```
# ANÁLISE DE AGRUPAMENTO COM DISTÂNCIA EUCLIDIANA E LIGAÇÃO SIMPLES

using LinearAlgebra; Statistics

# DADOS DE ENTRADA
X = [
   -1.0200   -1.3000    1.3000    2.3600;
   -0.9500   -1.1600    1.1900    1.5700;
   -0.8900   -1.1300    1.1100    0.9900;
   -0.3100   -1.1300    1.1100    0.9900;
    0.1300   -1.1200    1.3600    1.8800;
    0.5300   -1.1300    1.1100    0.9900;
   -1.3500   -1.1300    1.1100    0.9900;
   -0.4700   -1.1300    1.1100    0.9900;
    0.2700   -0.4200    0.1700    0.5900;
    0.8200   -0.5900    0.9000    0.7000;
   -0.7300    0.3600   -0.1000   -0.6100;
   -0.2000    0.6400   -0.5300   -0.3500;
   -0.6100    0.7100   -0.7100    0.1000;
    0.7600   -0.0400   -0.0700   -0.6200;
    0.6100   -0.0100   -0.0100   -0.2800;
    0.4600   -0.7500    0.7200    0.1300;
   -0.1700    0.3800   -0.5800   -0.4900;
    0.7500    0.0000   -0.1100   -0.1200;
    0.3600    0.2600   -0.3400   -0.5600;
   -0.2100    0.0800   -0.5200    0.0400;
    0.7000    2.9500   -2.9200   -1.6500;
```

```
    0.9500    0.6200   -0.6400   -1.6200;
   -0.2900    0.4700   -0.3500   -0.5600;
    2.1900    1.5300   -1.5500   -0.9300;
    0.2400    0.5500   -0.5100   -0.6600;
   -0.1400   -0.0700    0.2800   -0.5500;
   -3.3900   -0.2900    0.1500   -0.6500;
    1.1800    0.3200   -0.3200   -0.7700;
   -0.1700    1.7600   -1.5600   -0.9600;
    0.9700    0.8000   -0.8100   -0.9600
]
numero_observacoes = size(X, 1)

# MATRIZ DE DISTÂNCIAS EUCLIDIANAS
matriz_distancias = zeros(numero_observacoes, numero_observacoes)
for i in 1:numero_observacoes
    for j in i+1:numero_observacoes
        matriz_distancias[i,j] = norm(X[i,:] - X[j,:])
        matriz_distancias[j,i] = matriz_distancias[i,j]
    end
end

# INICIALIZAÇÃO
agrupamentos = [[i] for i in 1:numero_observacoes]
distancias_atuais = copy(matriz_distancias)
matriz_ligacao = zeros(numero_observacoes - 1, 3)
matriz_cofenetica = zeros(numero_observacoes, numero_observacoes)
etapa = 0

# ALGORITMO HIERÁRQUICO (LIGAÇÃO SIMPLES)
while length(agrupamentos) > 1
    etapa += 1
    # IGNORAR A DIAGONAL
    matriz_temporaria = copy(distancias_atuais)
    for k in 1:size(matriz_temporaria, 1)
        matriz_temporaria[k,k] = Inf
    end
    # MENOR DISTÂNCIA
    indice = argmin(matriz_temporaria)
    i, j = Tuple(indice)
    menor_distancia = matriz_temporaria[i,j]
    agrupamento_A = agrupamentos[i]
    agrupamento_B = agrupamentos[j]
    # ATUALIZAR MATRIZ COFENÉTICA
    for p in agrupamento_A
        for q in agrupamento_B
            if matriz_cofenetica[p,q] == 0.0
                matriz_cofenetica[p,q] = menor_distancia
                matriz_cofenetica[q,p] = menor_distancia
            end
        end
    end
```

```
    # MATRIZ DE LIGAÇÃO
    matriz_ligacao[etapa, :] = [
        agrupamento_A[1],
        agrupamento_B[1],
        menor_distancia]

    # NOVO AGRUPAMENTO
    novo_agrupamento = vcat(agrupamento_A, agrupamento_B)

    # REMOVER AGRUPAMENTOS ANTIGOS
    # (ORDEM DECRESCENTE)

    for indice_agrupamento in sort([i, j], rev=true)
        deleteat!(agrupamentos, indice_agrupamento)
        distancias_atuais = distancias_atuais[
            setdiff(1:end, indice_agrupamento),
            setdiff(1:end, indice_agrupamento)]
    end

    # CALCULAR NOVAS DISTÂNCIAS
    # (LIGAÇÃO SIMPLES)
    numero_agrupamentos = length(agrupamentos)
    novas_distancias = zeros(numero_agrupamentos)
    for k in 1:numero_agrupamentos
        lista_distancias = Float64[]
        for a in novo_agrupamento
            for b in agrupamentos[k]
                push!(
                    lista_distancias,
                    matriz_distancias[a,b])
            end
        end
        novas_distancias[k] = minimum(lista_distancias)
    end

    # ATUALIZAR MATRIZ DE DISTÂNCIAS
    distancias_atuais = [
        distancias_atuais   novas_distancias;
        novas_distancias'   0.0]
    push!(agrupamentos, novo_agrupamento)
end

# CORRELAÇÃO COFENÉTICA
mascara = triu(trues(numero_observacoes, numero_observacoes), 1)
distancias_originais = matriz_distancias[mascara]
distancias_cofeneticas = matriz_cofenetica[mascara]

correlacao_cofenetica = cor(distancias_originais, distancias_cofeneticas)

println("Correlação cofenética = ", correlacao_cofenetica)
```

```
# MAIOR DISTÂNCIA NO DENDROGRAMA
maior_distancia = maximum(matriz_ligacao[:,3])
println("Maior distância no dendrograma = ", maior_distancia)

# OBTENÇÃO DO DENDROGRAMA
using Clustering; Distances; StatsBase; StatsPlots; LinearAlgebra; Statistics

# MATRIZ DE DISTÂNCIAS EUCLIDIANAS
matriz_distancias = pairwise(Euclidean(), X; dims=)

# AGRUPAMENTO HIERÁRQUICO
# MÉTODO DE LIGAÇÃO SIMPLES
agrupamento_hierarquico = hclust(matriz_distancias, linkage=:single)

# REPRESENTAÇÃO GRÁFICA DO DENDROGRAMA
plot(agrupamento_hierarquico)
```

8.6 ANÁLISE DE AGRUPAMENTO NÃO HIERÁRQUICO - MÉTODO DO K-MEANS NO JULIA

```
using LinearAlgebra; Statistics

# Dados de entrada
X = [
   -1.0200   -1.3000    1.3000    2.3600
   -0.9500   -1.1600    1.1900    1.5700
   -0.8900   -1.1300    1.1100    0.9900
   -0.3100   -1.1300    1.1100    0.9900
    0.1300   -1.1200    1.3600    1.8800
    0.5300   -1.1300    1.1100    0.9900
   -1.3500   -1.1300    1.1100    0.9900
   -0.4700   -1.1300    1.1100    0.9900
    0.2700   -0.4200    0.1700    0.5900
    0.8200   -0.5900    0.9000    0.7000
   -0.7300    0.3600   -0.1000   -0.6100
   -0.2000    0.6400   -0.5300   -0.3500
   -0.6100    0.7100   -0.7100    0.1000
    0.7600   -0.0400   -0.0700   -0.6200
    0.6100   -0.0100   -0.0100   -0.2800
    0.4600   -0.7500    0.7200    0.1300
   -0.1700    0.3800   -0.5800   -0.4900
    0.7500         0   -0.1100   -0.1200
    0.3600    0.2600   -0.3400   -0.5600
   -0.2100    0.0800   -0.5200    0.0400
    0.7000    2.9500   -2.9200   -1.6500
    0.9500    0.6200   -0.6400   -1.6200
   -0.2900    0.4700   -0.3500   -0.5600
    2.1900    1.5300   -1.5500   -0.9300
    0.2400    0.5500   -0.5100   -0.6600
   -0.1400   -0.0700    0.2800   -0.5500
   -3.3900   -0.2900    0.1500   -0.6500
```

```
    1.1800    0.3200   -0.3200   -0.7700
   -0.1700    1.7600   -1.5600   -0.9600
    0.9700    0.8000   -0.8100   -0.9600
]

n, p = size(X)

# Definição do número de grupos
k = 3

# Inicialização dos centroides
# Índices escolhidos arbitrariamente
centroides = X[[1, 11, 21], :]

# Função distância euclidiana
function dist_euclidiana(x, c)
    return sqrt(sum((x .- c).^2))
end

# Iterações do k-means
maxiter = 100
tol = 1e-6
grupos = zeros(Int, n)
centroides_antigos = copy(centroides)
for iter in 1:maxiter
    # PASSO 1: Atribuição aos grupos
    for i in 1:n
        distancias = zeros(k)
        for g in 1:k
            distancias[g] = dist_euclidiana(X[i, :], centroides[g, :])
        end
        grupos[i] = argmin(distancias)
    end
    # PASSO 2: Atualização dos centroides
    for g in 1:k
        idx = findall(grupos .== g)
        centroides[g, :] = mean(X[idx, :], dims=1)
    end
    # PASSO 3: Critério de parada
    deslocamento = maximum(abs.(centroides .- centroides_antigos))
    if deslocamento < tol
        println("Convergência atingida na iteração ", iter)
        break
    end
    centroides_antigos = copy(centroides)
end

#Resultados finais
println("\nGrupos atribuídos a cada observação:")
println(grupos)
println("\nCentroides finais:")
```

```
println(centroides)

# Número de observações em cada grupos
println("Número de observações em cada grupo:")

for g in 1:k
    println("Grupo $g: ", sum(grupos .== g), " observações")
end
```

9 ANÁLISE DE CORRELAÇÃO CANÔNICA

9.1 INTRODUÇÃO

A análise de correlação canônica consiste em um modelo estatístico multivariado voltado à investigação das inter-relações existentes entre conjuntos de múltiplas variáveis dependentes e múltiplas variáveis independentes. Sua aplicação resulta no desenvolvimento de diversas funções canônicas independentes, construídas de modo a maximizar a correlação entre composições lineares das variáveis originais, denominadas variáveis estatísticas canônicas. Cada uma dessas funções é fundamentada na correlação entre um par de variáveis estatísticas canônicas, sendo uma associada ao conjunto de variáveis dependentes e a outra ao conjunto de variáveis independentes, permitindo uma compreensão mais abrangente das relações simultâneas entre os dois conjuntos de dados.

Na análise de correlação canônica, as variáveis estatísticas canônicas são obtidas de forma a maximizar a correlação entre si, permitindo identificar as relações mais fortes entre os conjuntos de variáveis analisadas. Este procedimento não se limita à derivação de uma única relação entre os conjuntos de variáveis, pois possibilita a obtenção de diversas funções canônicas, cada uma composta por um par de variáveis estatísticas canônicas mutuamente independentes. Dessa forma, as correlações canônicas constituem um instrumento robusto para explicar e interpretar as relações existentes entre os conjuntos de variáveis dependentes e independentes em um conjunto de dados multivariado.

Frequentemente, em ajustamento de variáveis, o pesquisador tem dois conjuntos de variáveis, y e x, disponíveis para uma amostra de elementos, cada variável vetor representando múltiplas medidas de um espaço finito. O que pode acontecer é que y seja um conjunto de medidas estimadas, ou medidas de variáveis dependentes em uma pesquisa, e x seja um conjunto de medidas de variáveis obtidas segundo determinados critérios, ou medidas independentes. Por exemplo, um conjunto de variáveis x formado pelo diâmetro e a altura da árvore e um conjunto y que representa a biomassa do fuste e da copa.

Outra possibilidade é que os dois domínios sejam conceitualmente diferentes, apesar de terem sido observações medidas simultaneamente. Seja y um conjunto de testes de habilidades para operar uma máquina florestal e x seja um conjunto de escalas de produtividade. A questão da pesquisa então será mostrar as interrelações das habilidades e a produtividade. É evidente, que as correlações bivariadas entre os pares de medições tomadas em cada domínio são de interesse, mas existe uma grande quantidade deles.

Por exemplo, se existir 10 testes de habilidade de operação das máquinas e 10 escalas de produtividade, existirão 100 correlações bivariadas entre pares de medidas de produtividade e habilidades. Tentar entender todas estas correlações simultaneamente é extremamente difícil se houver uma tentativa de generalizar sobre a extensão e a natureza das interrelações dos domínios. Se existirem p_1 elementos em y e q_1 elementos em x, e $p = p_1 + q_1$, a matriz correlação simétrica R de ordem p é subdividida em:

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right]$$

tal que R_{11} contém as intercorrelações entre os elementos de y e é de ordem p_1, R_{22} são as intercorrelações cruzadas entre os elementos de x e q_1 e de ordem $p_1 \cdot q_1$.

As interrelações entre dois conjuntos de medidas feitas nos mesmos objetos podem ser estudadas por meio do método da correlação canônica. Originalmente desenvolvido por Hotelling (1935,1936), a correlação canônica é a máxima correlação

entre funções lineares de duas variáveis vetores. Contudo, após o par de funções lineares que correlaciona maximamente seja encontrado, então existe a oportunidade para se obter pares de funções adicionais, que correlacionam maximamente, sujeitos à restrição que s funções e cada novo par precisa ser não correlacionada com todas as funções previamente localizadas em ambos os domínios. Cada par de funções é determinado para maximizar a correlação entre o novo par de variáveis canônicas, sujeito à restrição que elas sejam inteiramente ortogonais em relação a todas as combinações lineares derivadas.

O modelo da correlação canônica usa a mesma abordagem analítica para mostrar a estrutura das relações por meio dos domínios de medições, cujo modelo fatorial usa para mostrar a estrutura de relação dentro de um domínio. O procedimento envolve reduzir a dimensionalidade para poucas funções lineares de medições em estudo. O modelo fatorial seleciona funções lineares de testes, que têm variâncias máximas, sujeitas às restrições de ortogonalidade. O modelo canônico seleciona funções lineares, que têm covariâncias entre domínios, sujeitos novamente à restrições de ortogonalidade. Geometricamente, o modelo canônico pode ser considerado uma exploração da extensão em que indivíduos ocupem as mesmas posições relativas em um espaço de medição, assim como também no outro. O que se procura, explicitamente, é que a variável do vetor y mapeie sob as mesmas condições que a variável vetor x.

9. 2 DESENVOLVIMENTO

Na análise de correlação canônica o objetivo é dividir as variáveis em dois grupos e encontrar as combinações lineares do primeiro e segundo grupo, tal que a correlação das duas combinações lineares seja a maior possível. Uma vez encontrando este melhor conjunto de combinações lineares é possível continuar a investigar qual será o próximo melhor conjunto sujeito à limitação, que o segundo conjunto seja não correlacionado com o primeiro. O processo pode continuar até que se tenha tantos pares de combinações lineares quanto existem variáveis no conjunto menor. Isto pode ser visto como uma extensão da regressão múltipla para o caso no qual existem muitas variáveis dependentes.

Inicia-se dividindo a matriz de dados originais em duas partes, correspondentes para cada um dos dois conjuntos de variáveis. A partir de um conjunto inicial de dados obtém-se $Y_{n \cdot p}$ e $X_{n \cdot q}$. É possível calcular três matrizes de variância-covariância:

$$A_{p \cdot p} = Cov\ Y$$
$$B_{g \cdot g} = Cov\ X$$
$$A_{p \cdot g} = Cov\ XY$$

O que se procura é a combinação linear das variáveis Y:

$$Y_{(n \cdot p)} g_{(p \cdot 1)} = Y_{g(n \cdot 1)}$$

e a combinação linear das variáveis X:

$$X_{(n \cdot g)} h_{(g \cdot 1)} = X_{h(n \cdot 1)}$$

tal que a correlação entre as duas combinações lineares seja tão grande quanto possível.

Assumindo vetores de média zero por conveniência, a covariância entre as combinações lineares torna-se:

$$\left(Y_g\right)' (X_h) = \frac{g' Y' X h}{(n-1)}$$

Para obter a correlação, é necessário dividir esta covariância pelo produto dos desvios padrões. Assumindo vetores de média zero, as variâncias para as duas combinações lineares são:

$$\left(Y_g\right)' \left(Y_q\right) = \frac{g' Y' Y g}{(n-1)}$$

e

$$(X_h)'(X_h) = \frac{h'X'Xh}{(n-1)}$$

sendo que o quadrado da correlação entre as duas combinações lineares é:

$$r^2 = \frac{(g'Y'X_h')^2}{(g'Y'Yg)(h'X'X_h)}$$

Maximizar o quadrado da correlação maximizará a correlação e pode-se tomar isto como o critério a ser maximizado concernente aos componentes de g e h.

Derivando-se a equação em relação aos componentes g e h e igualando-a parcialmente a zero, resulta:

$$\{g'Ag\}C_h - \{g'C_h\}Ag = 0$$

e

$$\{h'Bh\}C_g' - \{g'C_h\}B_h = 0$$

em que as quantidades nos colchetes são escalares.

Após simplificar as equações resultante e eliminar g e h, elas se reduzem a:

$$(C'A^{-1}C - r^2B)h = 0$$
$$(CB^{-1}C' - r^2A)g = 0$$

Em que: r^2 é o valor máximo da correlação ao quadrado, sendo que r^2 é um escalar,

$$r^2B = Br^2$$

e

$$r^2A = Ar^2$$

Assim, pode-se modificar o que foi acima mostrado, como:

$$(B^{-1}C'A^{-1}C - \lambda I)h = 0$$
$$(A^{-1}CB^{-1}C' - \lambda I)g = 0$$

Em que: $\lambda = r^2$.

Isto gera um conjunto de dois valores e vetores característicos, resultante das soluções acima dispostas como matrizes G e H em colunas, então:

$$G'AG = D_1$$
$$H'BH = D_2$$
$$G'CH = D_3$$

Em que: D_i é matriz diagonal.

As combinações lineares definidas pelos vetores característicos, após a obtenção do primeiro autovetor, são não correlacionadas dentro e entre o conjunto de variáveis. Eles satisfazem o critério de valores imutáveis das correlações. As outras transformações constituem os pares remanescentes das variáveis canônicas.

Existe também um procedimento alternativo para o cálculo do segundo conjunto de vetores característicos. Para tanto, deve-se fazer o particionamento das variáveis de tal maneira, que a matriz B corresponda ao menor conjunto. Então,

$$G_{(p \cdot g)} = A^{-1}C\ H$$

com o remanescente das colunas $p - g$ de G sendo irrelevantes, pois elas correspondem às correlações canônicas zero. Os vetores obtidos desta equação precisam então ser normalizados.

As variáveis canônicas podem ser calculadas aplicando-se as matrizes de transformação para os dados originais.

Dado que as correlações não são alteradas pela escala das variáveis, o mesmo valor para as correlações canônicas será obtido usando-se as matrizes de correlação em substituição às matrizes variância-covariância. Neste caso, contudo, é necessário padronizar a matriz de dados antes de se calcular as variáveis canônicas, que não serão as mesmas que as variáveis calculadas da matriz variância-covariância.

Para as variáveis canônicas terem média zero, como descrito por Marriott (1974), é necessário usar as matrizes de correlação e dados padronizados. Marriott (1974) apresenta uma boa discussão para se testar as hipóteses na análise de correlação canônica.

O cálculo das correlações canônicas e variáveis canônicas não exigem hipóteses de distribuição, mas para testar hipóteses assume-se normalidade multivariada. Então, pode-se utilizar os seguintes passos para a condução da correlação canônica:

- *Obtenção da matriz variância-covariância dos dados originais*

A primeira matriz a ser calculada é a matriz variância-covariância a partir dos dados originais, conforme o seguinte esquema:

$A_{(p\cdot p)}$	$C_{(p\cdot q)}$
$C'_{(p\cdot p)}$	$B_{(q\cdot q)}$

Em que: A é a matriz variância-covariância das variáveis $X's$ (primeiro conjunto de variáveis, que também pode ser denominado de $X^{(1)}$ ou X^1); B é a matriz variância-covariância das variáveis $Y's$ (segundo conjunto de variáveis, que também pém pode ser denominado de $X^{(2)}$ ou X^2); C é a matriz variância-covariância das variáveis $X's$ e $Y's$; C' é a matriz variância-covariância cruzada transposta entre as variáveis $X's$ e $Y's$.

- *Dividir as matrizes*

Primeiro isolar a matriz A, em seguida a matriz B, depois a matriz C, a matriz C' e encontrar a inversa da matriz A e B. A razão deste procedimento é encontrar as variáveis canônicas para transformar os dados em um novo espaço canônico e, também, encontrar a correlação entre as duas combinações lineares.

Assim, para transformar as variáveis $Y's$ para o novo espaço canônico foi feito o seguinte cálculo:

- $(B^{-1}C'A^{-1}C)$: para se obter a matriz denominada de TRÊS.
- $(A^{-1}CA^{-1}C')$: para transformar as variáveis $X's$ no novo espaço canônico, denominada de QUATRO.

Com isto extrai-se os valores característicos e vetores característicos das matrizes TRÊS e QUATRO, nos quais os valores característicos são os quadrados dos coeficientes de correlação, que correspondem às correlações canônicas.

Os vetores característicos são os correspondentes pares de combinações lineares para transformar as variáveis $X's$ e $Y's$ no novo espaço canônico.

- *Verificar se os dois conjuntos não são correlacionados.*

Se não existe correlação entre os dois conjuntos, então para a matriz A faz-se

$$D_1 = G'AG$$

para a matriz B

$$D_2 = H'BH$$

e para a matriz C

$$D_3 = G'CH$$

A não existência de correlação será obtida quando as três matrizes forem diagonais.

- *Encontrar as variáveis canônicas entre X e Y*

Após encontrar as variáveis canônicas, calcula-se as matrizes variância-covariância e as correlações. Este procedimento é efetuado para verificar se estas matrizes não são correlacionadas entre elas.

Para a covariância cruzada e para verificar se ambos os conjuntos não são correlacionados entre si, são juntadas as variáveis canônicas para X e Y, extraída a variância-covariância e a correlação das matrizes.

- *Testes de significância*

Após realizados os passos acima, é possível então aplicar os testes de significância para determinar se as relações entre os dois conjuntos de variáveis podem ser descritas por um conjunto de S componentes. É importante notar que, para descrever completamente as relações de p medidas de X_j e q medidas de Y_k, correlações $S = \min(p, q)$ são necessárias. Cada uma dessas é uma correlação produto-momento de uma combinação linear de X_j. Os pesos que definem as funções lineares são escolhidos para maximizar a medida de associação.

Para o teste geral, pode-se utilizar o critério de Wilks. Para tanto, calcula-se a matriz de resíduos sobre a regressão, usando-se X como estimadores para os $Y's$, ou seja:

$$v_{11}^{*} = v_{11} - v_{21}v_{22}^{-1}v_{12}$$

Em que: v_{11}^{*} é a matriz de resíduos da regressão.

Em seguida, calcula-se a SQTotal e SQResíduo. O critério de Wilks passa a ser calculado como:

$$L = \frac{|V_{11}^{*}|}{|V_{11}|} \begin{matrix} \rightarrow SQResíduo \\ \rightarrow SQTotal \end{matrix}$$

Para testar sua significância, obtém-se o valor do Qui-quadrado a partir da seguinte fórmula:

$$\chi^2 = \left\{-n + \frac{1}{2}(p + q + 1)\right\} lnL$$

com $(p-1)(q-1)$ gruas de liberdade para um nível de significância α (Cooley e Lohnes, 1971). Se o teste for significante conclui-se que variáveis $X's$ são boas estimadoras das variáveis $Y's$.

- *Cálculos realizados com a matriz de correlação*

Outro ponto que dever ser levado em consideração é que quando se usa a matriz variância-covariância, existe o problema de escala (variáveis com unidades métricas diferentes), o que não acontece com a matriz correlação. Assim, os mesmos procedimentos apresentados anteriormente também podem ser efetuados usando a matriz correlação. A única diferença é que as variáveis devem ser padronizadas antes da obtenção das variáveis canônicas e que também é necessário a padronização dos dados antes de se multiplicar pelas variáveis canônicas para se obter os coeficientes canônicos (*variates*).

Observa-se as seguintes condições:

- A matriz variância-covariância é normalmente distribuída com uma determinada variância, $N\sim(\mu, \sigma^2)$
- A matriz correlação é normalmente distribuída com média zero e variância um, $N\sim(0,1)$
- Quando os dados são padronizados, eles seguem a distribuição Z. Isto é, padroniza-se uma variável com a seguinte fórmula:

$$x_p = \frac{x - \bar{x}}{s}$$

Em que: x_p é variável padronizada, x é a variável original; $\bar{x}$ é a média e s é o desvio padrão do conjunto de dados amostrais.

Outra abordagem da análise de correlação canônica é apresentada a seguir.

Considere dois conjuntos de variáveis aleatórias. O primeiro conjunto é composto por p variáveis e representado pelo vetor aleatório $X^{(1)}$ de dimensão $(p\ x\ 1)$. O segundo conjunto é formado por q variáveis, representado pelo vetor aleatório $X^{(2)}$de dimensão q, sendo assumido que $p \leq q$.

Para esses vetores aleatórios, têm-se os seguintes parâmetros populacionais:

$$E\left(X^{(1)}\right) = \mu^{(1)} \text{ e } Cov\left(X^{(1)}\right) = \sum_{11}$$

$$E\left(X^{(2)}\right) = \mu^{(2)} \text{ e } Cov\left(X^{(2)}\right) = \sum_{22}$$

e a matriz de covariância cruzada entre os dois conjuntos é dada por:

$$Cov\left(X^{(1)}, X^{(2)}\right) = \sum_{12} = \sum_{21}'$$

Considerando os vetores $X^{(1)}$ e $X^{(2)}$conjuntamente, define-se o vetor aleatório

$$X = \begin{bmatrix} X^{(1)} \\ - \ - \ - \\ X^{(2)} \end{bmatrix} = \begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_p^{(1)} \\ - - - \\ X_1^{(2)} \\ X_2^{(2)} \\ \vdots \\ X_q^{(2)} \end{bmatrix}$$

com vetor de médias

$$\mu = \begin{bmatrix} \mu^{(1)} \\ - \ - \ - \\ \mu^{(2)} \end{bmatrix}$$

e matriz de variância-covariâncias

$$\sum = \begin{bmatrix} \sum_{11} & | & \sum_{12} \\ - - - & | & - - - \\ \sum_{21} & | & \sum_{22} \end{bmatrix}$$

A análise de correlação canônica baseia-se na definição de combinações lineares dos dois conjuntos de variáveis, dadas por

$$U = a'X^{(1)}$$

$$\mathrm{V} = b'X^{(2)}$$

Em que: a e b são vetores de coeficientes.

As variâncias e covariâncias dessas combinações lineares são:

$$V(U) = a'\sum_{11} a$$

$$V(V) = b'\sum_{22} \mathrm{b}$$

$$Cov(U,V) = a'\sum_{12} b$$

O objetivo do método é determinar os vetores a e b que maximizam a correlação entre U e V, que seja a maior possível e definida por:

$$Corr(U,V) = \frac{a'\sum_{12} b}{\sqrt{a'\sum_{11} a}\,\sqrt{b'\sum_{22} b}}$$

O procedimento inicia-se com a determinação do primeiro par de variáveis canônicas, U_1 e V_1, constituído por combinações lineares com variâncias unitárias, que maximizam essa correlação. O segundo par de variáveis canônicas, U_2 e V_2, é obtido de forma análoga, impondo-se adicionalmente que seja não correlacionado com o primeiro

par. Este processo continua sucessivamente até que se encontre o $p-ésimo$ par. A correlação associada ao $k-ésimo$ par é denominada k-ésima correlação canônica.

A maximização da correlação canônica resulta em:

$$Corr(U,V) = \frac{a'\sum_{12} b}{\sqrt{a'\sum_{11} a}\sqrt{b'\sum_{22} b}}:$$

$$\max_{a,b} Corr(U,V) = \rho_1^*$$

sendo o primeiro par de variáveis canônicas expresso como

$$U_1 = a_1' X^{(1)} = e_1' \sum_{11}^{-0,5} X^{(1)}$$
$$V_1 = b_1' X^{(2)} = f_1' \sum_{22}^{-0,5} X^{(2)}$$

De modo geral, para $k = 2,3,\ldots,p$, o $k-ésimo$ par de variáveis canônicas é dado por:

$$U_k = a_k' X^{(1)} = e_k' \sum_{11}^{-0,5} X^{(1)}$$
$$V_k = b_k' X^{(2)} = f_k' \sum_{22}^{-0,5} X^{(2)}$$

maximizando a correlação canônica $Corr(U,V) = \rho_k^*$ entre todas as combinações lineares não correlacionadas com as anteriores.

Os quadrados das correlações canônicas que satisfazem:

$$\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_k^{*2}$$

são os autovalores de

$$\sum_{11}^{-0,5} \sum_{12} \sum_{22}^{-1} \sum_{21} \sum_{11}^{-0,5}$$

com autovetores associados

$$\underline{e_1}, \underline{e_2}, \ldots, \underline{e_p}$$

As quantidades $\rho_1^{*2}, \rho_2^{*2}, \ldots, \rho_p^{*2}$ são também os p autovalores da matriz $\sum_{22}^{-0,5} \sum_{21} \sum_{11}^{-1} \sum_{12} \sum_{22}^{-0,5}$ com os correspondentes autovetores $\underline{f_1}, \underline{f_2}, \ldots, \underline{f_p}$.

No caso de variáveis padronizadas, considera-se os vetores

$$Z^{(1)} = \left[Z_1^{(1)}, Z_2^{(1)}, \ldots, Z_p^{(1)}\right]'$$
$$Z^{(2)} = \left[Z_1^{(2)}, Z_2^{(2)}, \ldots, Z_q^{(2)}\right]'$$

com matriz de correlações

$$Cov(Z) = \begin{bmatrix} \rho_{11} & | & \rho_{12} \\ --- & | & --- \\ \rho_{21} & | & \rho_{22} \end{bmatrix}$$

Nesse contexto, as variáveis canônicas assumem a forma:

$$U_k = a_k' Z^{(1)} = e_k' \rho_{11}^{-0,5} Z^{(1)}$$
$$V_k = b_k' Z^{(2)} = f_k' \rho_{22}^{-0,5} Z^{(2)}$$

Em que: e_k e f_k são os autovetores de $\rho_{11}^{-0,5} \rho_{12} \rho_{22}^{-1} \rho_{21} \rho_{11}^{-0,5}$ e $\rho_{22}^{-0,5} \rho_{21} \rho_{11}^{-1} \rho_{12} \rho_{22}^{-0,5}$, respectivamente.

A interpretação das variáveis canônicas U_k e V_k pode ser aprimorada por meio do cálculo das correlações entre as variáveis canônicas e as variáveis originais pertencentes aos conjuntos $X^{(1)}$ e $X^{(2)}$. Essas correlações permitem identificar quais variáveis originais contribuem de forma mais relevante para cada variável canônica e como se dá a associação entre os dois conjuntos analisados.

Nesse contexto, distinguem-se dois tipos de correlações: as correlações dentro do grupo e as correlações entre os grupos de variáveis. As correlações dentro do grupo avaliam o grau de associação entre uma variável canônica e as variáveis originais do mesmo conjunto ao qual ela pertence. Assim, a correlação $Corr\left(U_k, X^{(1)}\right)$ representa a

correlação de U_k com $X^{(1)}$, o que expressa a relação entre a $k-ésima$ variável canônica U_k e as variáveis do conjunto $X^{(1)}$. A correlação V_k com $X^{(2)}$ representa a correlação entre a $k-ésima$ variável canônica V_k e as variáveis do conjunto $X^{(2)}$.

Por outro lado, as correlações entre os grupos permitem analisar como cada variável canônica se relaciona com o conjunto oposto de variáveis originais. Nesse caso, a correlação U_k com $X^{(2)}$ mede o grau de associação entre a variável canônica U_k e as variáveis do conjunto $X^{(2)}$, enquanto a correlação de V_k com $X^{(1)}$ avalia a relação entre a variável canônica V_k e as variáveis do conjunto $X^{(1)}$.

Em resumo, tem-se as correlações dentro de cada grupo:

- corr(U_1, X^1): Contribuição das variáveis de X^1 na sua própria variável canônica
- corr(V_1, X^2): Contribuição das variáveis de X^2 na sua própria variável canônica

e as correlações entre grupos de variáveis:

- corr(V_1, X^1): Como X^1 se relaciona com a variável canônica de X^2
- corr(U_1, X^2): Como X^2 se relaciona com a variável canônica de X^1

Essas correlações são importantes para a interpretação da análise de correlação canônica, indicando quais variáveis originais são mais fortemente associadas a cada par de variáveis canônicas e, também, como cada conjunto de variáveis assimila a estrutura de dependência absorvida pelas correlações canônicas.

9.3 APLICAÇÃO PRÁTICA

9.3.1 Introdução

Em regressão múltipla mede-se o grau de relação entre um conjunto de variáveis independentes $(X's)$ e uma variável dependente (Y), a qual é dita como variável resposta. A técnica conhecida como correlação canônica é a que procura avaliar o grau de relação entre um conjunto p de variáveis estimadoras e um conjunto q de variáveis de saída (estimadas). Na realidade, correlação canônica é uma técnica perfeitamente simétrica, em que a distinção entre o conjunto estimador e o conjunto de saída (estimado) não é espelhado por qualquer diferença no tratamento estatístico dos dois conjuntos e a distinção então não necessariamente deve ser feita.

O processo em regressão múltipla é efetuado para se obter um estimador combinado, ou seja, uma combinação linear de escores nas p variáveis originais estimadoras $x_1, x_2, \dots, x_p$, e correlacioná-la combinada com um único escore Y. Na correlação canônica obtém-se duas combinações lineares: uma combinação de p variáveis estimadoras e uma combinação de q variáveis de saída (estimadas), $y_1, y_2, \dots, y_q$.

Assim, conhecendo-se a matriz de dispersão do conjunto completo das p e q variáveis, é possível calcular a correlação de qualquer combinação linear de variáveis $x's$ com uma dada combinação linear de variáveis $y's$. De todos os possíveis pares de combinações lineares, uma resulta com máxima correlação, a qual é denominada como a primeira correlação canônica e os pares correspondentes das combinações lineares das variáveis $x's$ e as variáveis $y's$ são denominadas de primeiras variáveis canônicas.

9.3.2 O conjunto de dados

O conjunto de dados é composto por 20 árvores de acácia-negra. Em cada uma das árvores foram medidos dois conjuntos de variáveis. O primeiro conjunto foi composto pelas variáveis biomassa do fuste, biomassa da copa e biomassa total. O segundo grupo foi composto pelas variáveis diâmetro, altura, comprimento de copa e massa específica do fuste. Assim, pode-se destacar que o primeiro conjunto é caracterizado por variáveis de biomassa e o segundo de variáveis dendrométricas e tecnológicas. O objetivo de se

efetuar uma análise de correlação canônica é avaliar a possibilidade de estimar as variáveis do grupo 1 por meio das variáveis do grupo 2.

Tabela 9.1: Dados de 20 árvores de acácia-negra, onde foram medidas as variáveis biomassa do fuste (BF, em kg), biomassa da copa (BC, em kg), biomassa total (BT, em kg), diâmetro (DAP, em cm), a altura (H, em m), comprimento de copa (CC, em m) e massa específica do fuste (MEF, g/cm^3).

n	X(1)			X(2)			
	BF	BC	BT	DAP	H	CC	MEF
1	174,45	31,64	206,09	19,74	20,20	7,70	0,65
2	84,41	8,76	93,17	15,12	19,10	4,90	0,50
3	60,97	4,91	65,88	12,89	17,30	3,00	0,57
4	47,25	2,06	49,30	11,62	16,90	7,20	0,52
5	114,06	27,95	142,00	16,55	20,50	11,80	0,56
6	157,39	37,48	194,87	19,10	21,90	7,00	0,60
7	72,03	15,14	87,17	14,32	18,40	9,20	0,56
8	88,98	15,66	104,63	15,12	19,30	12,10	0,58
9	45,29	5,96	51,25	11,94	16,70	9,50	0,50
10	101,20	18,46	119,65	15,28	19,50	7,70	0,59
11	94,16	12,22	106,38	16,23	18,30	4,60	0,52
12	83,79	10,13	93,92	15,12	19,30	3,80	0,59
13	200,00	54,52	254,51	20,53	19,50	9,50	0,58
14	49,66	6,04	55,70	13,21	18,40	2,80	0,57
15	20,96	3,59	24,55	8,75	14,40	11,20	0,48
16	71,90	15,34	87,24	13,05	18,00	3,50	0,62
17	115,57	21,25	136,82	16,23	19,60	9,00	0,62
18	69,30	11,50	80,80	12,25	17,90	3,80	0,63
19	16,46	2,49	18,94	7,64	9,90	6,70	0,51
20	107,51	19,67	127,17	15,44	19,20	6,80	0,56

Os principais resultados obtidos estão sumarizados na seguinte tabela.

Tabela 9.2: Resumo da análise de correlação canônica entre um grupo de variáveis de biomassa e outro grupo de variáveis dendrométrica e tecnológicas.

Variáveis de biomassa X^1	U_1	V_1	Variáveis Dendrométricas / Tecnológicas X^2	U_1	V_1
BF	0,995	0,981	DAP	0,971	0,985
BC	0,907	0,894	H	0,767	0,777
BT	0,986	0,972	CC	0,140	0,142
			ME	0,589	0,598

A análise revelou uma correlação canônica elevada no primeiro par de variáveis canônicas, com valor de 0,98, indicando uma associação linear muito forte entre o conjunto de variáveis de biomassa e o conjunto de variáveis dendrométricas e

tecnológicas. As correlações canônicas subsequentes apresentaram valores menores, evidenciando que a maior parte da relação multivariada entre os conjuntos é explicada essencialmente pela primeira dimensão canônica.

O teste de Wilks confirmou a significância estatística dessa associação. O valor da estatística Lambda de Wilks foi baixo, indicando pequena proporção de variância não explicada pela relação entre os conjuntos. O teste Qui-quadrado resultou em um valor elevado, maior que o valor crítico, resultando em rejeição da hipótese nula de independência multivariada, ou seja, existe uma relação estatisticamente significante entre biomassa (primeiro conjunto) e variáveis dendrométricas e tecnológicas (segundo conjunto).

No conjunto de biomassa, as correlações entre as variáveis originais e a primeira variável canônica associada a esse conjunto (U_1) foram altas e positivas. A biomassa do fuste teve a maior associação (r = 0,995), seguida da biomassa total (r = 0,986) e da biomassa da copa (r = 0,907). Esse padrão indica que U_1 representa adequadamente um índice global de biomassa, sendo fortemente dominado pela biomassa total e do fuste. Ainda, concernente ao primeiro conjunto, as correlações dessas variáveis com a variável canônica V_1 (associada ao segundo conjunto) também foram elevadas, com valores de 0,981 para biomassa do fuste, 0,972 para biomassa total e 0,894 para biomassa da copa. Isso indica que a biomassa está diretamente associada às características dendrométricas e tecnológicas das árvores.

No segundo conjunto (X^2), composto por DAP, H, CC e ME, observou-se que o DAP apresentou a maior correlação com a variável canônica U_1 (r = 0,971), indicando que é a principal determinante estrutural da relação com a biomassa, seguido da Altura (r = 0,767). Essas duas variáveis são as mais influentes na explicação da biomassa, indicando que árvores com maior diâmetro e maior altura também detém os maiores valores de biomassa.

A massa específica do fuste apresentou uma correlação moderada (r = 0,589) com U_1, sugerindo que, embora contribua para a biomassa, seu efeito é menor quando comparado com o DAP e H. O comprimento da copa apresentou correlação baixa (r = 0,140), indicando que, neste conjunto de dados, exerce pouca influência sobre a biomassa, sugerindo menor associação direta com o fator de porte das árvores. As correlações com a variável canônica V_1 confirmam esse padrão: DAP (0,985) e H (0,777), que são novamente as variáveis mais associadas, enquanto CC e ME contribuem pouco para explicar a biomassa.

De forma geral, os resultados mostram que a biomassa das árvores é fortemente explicada por variáveis dimensionais, principalmente DAP e H, enquanto as demais têm papel secundário nesse conjunto amostral. Esse resultado indica que árvores com maiores valores de DAP e H também apresentam maiores valores de biomassa total, de fuste e de copa, conforme o que se espera nessas relações. Por outro lado, a contribuição relativamente menor da massa específica do fuste evidencia que a densidade da madeira varia de forma mais independente em relação ao porte da árvore, não acompanhando necessariamente o aumento do tamanho das árvores e da biomassa, o que acontece também para a altura de copa. Assim, para esse conjunto de dados, essas duas variáveis têm pouca importância para a modelagem, podendo-se sugerir um segundo conjunto de dados apenas dendrométrico e composto por DAP e H.

A análise com variáveis padronizadas revelou as mesmas respostas. Mesmo após a remoção do efeito de escala, as correlações canônicas permaneceram elevadas e a hierarquia das cargas canônicas se manteve praticamente inalterada, com o DAP e a biomassa total permanecendo como as variáveis mais importantes.

9.4 ANÁLISE DE CORRELAÇÃO CANÔNICA NO MATLAB

```
% ANÁLISE DE CORRELAÇÃO CANÔNICA

% MATRIZ DE DADOS X (N × (P+Q))
X =
```

174.45	31.64	206.09	19.74	20.2	7.7	0.65
84.41	8.76	93.17	15.12	19.1	4.9	0.5
60.97	4.91	65.88	12.89	17.3	3	0.57
47.25	2.06	49.3	11.62	16.9	7.2	0.52
114.06	27.95	142	16.55	20.5	11.8	0.56
157.39	37.48	194.87	19.1	21.9	7	0.6
72.03	15.14	87.17	14.32	18.4	9.2	0.56
88.98	15.66	104.63	15.12	19.3	12.1	0.58
45.29	5.96	51.25	11.94	16.7	9.5	0.5
101.2	18.46	119.65	15.28	19.5	7.7	0.59
94.16	12.22	106.38	16.23	18.3	4.6	0.52
83.79	10.13	93.92	15.12	19.3	3.8	0.59
200	54.52	254.51	20.53	19.5	9.5	0.58
49.66	6.04	55.7	13.21	18.4	2.8	0.57
20.96	3.59	24.55	8.75	14.4	11.2	0.48
71.9	15.34	87.24	13.05	18	3.5	0.62
115.57	21.25	136.82	16.23	19.6	9	0.62
69.3	11.5	80.8	12.25	17.9	3.8	0.63
16.46	2.49	18.94	7.64	9.9	6.7	0.51
107.51	19.67	127.17	15.44	19.2	6.8	0.56

```
% DEFINIÇÃO DOS CONJUNTOS DE VARIÁVEIS
% X1: PRIMEIRO CONJUNTO (N × P)
X1 = X(:,1:3)
X1 =
```

174.4500	31.6400	206.0900
84.4100	8.7600	93.1700
60.9700	4.9100	65.8800
47.2500	2.0600	49.3000
114.0600	27.9500	142.0000
157.3900	37.4800	194.8700
72.0300	15.1400	87.1700
88.9800	15.6600	104.6300
45.2900	5.9600	51.2500
101.2000	18.4600	119.6500
94.1600	12.2200	106.3800
83.7900	10.1300	93.9200
200.0000	54.5200	254.5100
49.6600	6.0400	55.7000
20.9600	3.5900	24.5500
71.9000	15.3400	87.2400
115.5700	21.2500	136.8200
69.3000	11.5000	80.8000
16.4600	2.4900	18.9400
107.5100	19.6700	127.1700

```
% X2: SEGUNDO CONJUNTO (n × q)
X2 = X(:,4:7)
X2 =
```

19.7400	20.2000	7.7000	0.6500
15.1200	19.1000	4.9000	0.5000
12.8900	17.3000	3.0000	0.5700
11.6200	16.9000	7.2000	0.5200
16.5500	20.5000	11.8000	0.5600
19.1000	21.9000	7.0000	0.6000
14.3200	18.4000	9.2000	0.5600

```
   15.1200   19.3000   12.1000    0.5800
   11.9400   16.7000    9.5000    0.5000
   15.2800   19.5000    7.7000    0.5900
   16.2300   18.3000    4.6000    0.5200
   15.1200   19.3000    3.8000    0.5900
   20.5300   19.5000    9.5000    0.5800
   13.2100   18.4000    2.8000    0.5700
    8.7500   14.4000   11.2000    0.4800
   13.0500   18.0000    3.5000    0.6200
   16.2300   19.6000    9.0000    0.6200
   12.2500   17.9000    3.8000    0.6300
    7.6400    9.9000    6.7000    0.5100
   15.4400   19.2000    6.8000    0.5600

% NÚMERO DE OBSERVAÇÕES
n = size(X1,1)
n =
    20

% MATRIZES DE VARIÂNCIA-COVARIÂNCIA

% Calcula a matriz de covariância conjunta de todas as variáveis
% Dimensão: (p+q) × (p+q)
% As p primeiras colunas/linhas correspondem a X1
% As q últimas colunas/linhas correspondem a X2
C = cov([X1 X2])
C =
     2265.6       596.1      2861.7      149.89      89.048        26.9      1.3068
      596.1      176.58      772.67      37.311      20.553      12.739     0.31925
     2861.7      772.67      3634.3       187.2       109.6      39.633       1.626
     149.89      37.311       187.2      10.749      7.1061      1.1991    0.085557
     89.048      20.553       109.6      7.1061      6.3603     0.34279    0.068071
       26.9      12.739      39.633      1.1991     0.34279      8.7136   -0.027521
     1.3068     0.31925       1.626    0.085557    0.068071   -0.027521   0.0022787

% NÚMERO DE VARIÁVEIS EM CADA CONJUNTO
p = size(X1,2);     % quantidade de variáveis no primeiro conjunto
(X1)
p =
     3
q = size(X2,2);     % quantidade de variáveis no segundo conjunto
(X2)
q =
     4

% MATRIZ DE COVARIÂNCIA DO PRIMEIRO CONJUNTO DE VARIÁVEIS
% Dimensão: p × p
VC11 = C(1:p, 1:p)
VC11 =
      2265.6       596.1      2861.7
       596.1      176.58      772.67
      2861.7      772.67      3634.3

% MATRIZ DE COVARIÂNCIA DO SEGUNDO CONJUNTO DE VARIÁVEIS
% Dimensão: q × q
VC22 = C(p+1:p+q, p+1:p+q)
VC22 =
      10.749      7.1061      1.1991    0.085557
      7.1061      6.3603     0.34279    0.068071
```

```
       1.1991       0.34279        8.7136    -0.027521
     0.085557      0.068071    -0.027521     0.0022787

% MATRIZ DE COVARIÂNCIA CRUZADA ENTRE X1 E X2
% Cada elemento representa cov(X1_i, X2_j)
% Dimensão: p × q
VC12 = C(1:p, p+1:p+q)
VC12 =
       149.89        89.048          26.9        1.3068
       37.311        20.553        12.739       0.31925
        187.2         109.6        39.633         1.626

% TRANSPOSTA DA MATRIZ DE COVARIÂNCIA CRUZADA
% Representa cov(X2, X1)
% Dimensão: q × p
VC21 = VC12'
VC21 =
       149.89        37.311         187.2
       89.048        20.553         109.6
         26.9        12.739        39.633
       1.3068       0.31925         1.626

% MATRIZ DE COVARIÂNCIA TOTAL
% Dimensão: (p+q) × (p+q)
VC = C
VC =
      2265.6        596.1       2861.7       149.89       89.048         26.9       1.3068
       596.1       176.58       772.67       37.311       20.553       12.739      0.31925
      2861.7       772.67       3634.3        187.2        109.6       39.633        1.626
      149.89       37.311        187.2       10.749       7.1061       1.1991     0.085557
      89.048       20.553        109.6       7.1061       6.3603      0.34279     0.068071
        26.9       12.739       39.633       1.1991      0.34279       8.7136    -0.027521
      1.3068      0.31925        1.626     0.085557     0.068071    -0.027521    0.0022787

% DETERMINAÇÃO DOS COEFICIENTES CANÔNICOS

% MATRIZ ASSOCIADA AOS COEFICIENTES A
PARESA = inv(sqrtm(VC11)) * VC12 * inv(VC22) * VC21 *
inv(sqrtm(VC11))
PARESA =
       0.43754   -0.0077452       0.47332
    -0.0077452       0.38777      0.026234
       0.47332      0.026234       0.55317

% AUTOVALORES E AUTOVETORES DE A
[VA,DA] = eig(PARESA)
VA =
       0.6622      -0.74666    -0.063224
      0.024826    -0.062466       0.99774
       0.74892       0.66227     0.022828
DA =
       0.97255             0             0
             0      0.017064             0
             0             0       0.38886

% COEFICIENTES CANÔNICOS A
a = inv(sqrtm(VC11)) * VA
```

```
a =
      -7.7227      -184.37       114.06
      -7.7666      -184.15       114.25
       7.7485       184.33       -114.1

% MATRIZ ASSOCIADA AOS COEFICIENTES B
PARESB = inv(sqrtm(VC22)) * VC21 * inv(VC11) * VC12 *
inv(sqrtm(VC22))
PARESB =
      0.86713      0.26639     0.089488      0.10774
      0.26639      0.16979     -0.13268    0.0034504
     0.089488     -0.13268      0.30132       0.0625
      0.10774    0.0034504       0.0625     0.040236

% AUTOVALORES E AUTOVETORES DE B
[VB,DB] = eig(PARESB)
VB =
      0.94353     0.054497      -0.3115    -0.098755
      0.30086     -0.45852      0.83606    -0.015679
     0.077058      0.87088       0.4463     -0.19093
      0.11532      0.16843     0.069185       0.9765
DB =
      0.97255            0            0            0
            0      0.38886            0            0
            0            0   1.3395e-13            0
            0            0            0     0.017064

% COEFICIENTES CANÔNICOS B
b = inv(sqrtm(VC22)) * VB
b =
      0.35568      0.13207     -0.46069     -0.15697
      -0.1214     -0.35083      0.69238     -0.13252
     0.013166      0.30518      0.18665      0.04174
       2.9489        4.688       1.8175       26.263

% CORRELAÇÕES CANÔNICAS
% As correlações canônicas são as raízes dos autovalores
RUV = sqrt(diag(DA))
RUV =
      0.98618
      0.13063
      0.62359

% ORDENAÇÃO EM ORDEM DECRESCENTE
[RUV,idx] = sort(RUV,'descend')
RUV =
      0.98618
      0.62359
      0.13063
idx =
     1
     3
     2
a = a(:,idx)
a =
      -7.7227       114.06      -184.37
      -7.7666       114.25      -184.15
       7.7485       -114.1       184.33
```

```
% PRIMEIRA FUNÇÃO CANÔNICA
a1 = a(:,1)
a1 =
      -7.7227
      -7.7666
       7.7485
[~,idxB] = sort(diag(DB),'descend');
idxB =
     1
     2
     4
     3
b1 = b(:,idxB(1));
b1 =
      0.35568
      -0.1214
     0.013166
       2.9489

% CORRELAÇÕES ENTRE VARIÁVEIS CANÔNICAS E VARIÁVEIS ORIGINAIS
DVC11   = diag(diag(VC11))
DVC11 =
       2265.6            0            0
            0       176.58            0
            0            0       3634.3
DVC22   = diag(diag(VC22))
DVC22 =
       10.749            0            0            0
            0       6.3603            0            0
            0            0       8.7136            0
            0            0            0    0.0022787
IRDVC11 = inv(sqrtm(DVC11))
IRDVC11 =
     0.021009            0            0
            0     0.075253            0
            0            0     0.016588
IRDVC22 = inv(sqrtm(DVC22))
IRDVC22 =
      0.30501            0            0            0
            0      0.39652            0            0
            0            0      0.33877            0
            0            0            0       20.949
RU1X1 = a1' * VC11 * IRDVC11   % U1 com X1
RU1X1 =
      0.99508       0.9069      0.98556
RU1X2 = a1' * VC12 * IRDVC22  % U1 com X2
RU1X2 =
      0.97117      0.76667      0.13979       0.5893
RV1X1 = b1' * VC21 * IRDVC11  % V1 com X1
RV1X1 =
      0.98132      0.89436      0.97194
RV1X2 = b1' * VC22 * IRDVC22  % V1 com X2
RV1X2 =
      0.98478      0.77741      0.14175      0.59756

% DADOS TRANSFORMADOS: VARIÁVEIS CANÔNICAS
U = X1 * a  % Variáveis canônicas do primeiro conjunto
```

```
U =
       3.9227       -2.045      -2.1239
       2.0169      -1.9777      -2.2198
       1.4825      -1.6424      -1.8614
       1.1031     -0.39515      -3.6034
       2.3555      0.89987        -1.91
        3.377     -0.49687     -0.29003
        1.582     -0.53112     -0.49911
        1.932     0.025771      -3.0137
       1.0593     -0.87105     -0.96172
       2.1962    -0.029406      -3.0562
       2.2054      -1.8368      -2.0306
        1.976      -1.7468      -1.9424
       4.0887       1.5497     -0.95715
       1.1705      -1.0298      -1.1446
      0.47514     -0.27761     -0.29369
        1.575     -0.49586     -0.45683
       2.5935      -1.3111      -1.3544
       1.5777     -0.97365      -1.0383
      0.30162       0.8733      -2.1338
       2.3369    -0.093481      -3.1211

V = X2 * b   % Variáveis canônicas do segundo conjunto
V =
       6.5871      0.91742       7.5106       11.617
       4.5982     -0.86456       8.0822       8.4317
       4.2049     -0.77927       7.6358       10.779
       3.7096       0.2407        8.637       9.8939
       5.2046       1.2202       9.7896       9.8855
       5.9964     -0.21155        8.761        10.15
       4.6322      0.86894       8.8777       10.405
       4.9046       1.6376       9.7099       10.807
        3.819      0.96129        8.744       9.4409
       4.9088      0.29268       8.9716       10.834
       5.1451     -0.43509       6.9972       8.8762
       4.8248     -0.84851       8.1789       10.723
       6.7703       1.4885       6.8708       9.8226
       4.1826       -1.184       8.2127       10.575
        2.927        1.772       8.9021       9.7921
       4.3309     -0.61673        8.231       11.996
       5.3401      0.92044       8.9003       11.514
       4.0919     -0.54887       8.6044       12.409
       3.1077       1.9714       5.5124       11.163
       4.9018    0.0037515       8.4677       10.023

U1 = U(:,1)   % Primeira variável canônica de X1
U1 =
       3.9227
       2.0169
       1.4825
       1.1031
       2.3555
        3.377
        1.582
        1.932
       1.0593
       2.1962
       2.2054
```

```
        1.976
       4.0887
       1.1705
      0.47514
        1.575
       2.5935
       1.5777
      0.30162
       2.3369

V1 = V(:,1)   % Primeira variável canônica de X2

% VERIFICAÇÃO DA CORRELAÇÃO CANÔNICA
cor_U1V1 = corr(U1,V1)
cor_U1V1 =
      0.98618

% ESTATÍSTICA DE WILKS E TESTE QUI-QUADRADO

% MATRIZ DE RESÍDUOS DA REGRESSÃO
V11_star = VC11 - VC12 * inv(VC22) * VC21
V11_star =
       75.466       31.976       107.43
       31.976       25.606       57.578
       107.43       57.578          165

% SOMAS DE QUADRADOS (DETERMINANTES)
SQTotal   = det(VC11)
SQTotal =
      0.95048
SQResiduo = det(V11_star)
SQResiduo =
     0.015673

% CRITÉRIO DE WILKS
L = SQResiduo / SQTotal
L =
      0.01649

% ESTATÍSTICA QUI-QUADRADO
p = size(X1,2)
p =
     3
q = size(X2,2)
q =
     4
chi2 = (-n + 0.5*(p+q+1)) * log(L)
chi2 =
        65.68
df   = (p-1)*(q-1)
df =
     6
p_valor_X = 1 - chi2cdf(chi2,df)
p_valor_X =
   3.1325e-12
```

```
% RESUMO DOS RESULTADOS — VARIÁVEIS ORIGINAIS
Resumo_X1 = [RU1X1' RV1X1']
Resumo_X1 =
      0.99508      0.98132
       0.9069      0.89436
      0.98556      0.97194
Resumo_X2 = [RU1X2' RV1X2']
Resumo_X2 =
      0.97117      0.98478
      0.76667      0.77741
      0.13979      0.14175
       0.5893      0.59756

% ANÁLISE DE CORRELAÇÃO A PARTIR DAS VARIÁVEIS PADRONIZADAS
% Padronização das variáveis
Z1 = (X1 - mean(X1)) ./ std(X1);
Z2 = (X2 - mean(X2)) ./ std(X2);

% MATRIZES DE CORRELAÇÃO
R11 = corr(Z1)
R11 =
           1      0.94245      0.99728
      0.94245            1      0.96452
      0.99728      0.96452            1
R22 = corr(Z2)
R22 =
           1      0.85941       0.1239      0.54667
      0.85941            1     0.046046      0.56543
       0.1239     0.046046            1     -0.19531
      0.54667      0.56543     -0.19531            1
R12 = corr(Z1,Z2)
R12 =
      0.96047      0.74182      0.19145      0.57513
      0.85639      0.61329      0.32477      0.50328
      0.94711       0.7209      0.22271      0.56504
R21 = R12'
R21 =
      0.96047      0.85639      0.94711
      0.74182      0.61329       0.7209
      0.19145      0.32477      0.22271
      0.57513      0.50328      0.56504
R   = corr([Z1 Z2])
R =
          1      0.94245      0.99728      0.96047      0.74182      0.19145      0.57513
      0.94245            1      0.96452      0.85639      0.61329      0.32477      0.50328
      0.99728      0.96452            1      0.94711       0.7209      0.22271      0.56504
      0.96047      0.85639      0.94711            1      0.85941       0.1239      0.54667
      0.74182      0.61329       0.7209      0.85941            1     0.046046      0.56543
      0.19145      0.32477      0.22271       0.1239     0.046046            1     -0.19531
      0.57513      0.50328      0.56504      0.54667      0.56543     -0.19531            1

% COEFICIENTES CANÔNICOS PADRONIZADOS
PARESA_Z = inv(sqrtm(R11)) * R12 * inv(R22) * R21 * inv(sqrtm(R11))
PARESA_Z =
       0.4382      0.23238      0.41356
      0.23238      0.42873      0.06994
      0.41356      0.06994      0.51155
[VAZ,DAZ] = eig(PARESA_Z)
```

```
VAZ =
      0.66389      0.74782    0.0021876
      0.36746     -0.32367      -0.8719
      0.65132     -0.57966      0.48967
DAZ =
      0.97255            0            0
            0     0.017064            0
            0            0      0.38886
a_Z = inv(sqrtm(R11)) * VAZ
a_Z =
      -367.58       8775.8        -5429
      -103.21         2447      -1518.2
       467.12       -11112       6878.5
PARESB_Z = inv(sqrtm(R22)) * R21 * inv(R11) * R12 * inv(sqrtm(R22))
PARESB_Z =
      0.75161      0.26716      0.11606      0.28067
      0.26716      0.18983     -0.12585      0.10609
      0.11606     -0.12585      0.31303     0.028413
      0.28067      0.10609     0.028413      0.12401
[VBZ,DBZ] = eig(PARESB_Z)
VBZ =
      0.87713     0.073264     -0.38758     -0.27397
      0.32766     -0.46675      0.79619     -0.20214
      0.10624      0.88122      0.45548    -0.068548
      0.33466    -0.014792     0.091687      0.93775
DBZ =
      0.97255            0            0            0
            0      0.38886            0            0
            0            0  -8.5864e-14            0
            0            0            0     0.017064
b_Z = inv(sqrtm(R22)) * VBZ
b_Z =
       1.1661      0.43301      -1.5104     -0.51463
     -0.30616     -0.88479       1.7462      -0.3342
     0.038863      0.90086      0.55096      0.12321
      0.14077      0.22379     0.086758       1.2537

% CORRELAÇÕES CANÔNICAS PADRONIZADAS
RUV_Z = sqrt(diag(DAZ))
RUV_Z =
      0.98618
      0.13063
      0.62359
[RUV_Z,idx_Z] = sort(RUV_Z,'descend')
RUV_Z =
      0.98618
      0.62359
      0.13063
idx_Z =
     1
     3
     2
a_Z = a_Z(:,idx_Z)
a_Z =
      -367.58        -5429       8775.8
      -103.21      -1518.2         2447
       467.12       6878.5      -11112
```

```
% PRIMEIRA FUNÇÃO CANÔNICA PADRONIZADA
a1_Z = a_Z(:,1)
a1_Z =
      -367.58
      -103.21
       467.12
[~,idxBZ] = sort(diag(DBZ),'descend');
idxBZ =
     1
     2
     4
     3
b1_Z = b_Z(:,idxBZ(1))
b1_Z =
       1.1661
     -0.30616
     0.038863
      0.14077

% CORRELAÇÕES ENTRE VARIÁVEIS CANÔNICAS E VARIÁVEIS PADRONIZADAS
RU1Z1 = a1_Z' * R11
RU1Z1 =
      0.99508       0.9069      0.98556
RU1Z2 = a1_Z' * R12
RU1Z2 =
      0.97117      0.76667      0.13979       0.5893
RV1Z1 = b1_Z' * R21
RV1Z1 =
      0.98132      0.89436      0.97194
RV1Z2 = b1_Z' * R22
RV1Z2 =
      0.98478      0.77741      0.14175      0.59756

% VARIÁVEIS CANÔNICAS PADRONIZADAS
U_Z = Z1 * a_Z
U_Z =
       1.9563       1.4247      0.42331
     0.050476       1.3575       0.5192
     -0.48388       1.0222      0.16077
     -0.86331      -0.2251       1.9028
      0.38914      -1.5201      0.20935
       1.4106     -0.12338      -1.4106
     -0.38437     -0.08914      -1.2015
     -0.03438     -0.64603       1.3131
     -0.90709      0.25079     -0.73889
      0.22981     -0.59085       1.3556
      0.23903       1.2165          0.33
    0.0096424       1.1265      0.24182
       2.1224        -2.17     -0.74347
      -0.7959      0.40954     -0.55598
      -1.4912     -0.34264      -1.4069
     -0.39135      -0.1244      -1.2438
      0.62709      0.69087     -0.34617
     -0.38871      0.35339     -0.66232
      -1.6648      -1.4936      0.43323
      0.37051     -0.52677       1.4205
```

```
V_Z = Z2 * b_Z
V_Z =
       1.8777       0.5771     -0.76921       1.1602
     -0.11121      -1.2049     -0.19768      -2.0252
     -0.50446      -1.1196     -0.64404      0.32245
     -0.99976    -0.099616      0.35714     -0.56304
      0.49523      0.87987       1.5098     -0.57141
        1.287     -0.55187      0.48115     -0.30703
    -0.077232      0.52862      0.59784    -0.051617
      0.19521       1.2973       1.4301      0.34986
     -0.89036      0.62098      0.46418       -1.016
       0.1994    -0.047643      0.69175      0.37721
      0.43574     -0.77541      -1.2826      -1.5807
      0.11543      -1.1888     -0.10095      0.26604
       2.0609       1.1482      -1.4091     -0.63436
     -0.52681      -1.5243    -0.067172      0.11811
      -1.7824       1.4316      0.62226     -0.66482
      -0.3785     -0.95705    -0.048886       1.5386
      0.63074      0.58012       0.6205        1.057
     -0.61746     -0.88919       0.3246       1.9526
      -1.6017       1.6311      -2.7675      0.70581
      0.19241     -0.33657      0.18782     -0.43361
U1_Z = U_Z(:,1)
U1_Z =
       1.9563
     0.050476
     -0.48388
     -0.86331
      0.38914
       1.4106
     -0.38437
     -0.03438
     -0.90709
      0.22981
      0.23903
    0.0096424
       2.1224
      -0.7959
      -1.4912
     -0.39135
      0.62709
     -0.38871
      -1.6648
      0.37051
V1_Z = V_Z(:,1)
V1_Z =
       1.8777
     -0.11121
     -0.50446
     -0.99976
      0.49523
        1.287
    -0.077232
      0.19521
     -0.89036
       0.1994
      0.43574
      0.11543
```

```
       2.0609
     -0.52681
      -1.7824
      -0.3785
      0.63074
     -0.61746
      -1.6017
      0.19241

% VERIFICAÇÃO DA CORRELAÇÃO CANÔNICA PADRONIZADA
cor_U1Z_V1Z = corr(U1_Z,V1_Z)
cor_U1Z_V1Z =
      0.98618

% ESTATÍSTICA DE WILKS (PADRONIZADA)
R11_star = R11 - R12 * inv(R22) * R21
R11_star =
      0.03331     0.050554     0.037441
     0.050554      0.14501     0.071874
     0.037441     0.071874     0.045401
SQTotal_Z   = det(R11)
SQTotal_Z =
   6.5372e-10
SQResiduo_Z = det(R11_star)
SQResiduo_Z =
    1.078e-11
L_Z = SQResiduo_Z / SQTotal_Z
L_Z =
      0.01649
chi2_Z = (-n + 0.5*(p+q+1)) * log(L_Z)
chi2_Z =
        65.68
df_Z   = (p-1)*(q-1)
df_Z =
     6
p_valor_Z = 1 - chi2cdf(chi2_Z,df_Z)
p_valor_Z =
   3.1325e-12

% RESUMO — VARIÁVEIS PADRONIZADAS
Resumo_Z1 = [RU1Z1' RV1Z1']
Resumo_Z1 =
      0.99508      0.98132
       0.9069      0.89436
      0.98556      0.97194
Resumo_Z2 = [RU1Z2' RV1Z2']
Resumo_Z2 =
      0.97117      0.98478
      0.76667      0.77741
      0.13979      0.14175
       0.5893      0.59756
```

9.5 ANÁLISE DE CORRELAÇÃO CANÔNICA NO JULIA

```
# Análise de correlação canônica

# Pacotes necessários

using LinearAlgebra; Statistics; Distributions

# Entrada dos dados

# Matriz de dados X (n × (p+q))
X = [
   174.45   31.64   206.09  19.74   20.20   7.70    0.65
84.41   8.76    93.17   15.12   19.10   4.90    0.50
60.97   4.91    65.88   12.89   17.30   3.00    0.57
47.25   2.06    49.30   11.62   16.90   7.20    0.52
114.06  27.95   142.00  16.55   20.50   11.80   0.56
157.39  37.48   194.87  19.10   21.90   7.00    0.60
72.03   15.14   87.17   14.32   18.40   9.20    0.56
88.98   15.66   104.63  15.12   19.30   12.10   0.58
45.29   5.96    51.25   11.94   16.70   9.50    0.50
101.20  18.46   119.65  15.28   19.50   7.70    0.59
94.16   12.22   106.38  16.23   18.30   4.60    0.52
83.79   10.13   93.92   15.12   19.30   3.80    0.59
200.00  54.52   254.51  20.53   19.50   9.50    0.58
49.66   6.04    55.70   13.21   18.40   2.80    0.57
20.96   3.59    24.55   8.75    14.40   11.20   0.48
71.90   15.34   87.24   13.05   18.00   3.50    0.62
115.57  21.25   136.82  16.23   19.60   9.00    0.62
69.30   11.50   80.80   12.25   17.90   3.80    0.63
16.46   2.49    18.94   7.64    9.90    6.70    0.51
107.51  19.67   127.17  15.44   19.20   6.80    0.56
]

# Definição dos conjuntos de variáveis
# X1: primeiro conjunto (n × p)
X1 = X[:, 1:3]

# X2: segundo conjunto (n × q)
X2 = X[:, 4:7]

# Número de observações
n = size(X1, 1)

# Matrizes de variância–covariância
VC11 = cov(X1)          # Covariância de X1 (p × p)
VC22 = cov(X2)          # Covariância de X2 (q × q)
VC12 = cov(X1, X2)      # Covariância cruzada (p × q)
VC21 = VC12'            # Transposta (q × p)
VC   = cov(X)           # Covariância total
```

```
# Determinação dos coeficientes canônicos

# Matriz associada aos coeficientes a
PARESA = inv(sqrt(VC11)) * VC12 * inv(VC22) * VC21 * inv(sqrt(VC11))

# Autovalores e autovetores de a
EigA = eigen(PARESA)

# Coeficientes canônicos a
a = inv(sqrt(VC11)) * EigA.vectors

# Matriz associada aos coeficientes b
PARESB = inv(sqrt(VC22)) * VC21 * inv(VC11) * VC12 * inv(sqrt(VC22))

# Autovalores e autovetores de b
EigB = eigen(PARESB)

# Coeficientes canônicos b
b = inv(sqrt(VC22)) * EigB.vectors

# Correlações canônicas
RUV = sqrt.(EigA.values)

# Ordenação em ordem decrescente
idx = sortperm(RUV, rev = true)
RUV = RUV[idx]

# Reordenar coeficientes a
a = a[:, idx]

# Primeira função canônica
a1 = a[:, 1]
b1 = b[:, sortperm(EigB.values, rev = true)[1]]

# Correlações entre variáveis canônicas e variáveis originais
DVC11   = diagm(diag(VC11))
DVC22   = diagm(diag(VC22))
IRDVC11 = inv(sqrt(DVC11))
IRDVC22 = inv(sqrt(DVC22))

RU1X1 = a1' * VC11 * IRDVC11   # U1 com X1
RU1X2 = a1' * VC12 * IRDVC22   # U1 com X2
RV1X1 = b1' * VC21 * IRDVC11   # V1 com X1
RV1X2 = b1' * VC22 * IRDVC22   # V1 com X2

# Dados transformados: variáveis canônicas
# As variáveis canônicas são combinações lineares das variáveis originais
# U = X1 * A  (variáveis canônicas do primeiro conjunto)
# V = X2 * B  (variáveis canônicas do segundo conjunto)
```

```
U = X1 * a          # Matriz (n × p) de variáveis canônicas de X1
V = X2 * b          # Matriz (n × q) de variáveis canônicas de X2

# Primeiro par de variáveis canônicas
U1 = U[:, 1]        # Primeira variável canônica associada a X1
V1 = V[:, 1]        # Primeira variável canônica associada a X2

# Verificação da correlação canônica
# Correlação entre o primeiro par canônico (U1, V1)
cor_U1V1 = cor(U1, V1)

# Estatística de Wilks e teste qui-quadrado
# V11* = V11 - V12 * V22^{-1} * V21
# V11 = VC11 (covariância de X1)
# V22 = VC22 (covariância de X2)
# V12 = VC12 (covariância cruzada)
# V21 = VC21 (transposta)
V11_star = VC11 - VC12 * inv(VC22) * VC21

# Somas de quadrados
SQTotal   = det(VC11)
SQResiduo = det(V11_star)

# Critério de Wilks
L = SQResiduo / SQTotal

# Estatística qui-quadrado
p = size(X1, 2)   # número de variáveis em X1
q = size(X2, 2)   # número de variáveis em X2
chi2 = (-n + 0.5 * (p + q + 1)) * log(L)
df = (p - 1) * (q - 1)
L        # Estatística de Wilks
chi2     # Valor do qui-quadrado
_df = df # Graus de liberdade
p_valor_X = ccdf(Chisq(_df), chi2)

# Resumo dos resultados para as variáveis originais
#  Conjunto X1
# Linhas: variáveis originais de X1
# Colunas: correlação com U1 (RU1X1) e V1 (RV1X1)

Resumo_X1 = [RU1X1' RV1X1']

# Conjunto X2
# Linhas: variáveis originais de X2
# Colunas: correlação com U1 (RU1X2) e V1 (RV1X2)

Resumo_X2 = [RU1X2' RV1X2']
```

```
# Análise de correlação a partir das variáveis padronizadas
# Padronização das variáveis
Z1 = (X1 .- mean(X1, dims=1)) ./ std(X1, dims=1)
Z2 = (X2 .- mean(X2, dims=1)) ./ std(X2, dims=1)

#Matrizes de correlação
R11 = cor(Z1)          # Correlação de Z1
R22 = cor(Z2)          # Correlação de Z2
R12 = cor(Z1, Z2)      # Correlação cruzada
R21 = R12'
R   = cor([Z1 Z2])     # Correlação total

# Determinação dos coeficientes canônicos padronizados
# Matriz associada aos coeficientes a*
PARESA_Z = inv(sqrt(R11)) * R12 * inv(R22) * R21 * inv(sqrt(R11))
EigA_Z = eigen(PARESA_Z)
a_Z = inv(sqrt(R11)) * EigA_Z.vectors

# Matriz associada aos coeficientes b*
PARESB_Z = inv(sqrt(R22)) * R21 * inv(R11) * R12 * inv(sqrt(R22))
EigB_Z = eigen(PARESB_Z)
b_Z = inv(sqrt(R22)) * EigB_Z.vectors

# Correlações canônicas padronizadas
RUV_Z = sqrt.(EigA_Z.values)

# Ordenação decrescente
idx_Z = sortperm(RUV_Z, rev=true)
RUV_Z = RUV_Z[idx_Z]

# Reordenar coeficientes
a_Z = a_Z[:, idx_Z]

# Primeira função canônica padronizada
a1_Z = a_Z[:, 1]
b1_Z = b_Z[:, sortperm(EigB_Z.values, rev=true)[1]]

# Correlações entre variáveis canônicas e variáveis originais padronizadas
RU1Z1 = a1_Z' * R11    # U1* com Z1
RU1Z2 = a1_Z' * R12    # U1* com Z2
RV1Z1 = b1_Z' * R21    # V1* com Z1
RV1Z2 = b1_Z' * R22    # V1* com Z2

# Variáveis canônicas padronizadas
U_Z = Z1 * a_Z
V_Z = Z2 * b_Z
U1_Z = U_Z[:, 1]
V1_Z = V_Z[:, 1]

# Verificação da correlação canônica padronizada
cor_U1Z_V1Z = cor(U1_Z, V1_Z)
```

```
# Estatística de Wilks (padronizada)
R11_star = R11 - R12 * inv(R22) * R21
SQTotal_Z   = det(R11)
SQResiduo_Z = det(R11_star)
L_Z = SQResiduo_Z / SQTotal_Z
chi2_Z = (-n + 0.5 * (p + q + 1)) * log(L_Z)
df_Z = (p - 1) * (q - 1)
p_valor_Z = ccdf(Chisq(df_Z), chi2_Z)

# Resumo dos resultados – variáveis padronizadas
# Conjunto Z1
Resumo_Z1 = [
    RU1Z1' RV1Z1']

#Conjunto Z2
Resumo_Z2 = [RU1Z2' RV1Z2']
```

REFERÊNCIAS

ANDERSON, T. W. *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1958. 374 p.

BARTLETT, M. S. Tests of significance in factor analysis. *British Journal of Psychology*, v. 3, p. 77–85, 1950.

CATTELL, R. B. The scree test for the number of factors. *Multivariate Behavioral Research*, v. 1, n. 2, p. 245–276, 1966.

CATTELL, R. B. (Ed.). *Handbook of multivariate experimental psychology*. Skokie, Ill.: Rand McNally, 1966. 959 p.

COOLEY, W. W.; LOHNES, P. R. *Multivariate data analysis*. New York: John Wiley & Sons, 1971. 364 p.

FINN, J. D. A general model for multivariate analysis. New York: Holt, Rinehart and Winston, 1974. 423 p.

FISHER, R. A.; YATES, F. *Statistical tables for biological, agricultural and medical research*. Edinburgh: Oliver and Boyd, 1938. 146 p.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, p. 179–188, 1936.

GOWER, J. C. Multivariate analysis and multidimensional geometry. *Statistician*, v. 17, p. 13–28, 1967.

GOWER, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, v. 53, p. 325–338, 1966.

HARTIGAN, J. A. Direct clustering of a data matrix. *Journal of the American Statistical Association*, v. 67, n. 337, p. 123–129, 1972.

HOTELLING, H. The most predictable criterion. *Journal of Educational Psychology*, v. 26, p. 139–142, 1935.

HOTELLING, H. Relations between two sets of variates. *Biometrika*, v. 28, p. 321–377, 1936.

JEFFERS, J. N. R. Principal component analysis in taxonomic research. *Forestry Commission Statistics Section Paper*, n. 83, p. 1–21, 1964.

KAISER, H. F. *The varimax method of factor analysis*. Ph.D. dissertation. University of California, 1956.

KAISER, H. F. Computer program for varimax rotation in factor analysis. *Educational and Psychological Measurement*, v. 19, n. 3, p. 413–420, 1959.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, v. 23, p. 187–240, 1958.

KENDALL. *A course in multivariate analysis*. Hafner Publishing Company, 1972. 185 p. Second edition.

KENDALL, M. G. *A course in multivariate analysis*. London: Charles Griffin and Company, 1957. 185 p.

KIRCHNER, F. F. *Discrimination of forest types through densitometric analysis of multilevel imagery in Parana State, Brazil*. Thesis (Ph.D.). Michigan State University, 1980.

MACHADO, S. do A.; FIGUEIREDO FILHO, A. *Dendrometria*. Curitiba: Ed. dos Autores, 2003. 309 p.

MARRIOTT, F. H. C. *The interpretation of multiple observations*. London: Academic Press, 1974. 117 p.

MORRISON, D. F. *Multivariate statistical methods*. New York: McGraw-Hill Company, 1976. 415 p.

PEARCE, S. C. *Biological statistics: An introduction*. McGraw-Hill, 1965. 212 p.

PEARCE, S. C.; HOLLAND, D. A. Some applications of multivariate methods in botany. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v. 9, n. 1, p. 1–7, 1960.

SCHOWENGERDT, R. A. *Remote sensing, models and methods for image processing*. 2. ed. Cambridge: Academic Press, 1997. 522 p.

SNEATH, P. H.; SOKAL, R. R. *Numerical taxonomy: The principles and practice of numerical classification*. 1. ed. San Francisco: W. H. Freeman, 1973. 573 p.

WILKS, S. S. Certain generalizations in the analysis of variance. *Biometrika*, v. 24, n. 3/4, p. 471–494, 1932.

BIBLIOGRAFIA COMPLEMENTAR

AFIFI, A.; MAY, S.; CLARK, V. A. Practical multivariate analysis. 5. ed. Boca Raton: Chapman & Hall/CRC, 2012. 517 p.

ANDERBERG, M. R. Cluster analysis for applications. New York: Academic Press, 1973. 359 p.

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A.; FREEMAN, J.; SHOESMITH, E. Statistics for business and economics. London: Thomson Learning, 2007. 904 p.

BASHAW, W. L. Mathematics for statistics. New York: John Wiley & Sons, 1969. 326 p.

BERTIN, J. A neográfica e o tratamento gráfico da informação. Tradução de Cecília Maria Westphalen. Curitiba: Editora da Universidade Federal do Paraná, 1986. 273 p.

BEZANSON, J.; EDELMAN, A.; KARPINSKI, S.; SHAH, V. B. Julia: a fresh approach to numerical computing. SIAM Review, v. 59, n. 1, p. 65-98, 2017.

CHASTON, I. Mathematics for ecologists. London: Butterworths, 1971. 132 p.

CHAVES NETO, A. Notas de aula da disciplina Análise Multivariada Aplicada à Pesquisa. [S. l.: s. n.], [s. d.].

DANCEY, C. P.; REIDY, J. Estatística sem matemática para psicologia. 5. ed. Porto Alegre: Penso, 2013. 608 p.

DE CAROLI, A. J.; CALLIOLI, C. A.; FEITOSA, M. O. Matrizes e sistemas lineares. Rio de Janeiro: Ao Livro Técnico, 1968. 123 p.

DORF, Richard C. Matrix algebra: a programmed introduction. New York: John Wiley & Sons, 1969. 260 p.

ENSLEIN, K.; RALSTON, A.; WILF, H. S. (eds.). Statistical methods for digital computers. New York: John Wiley & Sons, 1977. 454 p.

FIELD, A. *Discovering statistics using SPSS*. 2. ed. London: SAGE Publications, 2005. 779 p.

FERREIRA, Daniel Furtado. Estatística multivariada. 3. ed. Lavras: Editora UFLA, 2018. 624 p.

GNANADESIKAN, R. Methods for statistical data analysis of multivariate observations. New York: John Wiley & Sons, 1977. 311 p.

GOMES, F. P.; NOGUEIRA, I. R. Análise matemática. 2. ed. São Paulo: USP/ESALQ, 1973. 351 p.

GOTELLI, N. J.; ELLISON, A. M. Princípios de estatística em ecologia. Porto Alegre: Artmed, 2011. 527 p.

HAIR JR., J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. Análise multivariada de dados. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

HANSELMAN, D.; LITTLEFIELD, B. *MATLAB 6: curso completo*. São Paulo: Prentice Hall, 2003. 676 p.

JENNINGS, A. Matrix computation for engineers and scientists. New York: John Wiley & Sons, 1977. 330 p.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. Englewood Cliffs: Prentice-Hall, 1982. 594 p.

KACHIGAN, S. K. Multivariate statistical analysis: a conceptual introduction. 2. ed. New York: Radius Press, 1991. 303 p.

KERLINGER, F. N.; PEDHAZUR, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart and Winston, 1973. 534 p.

LOMAX, R. G. An introduction to statistical concepts for education and the behavioral sciences. 2. ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2001. 520 p.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. *Multivariate analysis*. London: Academic Press, 1979. 521 p.

MARQUES, J. M. Apostila de análise multivariada aplicada à pesquisa. Curitiba: Universidade Federal do Paraná, 2014.

MENDENHALL, W.; SCHEAFFER, R. L. Mathematical statistics with applications. North Scituate, MA: Duxbury Press, 1973. 561 p.

NAMBOODIRI, N. K.; CARTER, L. F.; BLALOCK, H. M. Applied multivariate analysis and experimental designs. New York: McGraw-Hill, 1975. 688 p.

NIE, N. H.; HULL, C. H.; STEINBRENNER, K.; BENT, Dale H. Statistical package for the social sciences. 2. ed. New York: McGraw-Hill, 1975. 675 p.

PALLANT, J. SPSS survival manual: a step by step guide to data analysis using SPSS for Windows. 3. ed. Maidenhead: Open University Press, 2007. 335 p.

QUEIROZ, W. T. de. Análise multivariada em inventário florestal contínuo. 1. ed. Belo Horizonte: Editora Poisson, 2021. 272 p.

RENCHER, A. C. *Methods of multivariate analysis*. 2. ed. New York: John Wiley & Sons, 2002. 738 p.

REYMENT, R. A.; BLACKITH, R. E.; CAMPBELL, N. A. Multivariate morphometrics. 2. ed. London: Academic Press, 1984. 233 p.

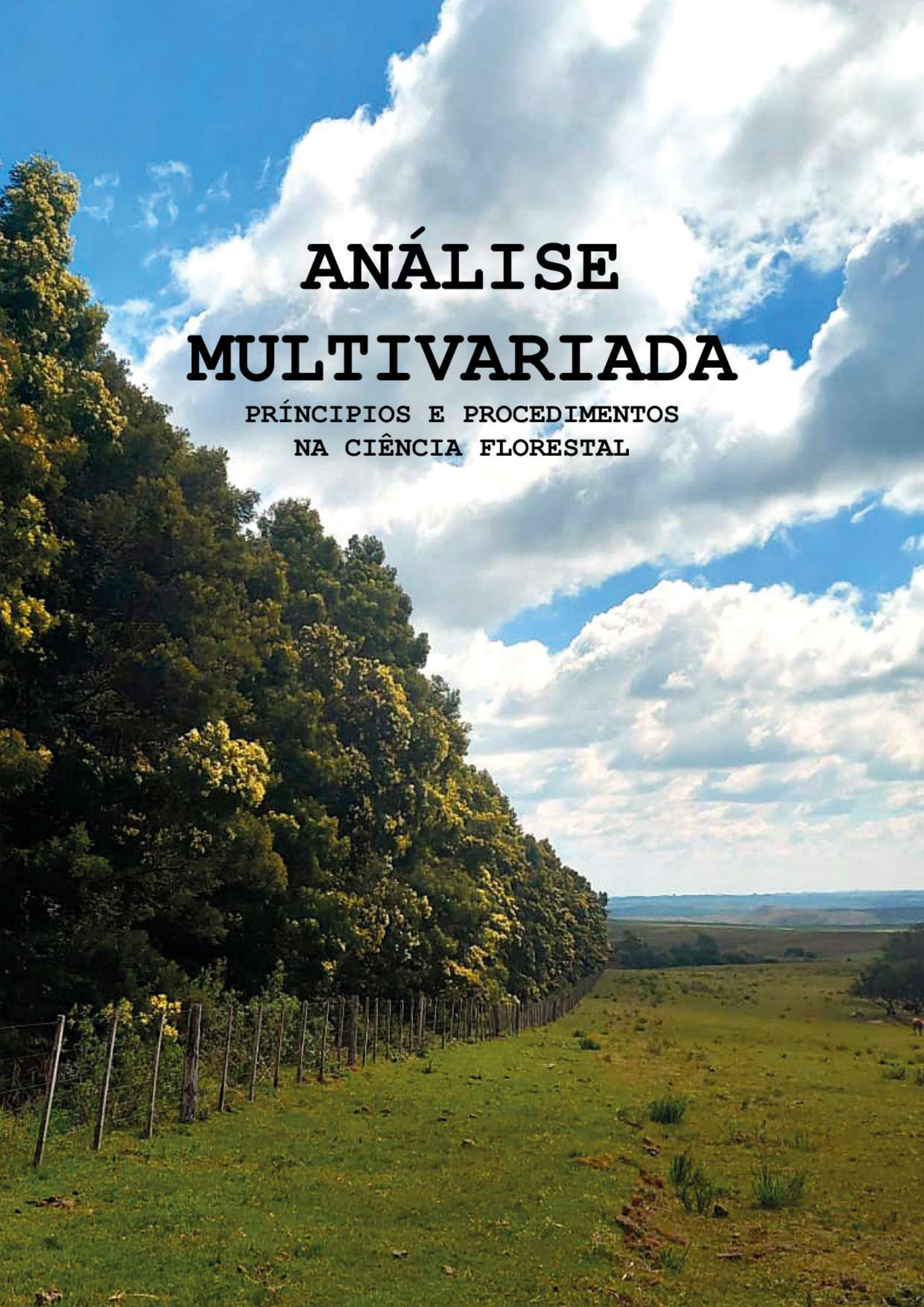
SNEATH, P. H. A.; SOKAL, R. R. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: W. H. Freeman, 1973. 573 p.

SPICER, J. Making sense of multivariate data analysis. Thousand Oaks: Sage Publications, 2005. 231 p.

SPSS INC. SPSS X user's guide. New York: McGraw-Hill, 1983. 806 p.

VELDMAN, D. J. Fortran programming for the behavioral sciences. New York: Holt, Rinehart and Winston, 1967. 406 p.

YAMANE, Taro. Mathematics for economists: an elementary survey. 2. ed. Englewood Cliffs: Prentice-Hall, 1968. 714 p.



ANÁLISE MULTIVARIADA

PRÍNCIPIOS E PROCEDIMENTOS NA CIÊNCIA FLORESTAL