Técnicas Multivariadas em Saúde

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

1. Introdução
2. Distribuições de Probabilidade Multivariadas
3. Representação de Dados Multivariados
4. Testes de Significância c/ Dados Multivariados
5. Análise de Componentes Principais
6. Análise Fatorial
7. Análise de Correlação Canônica
8. Análise de Conglomerados
9. Análise Discriminante
10. Análise de Correspondência
11. Referências

Representação de Dados Multivariados

Análise Exploratória de Dados

O que é Análise Exploratória de Dados?

- Uma filosofia/abordagem para análise de dados
- Emprega uma variedade de técnicas (a maioria gráficas)…trabalharemos com alguns deles:
  - √ Diagrama de dispersão
  - √ Boxplot
  - √ Gráficos para identificação de outliers
  - √ Curvas de crescimento
  - √ Etc.

- São técnicas que buscam:
  - √ maximizar o "insight" do conjunto de dados;
  - √ perceber a estrutura subjacente;
  - √ extrair variáveis importantes;
  - √ detectar valores atípicos (extremos) e anomalias;
  - √ testar hipóteses fundamentais;
  - √ desenvolver modelos parcimoniosos; e
  - √ determinar conjunto ótimo de fatores

Idéia Básica

- Modelo = Suave + Irregular (tosco)
  - √ Frequentemente, as técnicas gráficas podem separar o "suave" do "irregular" ("ruído")

Clássica vs Exploratória

- Seqüência Clássica:
  - √ Problema > Dados > Modelo> Análise > Conclusões
- Exploratória:
  - √ Problema > Dados > Análise > Modelo > Conclusões

Análise Descritiva

- Inicia-se pela verificação dos tipos disponíveis de variáveis
  - √ Elas podem ser resumidas por:
    - Gráficos
    - Medidas
    - Tabelas

Classificação

- Qualitativas (Categóricas)
  - √ Nominais:
  - √ Ordinais
- Quantitativas:
  - √ Discretas
  - √ Contínuas

Objetivos

- Familiarização com os dados
- Detecção de estruturas interessantes
- Presença de valores atípicos (*outliers*)

Razões para Uso de AED

- √ Identificação de erros e inconsistências
- √ Verificação de pressupostos do modelo
- √ Seleção preliminar de modelos apropriados
- √ Determinação das relações entre as variáveis explicativas
- √ Avaliação da direção e da dimensão das relações entre as variáveis explicativas e as variáveis respostas.

Técnicas Multivariadas em Saúde - 2015

Análise Multivariada

- Para um conjunto de variáveis correlacionadas:
  - √ Avaliar as relações entre as variáveis
  - √ Considerar os efeitos dos "tratamentos" sobre essas relações
  - √ Considerar como uma "resposta" depende dessas relações

- Métodos multivariados para redução de dados:
  - √ Resumir as correlações entre variáveis
  - √ Produzir um conjunto menor de variáveis (não correlacionadas) contendo as informações mais importantes
- Para um conjunto de objetos "relacionados"
  - √ Identificar grupos de objetos semelhantes
  - √ Identificar diferenças entre grupos de objetos semelhantes
    - (e o que faz com que os objetos sejam semelhantes)

Importante

- A Análise Exploratória de Dados é um passo inicial crítico em qualquer análise de dados.

Conjuntos de Dados

Conjunto de Dados #7

- Anderson (1935) e Fischer (1936)
- Conjunto de dados de flores de íris (gênero de iridácea)
  - √ Medidas morfológicas de 50 flores de cada espécie
  - √ Espécies:
    - Iris setosa (originária do Alasca)
    - Iris versicolor
    - Iris virginica
- Dados: *iris* {*datasets*}

- √ Variáveis:
  - Sepal.Length: comprimento da sépala, em cm.
  - Sepal.Width: largura da sépala, em cm.
  - Petal.Length: comprimento da pétala, em cm.
  - Petal.Width: largura da pétala, em cm.
  - Species: setosa, versicolor e virginica

Visualização de Dados Multivariados

Box-plot

- Pode ser utilizado para comparações entre diferentes grupos de dados
  - √ Variável quantitativa *vs.* variável categórica

Exemplo – Doenças Cardiovasculares

- Universo:
  - √ Homens doentes com idade entre 45 e 67 anos
- Amostra:
  - √ 100 casos coletados em 1969
- Variáveis
  - √ nível de glicose no sangue, em mg percentuais
  - √ atividade física em casa
    - 1 = sedentário; 2 = moderada

- Perguntas a partir do gráfico:
  - √ Há diferença no nível médio de glicose entre os grupos
  - √ Há diferença das variabilidades dos grupos?
  - √ Há outliers?

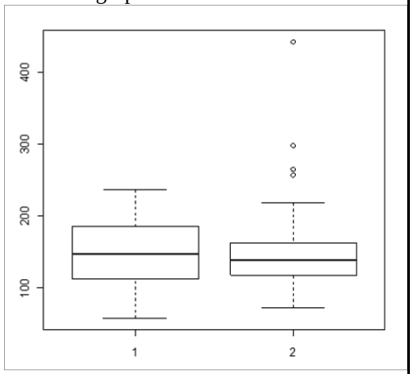
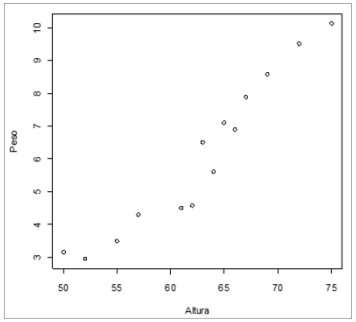


Diagrama de Dispersão

- Análise bivariada de componentes quantitativos
  - √ Pares ordenados (ponto) por item
  - √ Visualização mais simples
- O coeficiente de correlação deve sempre acompanhar o gráfico
- Altura e peso crianças
  - √ Até 1 ano

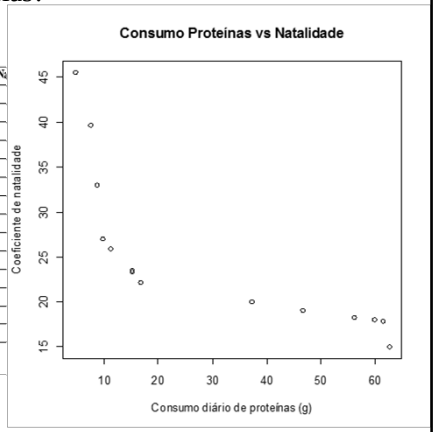


- Objetivo:
  - √ Identificação de tendências (lineares ou não)
  - √ Agrupamentos de itens
    - Há alguma variável categórica que explica?
  - √ Mudanças de variabilidade de uma variável em relação à outra
  - √ Identificação de valores atípicos (‘outliers’)

- Perguntas importantes para análise gráfica:
  - √ Qual a relação entre o peso e a estatura das pessoas?
  - √ Percebem-se ‘clusters’ no conjunto de dados?
  - √ Há diferenças na variabilidade de uma variável, considerados os valores da outra?
  - √ Há valores atípicos?

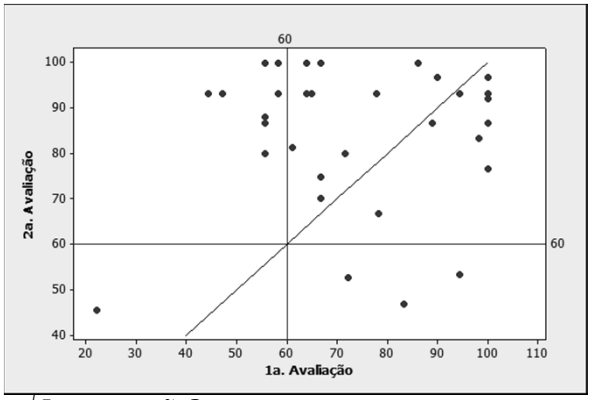
- Exemplo: Consumo de proteínas e natalidade
  - √ Há relação entre elas?

País	Consumo de Proteínas	Iciente de Na
Formosa	4,7	45,6
Malásia	7,5	39,7
Índia	8,7	33,0
Japão	9,7	27,0
Jugoslávia	11,2	25,9
Grécia	15,2	23,5
Itália	15,2	23,4
Bulgária	16,8	22,2
Alemanha	37,3	20,0
Irlanda	46,7	19,1
Dinamarca	56,1	18,3
Austrália	59,9	18,0
Estados Unidos	61,4	17,9
Suécia	62,6	15,0

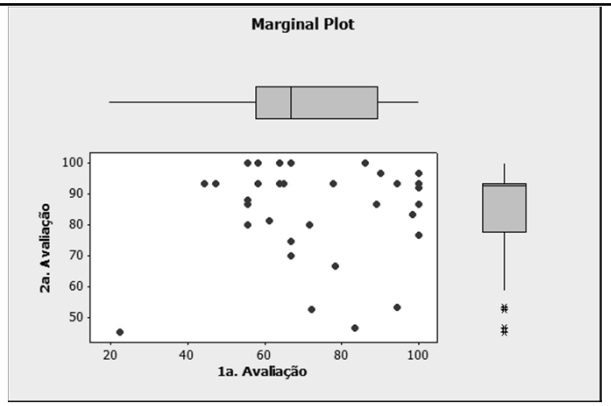


- Perguntas importantes:
  - √ Pode-se afirmar que há relação causal entre consumo de proteínas e natalidade ?
  - √ Há indícios de clusters?

Exemplo: Notas alunos



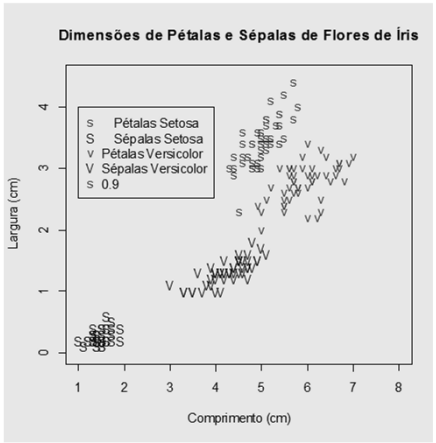
√ Interpretação?



√ Leituras gráficas?

Técnicas Multivariadas em Saúde - 2015

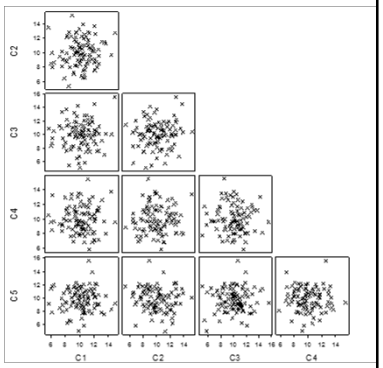
- Exemplo – Flores de íris:
  - √ Setosa é do Alasca
  - √ Há clusters?
  - √ Há outliers?



Técnicas Multivariadas em Saúde - 2015

Scatter Plot Matrix

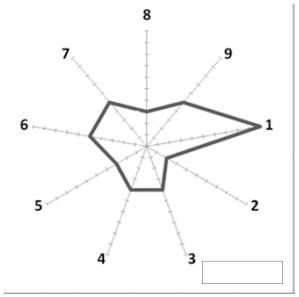
- Dados quantitativos
  - √ Ex.:
    - – Medidas diferentes de um conjunto de plantas
- Diagrama de dispersão para os pares de variáveis
  - √ Apresentação em forma matricial
- Calcular coeficiente de correlação de cada par de variáveis



48

Star Plot

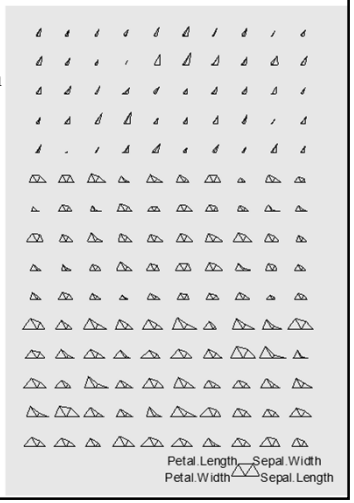
- Estrelas para visualização de dados
- Formação da estrela:
  - √ Raio para cada variável
  - √ Comprimento é proporcional à variável
- Útil para visualização de itens com número arbitrário de variáveis



Técnicas Multivariadas em Saúde - 2015

- Pode ser usado para responder as seguintes perguntas:
  - √ Quais variáveis são dominantes para uma determinada observação?
  - √ Quais observações são similares?
    (Existem agrupamentos de observações?)
  - √ Existem valores discrepantes?

- Exemplo: Flores de íris
  - √ Dados em sequência
    - Setosa, versicolar e virginica
  - √ Valores iniciais pequenos
    - Setosa é do Alasca!
  - √ Há outliers?



Exemplos de Aplicação

Exemplo

- Estudo poluição do ar
  - √ Amostra: 41 cidades americanas
  - √ Variáveis:
    - SO2: concentração no ar (mg/m3)
    - Temp: temperatura
    - Popul: população, em milhares (censo 1970)
    - Vento: velocidade média anual (milhas/hora)
    - Precip: precipitação média anual (pol)
    - Dias: número médio anual de dias de chuva
  - √ Dados: *USairpollution* {*HSAUR2*}

- Carregamento e preparação do conjunto de dados:

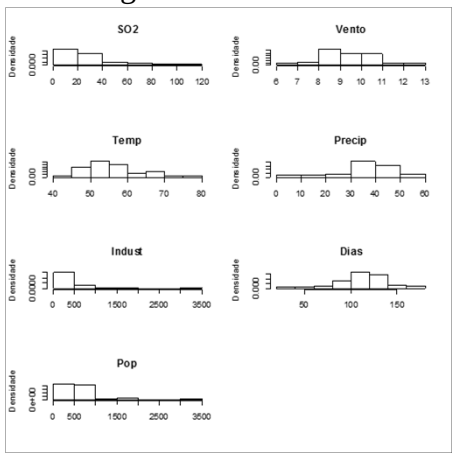
```
> library(MVA, HSAUR2)# carrega os pacotes
> data(USairpollution)# carrega o banco de dados
> help(USairpollution)# Descrição do banco de dados

> colunas<- c("SO2", "Temp", "Indust", "Pop",
+ "Vento", "Precip", "Dias")
> colnames(USairpollution)<- colunas
> dados <- USairpollution
```


- Medidas resumo - univariada

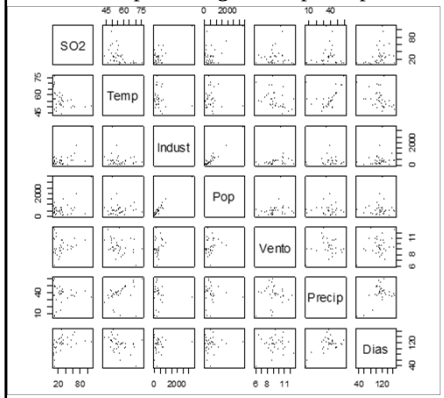
	SO2	Temp	Indust	Pop	Vento	Precip	Dias
Min.	8,0	43,5	35,0	71,0	6,0	7,1	36,0
1º Quartil	13,0	50,6	181,0	299,0	8,7	31,0	103,0
Mediana	26,0	54,6	347,0	515,0	9,3	38,7	115,0
Mean	30,1	55,8	463,1	608,6	9,4	36,8	113,9
3º Quartil	35,0	59,3	462,0	717,0	10,6	43,1	128,0
Max.	110,0	75,5	3344,0	3369,0	12,7	59,8	166,0
D. padrão	23,5	7,2	563,5	579,1	1,4	11,8	26,5
C. variação	78,1%	13,0%	121,7%	95,2%	15,1%	32,0%	23,3%

- Histogramas das variáveis



Verificação quanto a:
- √ Simetria
- √ Modalidade
- √ Pontos atípicos

- Estrutura de correlação entre as variáveis
  - √ Scatterplot matrix
    - Separar os gráficos para apresentá-los mais adequadamente



- Matriz de correlações

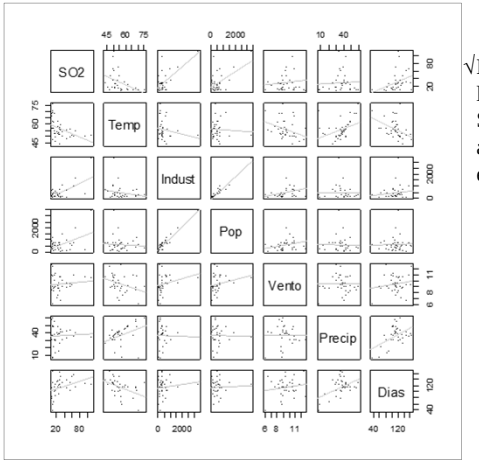
```
> round(cor(dados), 4)
             SO2    Temp  Indust     Pop   Vento  Precip    Dias
SO2       1.0000 -0.4336  0.6448  0.4938  0.0947  0.0543  0.3696
Temp     -0.4336  1.0000 -0.1900 -0.0627 -0.3497  0.3863 -0.4302
Indust    0.6448 -0.1900  1.0000  0.9553  0.2379 -0.0324  0.1318
Pop       0.4938 -0.0627  0.9553  1.0000  0.2126 -0.0261  0.0421
Vento     0.0947 -0.3497  0.2379  0.2126  1.0000 -0.0130  0.1641
Precip    0.0543  0.3863 -0.0324 -0.0261 -0.0130  1.0000  0.4961
Dias      0.3696 -0.4302  0.1318  0.0421  0.1641  0.4961  1.0000
```

√ Forte correlação entre SO2 e Industr e Popul
- Indust e Popul são fortemente correlacionadas
- Provavelmente predizem SO2 da mesma maneira

√ Correlação entre SO2 e Precip é muito pequena

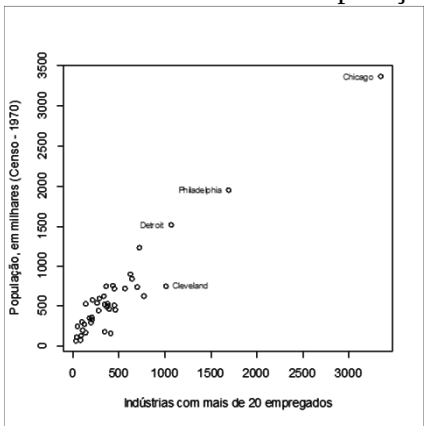
√ Correlação entre SO2 e Dias é moderada

- Scatterplot matrix com ajuste linear



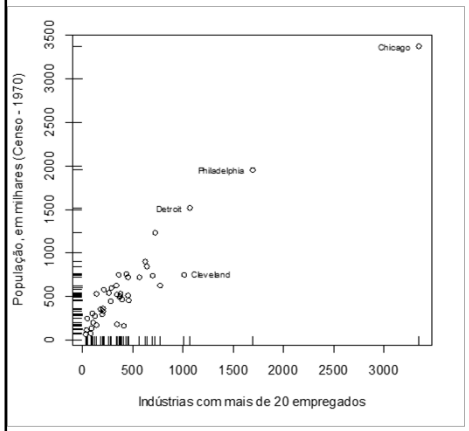
√É provável que modelo linear entre SO2 e Precip e SO2 e dias não perceberá adequadamente a relação entre cada par de variáveis

- Variáveis Indústria e População



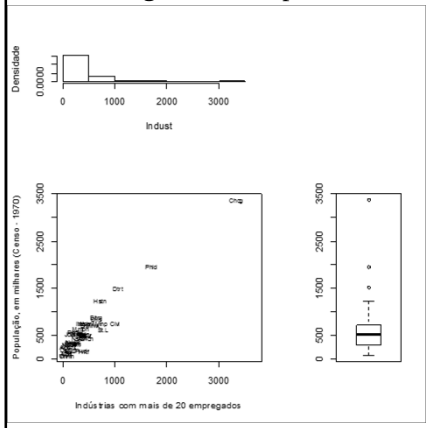
√Há pontos que se afastam do padrão dos dados

- Gráfico de dispersão com distribuição marginal



√Concentração de pontos na faixa inferior de ambas as variáveis

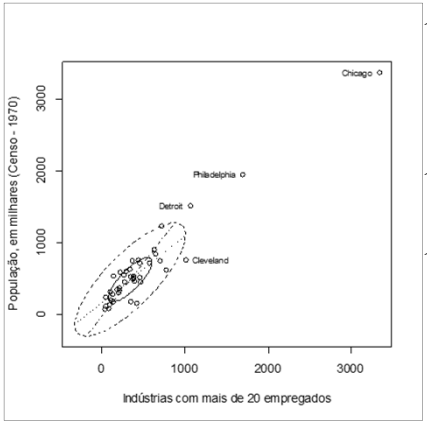
- Gráfico de dispersão com distribuição marginal
  - √ Histograma e boxplot



√Boxplot identifica alguns pontos extremos (univariado)

Técnicas Multivariadas em Saúde - 2015

- Boxplot bivariado
  - √ Análogo ao boxplot univariado

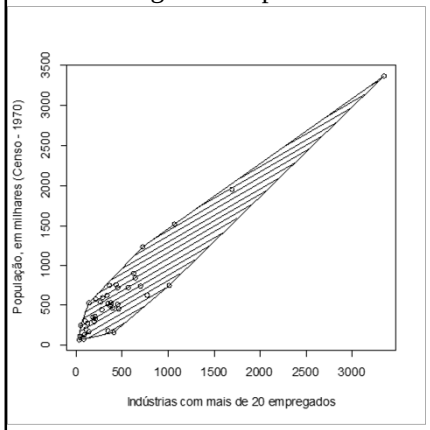


√Boxplot bivariado identifica pontos extremos do vetor de dados (bivariado)

√Correlação
Todas: 0,9553
Exceto identificadas: 0,7956
√A redução não é considerável

Técnicas Multivariadas em Saúde - 2015

- Envelope convexo dos dados
  - √ Análogo ao boxplot univariado

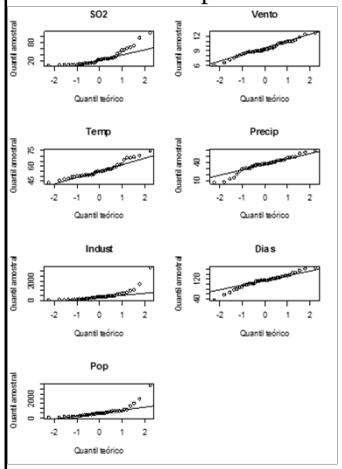


√Envelope convexo apara valores
√Estimação robusta da correlação

√Correlação
Todas: 0,9553
Exceto envoltória: 0,9225
√Correlação estimada após remoção é maior que a relacionado com pontos identificados pelo boxplot bivariado

Técnicas Multivariadas em Saúde - 2015

- Gráficos de probabilidade univariados



√ Gráficos para SO2 e Precipitação desviam-se consideravelmente da linearidade
√ Gráficos para Indústrias e População evidenciam outliers

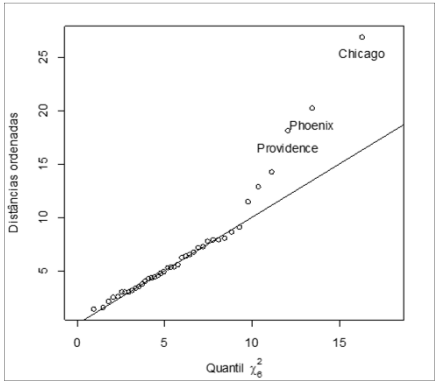
Técnicas Multivariadas em Ecologia - 2014

- Teste de normalidade
  - √ H_0: dados se ajustam à distribuição normal
  - √ H_1: dados não se ajustam à distribuição normal

```
> teste<- sapply(colnames(dados), function(x)shapiro.test(dados[[x]])$p.value)
> as.matrix(teste)
              [,1]
SO2     9.723376e-06
Temp    2.214972e-02
Indust 2.781101e-09
Pop     3.622798e-08
Vento   6.972580e-01
Precip 3.725311e-02
Dias    2.419457e-01
```

  - √ Não se rejeita a hipótese de normalidade para as variáveis Vento e Dias

- Gráfico das distâncias generalizadas (χ^2):



  - √ Gráfico detectou possíveis outliers nos dados multivariados
    - Desvio da variabilidade natural dos dados

Comentários

- É difícil construir um bom teste global de normalidade conjunta em mais de duas dimensões
- Hipótese de normalidade aparenta estar violada
  - √ As marginais aparentam ser normais? E aalgumas combinações lineares de componentes X_i?
  - √ Diagramas de dispersão de diferentes características têm aparência elíptica?
  - √ Há outliers que devessem ser verificados?

- Hipótese de normalidade individual é menos crucial em situações em que o tamanho amostral é grande e as técnicas dependem da média amostral (ou de distâncias envolvendo essa média)

Referências

Bibliografia Recomendada

- MANLY, B. J. F. *Métodos Estatísticos Multivariados: uma Introdução.* Bookman, 2008.
- JOHNSON, R. A.; WINCHERN, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007
- MINGOTI, D.C. *Análise de Dados através de Métodos de Estatística Multivariada*. Ed. UFMG, 2005.
- EVERITT, B.; HOTHORN, T. *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011.