Análise Multivariada

Lupércio França Bessegato
Dep. Estatística/UFJF

Roteiro

Análise de Componentes Principais

Introdução

- Objetivo:
  - √ Explicar a estrutura de variância e covariância de conjunto de variáveis através de algumas combinações lineares das mesmas
  - √ Busca-se:
    - Redução de dados
    - Interpretação

Componentes Principais Exatas

- Algebricamente:
  - √ Combinações lineares particulares das *p* variáveis aleatórias $X_1, X_2, ..., X_p$.
- Geometricamente:
  - √ Representam a seleção de um novo sistema de coordenadas obtidas por rotação do sistema original
  - √ Os novos eixos representam as direções com maior variabilidade
  - √ Fornecem descrição mais simples e mais parcimoniosa da estrutura de covariâncias

- Componentes principais:
  - √ São necessárias *p* componentes para reproduzir a variabilidade total do sistema
  - √ As componentes são não correlacionadas entre si
    - – Ortogonalidade entre as componentes
  - √ Variabilidade das p variáveis é aproximada pela variabilidade das k principais componentes
    - – Buscam-se situações em que haja quase tanta informação nas k componentes principais quanto nas p variáveis originais

- Análise de componentes principais:
  - √ Não pressupõe normalidade
    - – Componentes principais derivadas de populações normais têm interpretações úteis
  - √ Com frequência, revela relações insuspeitadas
    - – Pode permitir interpretações que não seriam obtidas preliminarmente
  - √ Em geral, é um passo intermediário para a aplicação de outras técnicas

Componentes Principais Exatas Extraídas da Matriz de Covariâncias

- Sejam o vetor aleatório

$$\mathbf{X}' = [X_1, X_2, \ldots, X_p].$$

com matriz de covariâncias é $\mathbf{\Sigma}$, cujos autovalores são $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

- Componentes principais de $\mathbf{\Sigma}$:

$$Y_1, Y_2, ..., Y_p.$$

  - √ Combinações lineares não correlacionadas do vetor aleatório, cujas variâncias são as maiores possíveis

- Definição – Componente principal:
  - √ Sistema cuja j-ésima combinação linear de **X** é definida como:

$$Y_j = \mathbf{a}_j'\mathbf{X} = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jp}X_p.$$

  - √ $\mathbf{e}_j$: autovetor correspondente ao j-ésimo autovalor
- Esperança e variância de Y_j:

$$\mathrm{E}[Y_j] = \mathrm{E}[\mathbf{e}_j'\mathbf{X}] = \mathbf{e}_j'\boldsymbol{\mu} = e_{j1}\mu_1 + e_{j2}\mu_2 + \cdots + e_{jp}\mu_p.$$

$$\mathrm{Var}[Y_j] = \mathrm{Var}[\mathbf{a}_j'\mathbf{X}] = \mathbf{a}_j'\boldsymbol{\Sigma}\mathbf{a}_j.$$

- Covariância entre duas componentes principais:

$$\mathrm{Cov}[Y_j, Y_k] = \mathbf{a}_j'\boldsymbol{\Sigma}\mathbf{a}_k,\ j \neq k,\ j = 1, 2, \ldots, p$$

Análise Multivariada - 2016 9

√ Buscam-se os valores dos coeficientes a_{ij}, tais que:

i. Y_1, Y_2, ..., Y_p tenham variância máxima e sejam não correlacionadas entre si
ii. Os vetores $\mathbf{a}_i$ tenham comprimento unitário:

$$\mathbf{a}_j'\mathbf{a}_k = \begin{cases} 1 & ,\ \text{se } j = k \\ 0 & ,\ \text{se } j \neq k \end{cases}$$

√ Pode-se demostrar que :
- A variância máxima de ($\mathbf{a}_i'$ **X**) é igual a λ_i.
- É obtida quando $\mathbf{a}_i = \mathbf{e}_i$.

Análise Multivariada - 2016 10

- Definição 2 – Componente principal:
  - √ A j-ésima componente principal da matriz **Σ** é definida como:

$$Y_j = \mathbf{e}_j'\mathbf{X} = e_{j1}X_1 + e_{j2}X_2 + \cdots + e_{jp}X_p.$$

  - √ $\mathbf{e}_j$: autovetor correspondente ao j-ésimo autovalor
- Esperança e variância de Y_j:

$$\mathrm{E}[Y_j] = \mathrm{E}[\mathbf{e}_j'\mathbf{X}] = \mathbf{e}_j'\boldsymbol{\mu} = e_{j1}\mu_1 + e_{j2}\mu_2 + \cdots + e_{jp}\mu_p.$$

$$\mathrm{Var}[Y_j] = \mathrm{Var}[\mathbf{e}_j'\mathbf{X}] = \mathbf{e}_j'\boldsymbol{\Sigma}\mathbf{e}_j = \mathbf{e}_j'\left(\sum_{i=1}^{p}\lambda_i\mathbf{e}_i\mathbf{e}_i'\right)\mathbf{e}_j = \lambda_j.$$

- Covariância entre duas componentes principais:

$$\mathrm{Cov}[Y_j, Y_k] = 0,\ j \neq k$$

Análise Multivariada - 2016 13

- Comentário:
  - √ Cada autovalor λ_j representa a variância de uma componente principal Y_j.
  - √ Autovalores estão ordenados em ordem decrescente
    - A primeira componente é a de maior variabilidade
    - A p-ésima componente é a de menor variabilidade

Análise Multivariada - 2016 14

- Variâncias total e generalizada de $\mathbf{\Sigma}$:
  - √ Total: $\mathrm{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i$
  - √ Generalizada de $\mathbf{\Sigma}$: $|\mathbf{\Sigma}| = \prod_{i=1}^{p} \lambda_i$
  - √ Em termos dessas duas medidas globais de variação, os vetores **X** e **Y** são equivalentes

- Proporção da variância total que é explicada pela j-ésima componente principal:

$$\frac{\mathrm{Var}[Y_j]}{\text{Variância total de } \mathbf{X}} = \frac{\lambda_j}{\mathrm{tr}(\mathbf{\Sigma})} = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

  - √ 1ª componente tem a maior proporção de explicação
- Proporção da variância total que é explicada pelas k primeiras componentes principais

$$\frac{\sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{Var}[Y_j]}{\text{Variância total de } \mathbf{X}} = \frac{\sum_{j=1}^{k} \lambda_j}{\mathrm{tr}(\mathbf{\Sigma})} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

  - √ Busca-se analisar um conjunto menor de variáveis sem perder muita informação sobre a estrutura de variabilidade original

- Aproximação de $\mathbf{\Sigma}$:
  - √ Analisando as k primeiras componentes principais

$$\mathbf{\Sigma}_{p\times p} \approx \sum_{i=1}^{k} \lambda_i \mathbf{e}_j \mathbf{e}_j'$$

  - √ Cada parcela da soma envolve uma matriz de dimensão pxp correspondente apenas à informação da j-ésima componente principal

Correlação entre Componente Principal e Variável Aleatória

- Os coeficientes de correlação entre a componente principal Y_i de S e a variáve1 X_k é

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

  - √ A magnitude de e_{ik} mede a contribuição da k-ésima variável na i-ésima componente (a despeito das outras variáveis).
    - – Não medem a importância de X_k na presença das outras variáveis.
    - – Alguns estatísticos recomendam que somente os valores e_{ik} (e não as correlações) sejam consideradas na interpretação dos componentes

Estimação das Componentes Principais – Matriz de Covariâncias

- Em geral, $\mathbf{\Sigma}$ é estimada por **S**:

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1p} \\ S_{12} & S_{22} & \dots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{1p} & S_{2p} & \dots & S_{pp} \end{bmatrix}.$$

√ Autovalores de **S**: $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p$

√ Autovetores de **S**: $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_p$

- j-ésima componente principal de **S**:

$$\hat{Y}_j = \hat{\mathbf{e}}_j'\mathbf{X} = \hat{e}_{j1}X_1 + \hat{e}_{j2}X_2 + \cdots + \hat{e}_{jp}X_p, \; j = 1, 2, \dots, p.$$

- Componentes principais amostrais – Propriedades
  i. Variância: $\mathrm{Var}[\hat{Y}_j] = \hat{\lambda}_j.$
  ii. Covariância entre as componentes: $\mathrm{Cov}(\hat{Y}_j, \hat{Y}_k) = 0, \; j \neq k$
  iii. Variância total estimada explicada pela componente:

$$\frac{\mathrm{Var}[\hat{Y}_j]}{\text{Variância total estimada de } \mathbf{X}} = \frac{\hat{\lambda}_j}{\mathrm{tr}(\mathbf{S})} = \frac{\hat{\lambda}_j}{\sum_{i=1}^{p} \hat{\lambda}_i}$$

  iv. Correlação estimada entre componente e variável:

$$r_{\hat{Y}_j, X_k} = \frac{\hat{e}_{jk}\sqrt{\hat{\lambda}_j}}{\sqrt{S_{kk}}}$$

- Decomposição espectral de **S**:

$$\mathbf{S} = \sum_{j=1}^{p} \hat{\lambda}_j \mathbf{e}_j \mathbf{e}_j'.$$

√ Aproximação de S pelas primeiras k componentes

$$\mathbf{S}_{p \times p} \approx \sum_{i=1}^{k} \hat{\lambda}_i \hat{\mathbf{e}}_j \hat{\mathbf{e}}_j'$$

- Scores das componentes

√ Valor das componentes para cada elemento amostral

√ Na prática, o uso das componentes relevantes se dá através dos scores

Exemplo 8.3

- Pesquisa com 5 variáveis socioeconômicas

√ X_1: População total (milhares)

√ X_2: Escolaridade mediana (anos concluídos)

√ X_3: Emprego total (milhares)

√ X_4: Empregos na área da saúde (centenas)

√ X_5: Valor mediano da habitação (x \$10.000)

- Dados: *BD_multivariada.xls/pesquisa*

- Vetor de médias amostral ($\bar{\mathbf{x}}$)

Variable	Mean
X1_Pop	4,323
X2_escol	14,014
X3_empregos	1,952
X4_saude	2,171
X5_habitacao	2,454

- Matriz de covariâncias amostral (S)

Covariances: X1_Pop; X2_escol; X3_empregos; X4_saude; X5_habitacao

	X1_Pop	X2_escol	X3_empregos	X4_saude	X5_habitacao
X1_Pop	4,307556				
X2_escol	1,683680	1,767473			
X3_empregos	1,802776	0,588026	0,800669		
X4_saude	2,155326	0,177978	1,064828	1,969475	
X5_habitacao	-0,253474	0,175549	-0,158339	-0,356807	0,504380

- A variação amostral pode ser resumida por uma ou duas componentes principais?

29

	Componentes Principais						
	1		2		3	4	5
	e1	r(y1,xk)	e2	r(y2,xk)	e3	e4	e5
População Total	0,781	0,99	0,071	-0,04	-0,004	-0,542	0,302
Escolaridade Mediana	0,306	0,61	0,764	-0,76	0,162	0,545	0,009
Total de Empregos	0,334	0,98	-0,083	0,12	-0,015	-0,051	-0,937
Empregos Área Saúde	0,426	0,80	-0,579	0,55	-0,220	0,636	0,172
Valor Mediano Habitação	-0,054	-0,20	0,262	0,49	-0,962	-0,051	-0,025
Variância	6,931		1,785		0,390	0,230	0,014
% Variância Total (acumulada)	74,1		93,2		97,4	99,8	100,0

- Variância amostral é bem resumida por 2 componentes
  - √ redução de 14 observações de 5 variáveis para 14 observações de 2 variáveis
  - √ 1ª. componente: média ponderada de 4 variáveis
  - √ 2ª. componente: contraste entre empregos saúde com média ponderada da escolaridade com valor habitação

30

- Correlação mede unicamente importância de uma variável individual sem considerar a influência das demais
  - √ No exemplo, os coeficientes de correlação confirmam a interpretação fornecida pelos coeficientes das componentes

31

Número de Componentes Principais

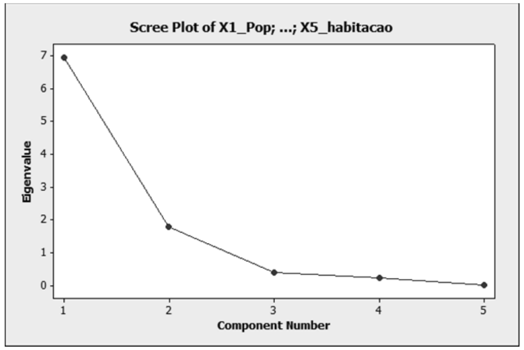
- Quantas componentes principais devem ser retidas?
  - √ Não há resposta definitiva
- Considerações a serem tomadas:
  - √ Quantidade explicada de variância amostral total
  - √ Tamanho relativo dos autovalores
    (variância das componentes amostrais)
  - √ Interpretação das componentes

32

Scree Plot

- Gráfico λ_i *vs. i*
  - √ Procura-se um ‘cotovelo’ no gráfico
  - √ São consideradas as componentes até o ponto em que os autovalores remanescentes são relativamente pequenos e todos aproximadamente do mesmo valor

- Exemplo 8.3



Exemplo 8.4

- Relação entre tamanho e forma de cascos de tartaruga
  - √ Comprimento
  - √ Largura
  - √ Espessura
  - √ Gênero: macho/fêmea
- Análise para as tartarugas macho
- Literatura sugere transformação logarítmica em estudos de relação entre tamanho e forma
- Dados: *BD_multivariada.xls/tartarugas*

- Vetor de médias amostral ($\bar{\mathbf{x}}$)

Descriptive Statistics: log_comp_male; log_larg_male; log_esp_male

Variable	Mean
log_comp_male	4,7254
log_larg_male	4,4776
log_esp_male	3,7032

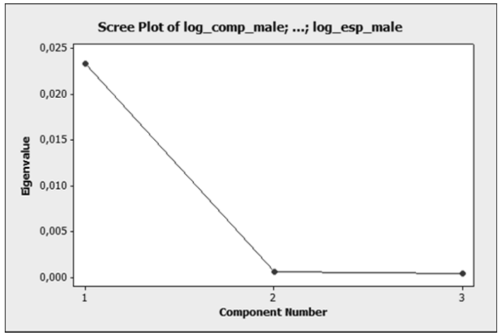
- Matriz de covariâncias amostral (S)

Covariances: log_comp_male; log_larg_male; log_esp_male

	log_comp_male	log_larg_male	log_esp_male
log_comp_male	0,01107200		
log_larg_male	0,00801914	0,00641673	
log_esp_male	0,00815965	0,00600527	0,00677276

- A variação amostral pode ser resumida por uma componente principal?

- Scree Plot



√ Uma componente principal é claramente dominante

38

- Componentes principais:

Principal Component Analysis: log_comp_male; log_larg_male; log_esp_male

Eigenanalysis of the Covariance Matrix

Eigenvalue	0,023303	0,000598	0,000360
Proportion	0,961	0,025	0,015
Cumulative	0,961	0,985	1,000

Variable	PC1	PC2	PC3
log_comp_male	0,683	-0,159	-0,713
log_larg_male	0,510	-0,594	0,622
log_esp_male	0,523	0,788	0,324

- Componente adotada:

$$\begin{aligned}\hat{y}_1 &= 0,683\ln(comp) + 0,510\ln(larg) + 0,523\ln(espes)\\ &= \ln\left[(comp)^{0,683}(larg)^{0,510}(esp)^{0,523}\right]\end{aligned}$$

√ ln(volume) de uma caixa com dimensões ajustadas

39

Componentes Principais de Variáveis Padronizadas

- Padronização do vetor aleatório **X**:

$$\mathbf{Z} = \left(\mathbf{V}^{1/2}\right)^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

√ **V**½: matriz diagonal de desvios-padrão

√ Variável padronizada: $Z_i = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$

√ Matriz de covariâncias de **Z**:

$$\text{Cov}(\mathbf{Z}) = \left(\mathbf{V}^{1/2}\right)^{-1}\boldsymbol{\Sigma}\left(\mathbf{V}^{1/2}\right)^{-1} = \mathbf{P}$$

√ Componentes principais de **Z**:

– Obtidas dos autovalores e autovetores de **P**.

44

- Componente principal das variáveis padronizadas:

√ A j-ésima componente principal da matriz **Σ** :

$$Y_j = \mathbf{e}_j'\mathbf{Z} = \mathbf{e}_j'\left(\mathbf{V}^{1/2}\right)^{-1}(\mathbf{X} - \boldsymbol{\mu}) = e_{j1}Z_1 + e_{j2}Z_2 + \cdots + e_{jp}Z_p.$$

√ $\mathbf{e}_j$: autovetor da matriz de correlações **P**.

- Variância total de **P**:

$$\sum_{j=1}^{p}\text{Var}[Y_j] = \sum_{j=1}^{p}\text{Var}[Z_j] = p$$

√ Proporção de variância populacional (padronizada) devido à j-ésima componente

$$\frac{\text{Var}[Y_j]}{\text{Variância total de }\mathbf{Z}} = \frac{\lambda_j}{\text{tr}(\mathbf{P})} = \frac{\lambda_j}{p},\ k = 1, 2, \ldots, p$$

√ Correlação entre Y_j e X_k: $\rho_{Y_j,X_k} = e_{jk}\sqrt{\lambda_j},\ i,k = 1, 2, \ldots, p$

45

Comentários

- As componentes principais de $\mathbf{\Sigma}$ são diferentes daquelas obtidas de $\mathbf{P}$.
  - √ Seus autovalores e autovetores são diferentes
  - √ Um conjunto de componentes principais não é simplesmente uma função do outro conjunto
- A padronização traz consequências
  - √ Variáveis deveriam ser padronizadas se elas são medidas em escalas com amplitudes muito diferentes
    - Ex. Vendas anuais e razão entre lucro/ativos

Padronização dos Componentes Principais Amostrais

- Frequentemente são padronizadas:
  - √ Variáveis medidas em diferentes escalas
  - √ Na mesma escala, mas com amplitudes bastante diferentes
- As componentes principais não são invariantes às mudanças na escala

- Padronização dos elementos amostrais:

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) = \begin{bmatrix} \frac{x_{j1}-\bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{j2}-\bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{jp}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix}, \ j = 1, 2, \ldots, n$$

  - √ D: matriz diagonal dos desvios-padrão amostrais
- Matriz de dados:

$$\mathbf{Z}_{n\times p} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix}.$$

Análise de Componentes Principais – Matriz de Correlações

- As componentes principais obtidas a partir da matriz de covariâncias são influenciadas pelas variáveis de maior variância
  - √ A padronização das variáveis ameniza esse problema
- Análise de componentes principais de variáveis padronizadas é equivalente a obter as componentes principais através da matriz de correlações

Estimação das Componentes Principais – Matriz de Correlação

- **P** é estimada por **R**:
  - √ Importante: $\mathbf{S_Z}$ = R

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{12} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & 1 \end{bmatrix}.$$

  - √ Autovalores de **R**: $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p$
  - √ Autovetores de **R**: $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_p$

- j-ésima componente principal de **R**:

$$\hat{Y}_j = \hat{\mathbf{e}}_j'\mathbf{Z} = \hat{e}_{j1}Z_1 + \hat{e}_{j2}Z_2 + \cdots + \hat{e}_{jp}Z_p, \ \ j = 1, 2, \dots, p.$$

- Componentes principais amostrais – Propriedades
  - i. Variância: $\mathrm{Var}[\hat{Y}_j] = \hat{\lambda}_j.$
  - ii. Covariância entre as componentes: $\mathrm{Cov}(\hat{Y}_j, \hat{Y}_k) = 0, \ j \neq k$
  - iii. Variância total estimada explicada pela componente:

$$\frac{\hat{\lambda}_j}{p}.$$

  - iv. Correlação estimada entre componente e variável:

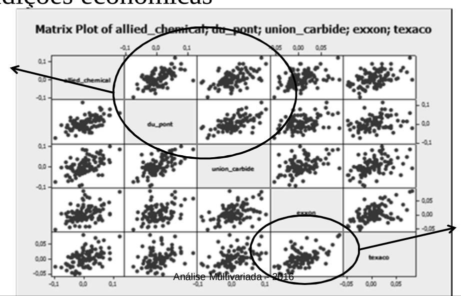
$$r_{\hat{Y}_j, X_k} = \hat{e}_{jk}\sqrt{\hat{\lambda}_j}$$

Exemplo 8.5

- Taxas de retorno de 5 ações negociadas na Bolsa de New York
  - √ Período: Jan./75 a Dez./76
  - √ Ações:
    - Allied Chemical
    - du Pont
    - Union Carbide
    - Exxon
    - Texaco
  - √ Dados: BD_multivariada.xls/

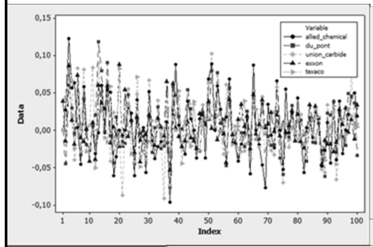
- Taxa de retorno: $\begin{pmatrix}\text{taxa de}\\ \text{retorno}\end{pmatrix} = \dfrac{\begin{pmatrix}\text{\$ fechamento}\\ \text{sexta atual}\end{pmatrix} - \begin{pmatrix}\text{\$ fechamento}\\ \text{sexta anterior}\end{pmatrix}}{\begin{pmatrix}\text{\$ fechamento}\\ \text{sexta anterior}\end{pmatrix}}$
- As taxas de retorno entre ativos estão correlacionadas
  - √ ações tendem a se mover juntas em resposta às condições econômicas

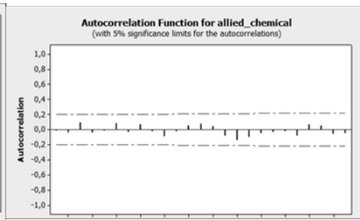
indústria química



indústria óleo & gás

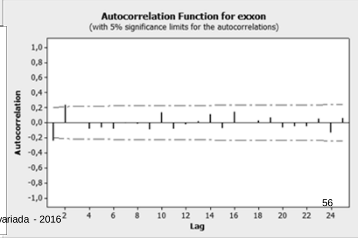
- Observações de 100 semanas aparentam estar distribuídas independentemente





Autocorrelation Function: exxon

Lag	ACF	T	LBQ
1	-0,239413	-2,39	5,91
2	0,233674	2,21	11,59
3	-0,000760	-0,01	11,59
4	-0,086924	-0,79	12,39
5	-0,068469	-0,62	12,90
6	-0,084782	-0,76	13,67
7	-0,012353	-0,11	13,69
8	-0,017693	-0,16	13,73
9	-0,093155	-0,83	14,70
10	0,134497	1,19	16,75
11	-0,087232	-0,76	17,62
12	-0,022943	-0,20	17,68
13	0,012750	0,11	17,70
14	0,109989	0,95	19,14



LFB1

- Vetor de médias amostral ($\overline{x}$)

Descriptive Statistics: allied_chemical; du_pont; union_carbide; exxon; texaco

Variable	Mean
allied_chemical	0,00543
du_pont	0,00483
union_carbide	0,00565
exxon	0,00629
texaco	0,00371

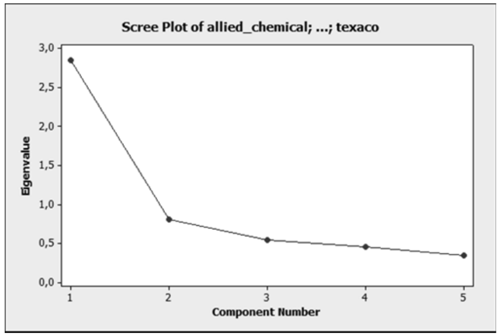
- Matriz de correlação amostral (**R**)

Correlations: allied_chemical; du_pont; union_carbide; exxon; texaco

	allied_chemical	du_pont	union_carbide	exxon
du_pont	0,577			
union_carbide	0,509	0,598		
exxon	0,387	0,390	0,436	
texaco	0,462	0,322	0,426	0,524

- A variação amostral pode ser resumida por uma ou duas componentes principais?

- Scree Plot



√ Aparentemente duas componentes principais resumem bem os dados

- Componentes principais:

Principal Component Analysis: allied_chemi; du_pont; union_carbid; exxon; texac

Eigenanalysis of the Correlation Matrix

Eigenvalue	2,8565	0,8091	0,5400	0,4513	0,3430
Proportion	0,571	0,162	0,108	0,090	0,069
Cumulative	0,571	0,733	0,841	0,931	1,000

Variable	PC1	PC2	PC3	PC4	PC5
allied_chemical	0,464	0,241	0,613	-0,381	-0,453
du_pont	0,457	0,509	-0,178	-0,211	0,675
union_carbide	0,470	0,261	-0,337	0,664	-0,396
exxon	0,422	-0,525	-0,539	-0,473	-0,179
texaco	0,421	-0,582	0,434	0,381	0,387

√ Duas primeiras componentes com 73% da variabilidade amostral padronizada total

Slide 57

LFB1 Calcular matriz de covariâncias amostral
Há domínio de variabilidade?
Lupércio Bessegato; 20/02/2013

- 1ª. componente principal:

$$\hat{y}_1 = 0,464z_1 + 0,457z_2 + 0,470z_3 + 0,421z_4 + 0,421z_5$$

√ Variáveis:
- z_1: retorno padronizado – Allied Chemical
- z_1: retorno padronizado – du Pont
- z_1: retorno padronizado – Union Carbide
- z_1: retorno padronizado – Exxon
- z_1: retorno padronizado – Texaco

√ Interpretação:
- soma ponderada (índice) das 5 ações
- pesos aproximadamente iguais
- Componente geral do mercado de ações
  (componente do mercado)

- 2ª. componente principal:

$$\hat{y}_2 = 0,240z_1 + 0,509z_2 + 0,260z_3 - 0,526z_4 - 0,582z_5$$

√ Interpretação:
- contraste entre ações de indústrias químicas e de óleo & gás
- Componente industrial

- Comentários:

√ A maioria das variações dos ativos devem-se às atividades de mercado (1ª. componente) e atividades industriais não correlacionadas (2ª. componente)

√ As componente remanescentes não são de simples interpretação
- coletivamente , representam variação que é provavelmente específica de cada ação

Variáveis Padronizadas – Regra Empírica

- Reter apenas as componentes cujas variâncias (λ_i) são maiores que a unidade

√ componente que explicam individualmente pelo menos 1/*p* da variância amostral padronizada total

- No caso do exemplo anterior (8.6), pareceu-se sensível reter uma componente (y_2) associada à autovalor menor que a unidade

Importante

√ Um valor pequeno incomum para o último autovalor da matriz de covariâncias (ou correlação) amostral pode indicar uma dependência linear não detectada no conjunto de dados

√ Valores grande de autovalores (e correspondentes autovetores são importantes em uma análise

√ Autovalores próximos de zero não devem ser ignorados

– Autovetores associados podem apontar dependências lineares no conjunto de dados

(problemas computacionais ou de interpretação)

Gráfico dos Componentes Principais

- Podem:
  - √ revelar observações suspeitas
  - √ fornecer verificações da hipótese de normalidade

- São combinações das variáveis originais:
  - √ Se as observações provém de população normal multivariada, é razoável esperar que as componentes sejam aproximadamente normais
  - √ Se forem usadas como entrada em análises adicionais
    - – Verificar se as 1ª.s componentes são aproximadamente normais
- As últimas componentes principais podem ajudar a apontar observações suspeitas

Resumo

- Procedimento auxiliar na verificação de normalidade
  - √ Construir diagrama de dispersão para os pares dos primeiros componentes principais
  - √ Construir Q-Q plots para os valores amostrais gerados por cada componente principal
- Identificação de observações suspeitas:
  - √ Construir diagramas de dispersão e Q-Q plots para as últimas componentes principais.

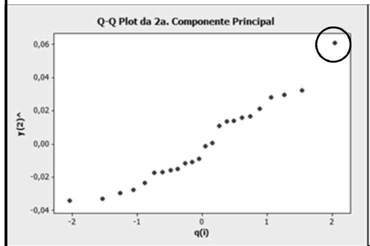
Exemplo 8.7

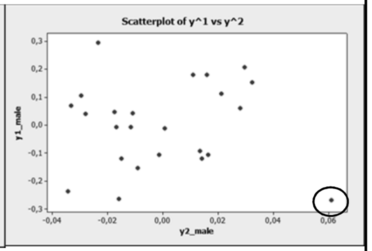
- Plotando os Componentes Principais dos dados das tartarugas macho:
  - √ x_1 = ln(comp)
  - √ x_2 = ln(larg)
  - √ x_3 = ln(esp)
- Componentes:

$$
\begin{aligned}
\hat{y}_1 &= 0,683\ln(x_1-4,725)+0,510\ln(x_2-4,478)+0,523\ln(x_3-3,703)\\
\hat{y}_2 &= -0,159\ln(x_1-4,725)-0,594\ln(x_2-4,478)+0,788\ln(x_3-3,703)\\
\hat{y}_3 &= -0,713\ln(x_1-4,725)+0,622\ln(x_2-4,478)+0,324\ln(x_3-3,703)
\end{aligned}
$$

- Comandos Minitab para Q-Q Plot

```
Name C30 "(j-1/2)/n"
Set C30
1( 1 : 24 / 1 )1
End.

Let C30 = (C30-0,5)/24    # Cálculo percentagens

Name C31 "q(i)"
Invcdf c30 c31;           # Cálculo quantis
Normal 0 1.

Name C32 "y(2)^"
Sort c25 c32              # Ordenação vetor de dados

Plot C32*C31;             # Scatter plot
Title "Q-Q Plot da 2ª. Componente Principal";
Symbol.
```



- Observação da 1ª. tartaruga é suspeita.
  - √ Checar registros ou verificar anomalias na tartaruga
- Excetuado esse dado o scatter plot aparenta estar razoavelmente elíptico
- Verificar os plots dos outros conjunto de componentes principais.

Propriedades Assintóticas

- Assuma que a amostra são observações aleatórias de população normal p-variada
  - √ Autovalores desconhecidos são distintos e positivos
  - √ Distribuição amostral autovalores

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}) \overset{\text{as.}}{\sim} \mathrm{N}_p(\mathbf{0},2\boldsymbol{\Lambda}^2) \qquad \hat{\lambda}_i \overset{\text{as.}}{\sim} \mathrm{N}\left(\lambda_i, 2\frac{\lambda_i^2}{n}\right)$$

  - √ Distribuição amostral dos autovetores

$$\sqrt{n}(\hat{\mathbf{e}}_i-\mathbf{e}_i) \overset{\text{as.}}{\sim} \mathrm{N}_p(\mathbf{0},\mathbf{E}_i). \qquad \mathbf{E}_i = \lambda_i \sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_k}{(\lambda_k-\lambda_i)^2}\mathbf{e}_k\mathbf{e}_k'.$$

  - √ Cada $\hat{\lambda}_i$ é independente dos elementos de $\hat{\mathbf{e}}_i$ associados

- Intervalo de confiança aproximado para os λ_i de amostras suficientemente grandes

$$\hat{\lambda}_i \overset{\text{as.}}{\sim} \mathrm{N}\left(\lambda_i, 2\frac{\lambda_i^2}{n}\right) \qquad \mathrm{P}\left\{|\hat{\lambda}_i - \lambda_i| \le z_{\alpha/2}\,\lambda_i\sqrt{\frac{2}{n}}\right\} = 1-\alpha$$

$$\frac{\hat{\lambda}_i}{1 + z_{\alpha/2}\sqrt{\frac{2}{n}}} \le \lambda_i \le \frac{\hat{\lambda}_i}{1 - z_{\alpha/2}\sqrt{\frac{2}{n}}}$$

- Intervalos de confiança simultâneos de Bonferroni para m λ_i's
  - √ Trocar $z_{\alpha/2}$ por $z_{\alpha/2m}$.

Componentes Principais para Matrizes de Covariâncias com Estruturas Especiais

- Matriz diagonal: $\mathbf{\Sigma}_{p\times p} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{pp} \end{bmatrix}$.
  - √ j-ésimo autovetor: $\mathbf{e}'_j = [0, \dots, 0, 1, 0, \dots, 0]$
    - – 1 na j-ésima posição
  - √ j-ésima componente principal: $Y_j = \mathbf{e}'_j\mathbf{X} = X_j$
  - √ Não há ganho extraindo as componentes principais
    - – A padronização não altera a situação

$$\mathbf{P} = \mathbf{I}$$

Exemplo

- Estudo de poluição do ar em 41 cidades dos EUA
  - √ Ano: 1970
- Dados: *Usairpollution*{*MVA*}

- Variáveis:
  - √ SO2: conteúdo de dióxido de enxofre no ar, em μg/m^3.
  - √ Temp: temperatura média anual (ºF)
  - √ Indust: quantidade de empresas manufatureiras empregando pelo menos 20 empregados.
  - √ Pop: população (censo 1970), em milhares.
  - √ Vento: velocidade média anual de vento, em milhas/h
  - √ Precip: precipitação média anual, em polegadas
  - √ Dias: número médio anual de dias com precipitação

- Variáveis
  - √ ‘Resposta’: SO2
  - √ Ambientais: Vento, Precip, Dias, Temp
  - √ Demográficas: Indust, Pop

- Comandos em R:
  - √ Carregamento dos dados:

```
> library(MVA)
> library(HSAUR2, ADGofTest)# carrega os pacotes
> data(USairpollution)# carrega o banco de dados
> poluicao <- USairpollution
> colunas<- c("SO2", "Temp", "Indust", "Pop",
+ "Vento", "Precip", "Dias")
> colnames(poluicao)<- colunas
>
> poluicao$negtemp <- poluicao$Temp * (-1)
> poluicao$Temp <- NULL
```

  - √ Comentários:
    - Extrair componentes da matriz de correlações:
      - Variáveis estão em escalas muito distintas
    - Ignora a variável SO2 (considera só as ‘explicativas’)
    - Uso da temperatura negativa
      - As 6 variáveis têm valores altos, de maneira que representam um ambiente menos atrativo

√ Matriz de correlações:

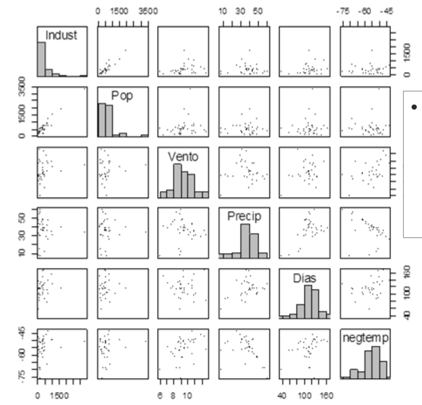
```
> # Descrição do conjunto de dados
>
> poluicao.corr <- cor(poluicao[,-1])
> poluicao.corr
             Indust         Pop       Vento      Precip       Dias     negtemp
Indust   1.00000000  0.95526935  0.23794683 -0.03241688 0.13182930  0.19004216
Pop      0.95526935  1.00000000  0.21264375 -0.02611873 0.04208319  0.06267813
Vento    0.23794683  0.21264375  1.00000000 -0.01299438 0.16410559  0.34973963
Precip  -0.03241688 -0.02611873 -0.01299438  1.00000000 0.49609671 -0.38625342
Dias     0.13182930  0.04208319  0.16410559  0.49609671 1.00000000  0.43024212
negtemp  0.19004216  0.06267813  0.34973963 -0.38625342 0.43024212  1.00000000
```

√ Comentário:
- Valores altos de correlação entre Indust e Pop

√ Matrix plot:
- Comandos em R:

```
> panel.hist <- function(x, ...){
+ usr <- par("usr"); on.exit(par(usr))
+ par(usr = c(usr[1:2], 0, 1.5) )
+ h <- hist(x, plot = FALSE)
+ breaks <- h$breaks; nB <- length(breaks)
+ y <- h$counts; y <- y/max(y)
+ rect(breaks[-nB], 0, breaks[-1], y, col = "grey", ...)
+     }
>
> pairs(poluicao[,-1], pch = ".", cex = 1.5)
> pairs(poluicao[,-1], diag.panel = panel.hist, pch = ".", cex = 1.5)
```

√ Gráfico matricial dos diagramas de dispersão



- Há pelo menos um outlier
  - √ Provavelmente mais de um
  - √ Chicago tem cerca do dobro de fábricas do que a 2ª. cidade com maior número

√ Análise de componentes principais:

```
> poluicao.acp <- princomp(poluicao[,-1], cor = TRUE)
> summary(poluicao.acp, loadings = TRUE)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
Standard deviation     1.4819456 1.2247218 1.1809526 0.8719099 0.33848287
Proportion of Variance 0.3660271 0.2499906 0.2324415 0.1267045 0.01909511
Cumulative Proportion  0.3660271 0.6160177 0.8484592 0.9751637 0.99425879
                            Comp.6
Standard deviation     0.185599752
Proportion of Variance 0.005741211
Cumulative Proportion  1.000000000
```

√ 3 primeiras componentes têm variâncias maiores que 1

- Compreendem 85% da variância total

√ Escores das componentes podem ser usados para visualização gráfica, com pouca perda de informação.

√ Interpretação das componentes principais:

```
Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Indust  -0.612  0.168 -0.273 -0.137  0.102  0.703
Pop     -0.578  0.222 -0.350               -0.695
Vento   -0.354 -0.131  0.297  0.869 -0.113
Precip         -0.623 -0.505  0.171  0.568
Dias    -0.238 -0.708        -0.311 -0.580
negtemp -0.330 -0.128  0.672 -0.306  0.558 -0.136
>
```

√ Componente 1

- Relaciona-se com algum índice de qualidade de vida
- Altos valores indicam um ambiente com menos qualidade

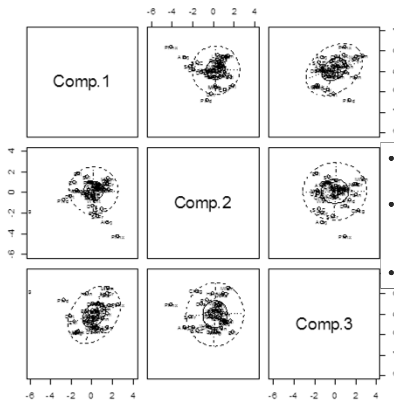
√ Componente 2

- Coeficientes altos para Precip e Dias
- Relaciona-se com chuva (clima úmido)

√ Componente 3

- Contraste entre Precip e negtemp
- Separa cidades c/ altas temperaturas e muita chuva (tipo de clima)

√ Boxplot bivariado para as 3 primeiras componentes

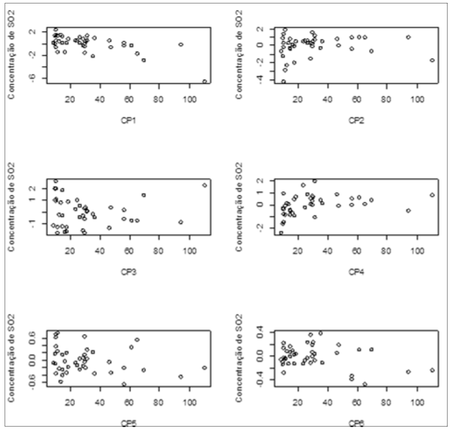


- Chicago é outlier
  - √ Suspeita-se de Phoenix e Philadelphia
- Phoenix aparenta ter a melhor qualidade de vida
  - √ Baseando-se nas 6 variáveis
- Bufallo tem clima mais seco

Questão Interessante

- Quais dentre as variáveis climáticas e ambientais são as melhores preditoras do grau de poluição do ar (concentração de SO_2)?
  - √ Esta questão é tratada com regressão linear múltipla
  - √ Potencial problema para aplicação dessa técnica:
    - Alta correlação entre Indust e Pop
  - √ Solução:
    - Retirar uma das variáveis
  - √ Alternativa:
    - Fazer regressão dos níveis de SO_2 com as componentes principais derivadas das 6 variáveis originais
    - Pode ser melhor regredir com todas as 6 componentes

√ SO2 dependendo das componentes principais



√ Regressão com as 6 componentes principais:

```
> poluicao.reg <- lm(SO2 ~ poluicao.acp$scores,
+  data = poluicao)
> summary(poluicao.reg)

Call:
lm(formula = SO2 ~ poluicao.acp$scores, data = poluicao)

Residuals:
    Min      1Q  Median      3Q     Max
-23.004  -8.542  -0.991   5.758  48.758

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 30.049      2.286  13.146 6.91e-15 ***
poluicao.acp$scoresComp.1   -9.942      1.542  -6.446 2.28e-07 ***
poluicao.acp$scoresComp.2    2.240      1.866   1.200  0.23845
poluicao.acp$scoresComp.3    0.375      1.935   0.194  0.84752
poluicao.acp$scoresComp.4    8.549      2.622   3.261  0.00253 **
poluicao.acp$scoresComp.5  -15.176      6.753  -2.247  0.03122 *
poluicao.acp$scoresComp.6  -39.271     12.316  -3.189  0.00306 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.64 on 34 degrees of freedom
Multiple R-squared:  0.6695,    Adjusted R-squared:  0.6112
F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

√ Escores da 1ª. componente predizem mais a resposta

√ Componentes com menor variância não têm necessariamente as menores correlações com a resposta

Exercício – Solo

- Análise de solo
  - √ 20 amostras
  - √ Variáveis:
    - areia (%)
    - sedimentos (%)
    - argila (%)
    - qte. material orgânico (%)
    - acidez do solo (pH)
  - √ Banco de dados: *BD_multivariada.xls/solo*

- Matriz de covariâncias amostral (S)

Covariances: areia; sedimentos; argila; morganico; ph

	areia	sedimentos	argila	morganico	ph
areia	138,32674				
sedimentos	-102,12274	79,73818			
argila	-36,20400	22,38455	13,81945		
morganico	-0,94221	1,52661	-0,58439	0,64345	
ph	-0,13579	0,11079	0,02500	0,03237	0,26263

- Autovalores de S

Eigenvalues

223,841	8,218	0,472	0,258	0,000

√ S é singular pois $\lambda_5 = 0$ (|S| = 0)

($X_1 + X_2 + X_3 = 100\%$)

- Componentes principais (p=5)

Principal Component Analysis: areia; sedimentos; argila; morganico; ph

Eigenanalysis of the Covariance Matrix

Eigenvalue	223,84	8,22	0,47	0,26	0,00
Proportion	0,962	0,035	0,002	0,001	0,000
Cumulative	0,962	0,997	0,999	1,000	1,000

Variable	PC1	PC2	PC3	PC4	PC5
areia	-0,785	0,223	-0,027	-0,004	-0,577
sedimentos	0,587	0,561	-0,086	-0,010	-0,577
argila	0,198	-0,784	0,113	0,014	-0,577
morganico	0,007	0,146	0,980	0,136	0,000
ph	0,001	0,002	0,137	-0,991	-0,000

√ y5 é constante para qualquer observação j

y5 = 0,577 (100)

√ Qualquer das três variáveis poderia ser eliminada

- Eliminada X_1 (areia)

√ maior variância amostral

tenderia dominar primeira componente

- Matriz de covariâncias amostral (S)

Covariances: sedimentos; argila; morganico; ph

79,7382	22,3846	1,52661	0,110789
22,3846	13,8194	-0,58439	0,025000
1,5266	-0,5844	0,64345	0,032368
0,1108	0,0250	0,03237	0,262632

- Autovalores de S

Eigenvalues

86,6403	7,0936	0,4714	0,2584

- Componentes principais (p = 4 – eliminada X_1)

Principal Component Analysis: sedimentos; argila; morganico; ph

Eigenanalysis of the Covariance Matrix

Eigenvalue	86,640	7,094	0,471	0,258
Proportion	0,917	0,075	0,005	0,003
Cumulative	0,917	0,992	0,997	1,000

Variable	PC1	PC2	PC3	PC4
sedimentos	0,956	-0,288	0,059	0,006
argila	0,294	0,945	-0,142	-0,018
morganico	0,015	-0,154	-0,979	-0,136
ph	0,001	-0,002	-0,137	0,991

√ Duas primeiras componentes explicam 99,2% da variância total

- 1ª. Componente: Índice de qualidade do solo em termos de % sedimentos e argila
  - sedimentos é a variável mais importante
- 2ª. Componente: Comparação entre % de sedimentos e % de argila
  - argila tem peso maior na componente
- 3ª. Componente: variável material orgânico

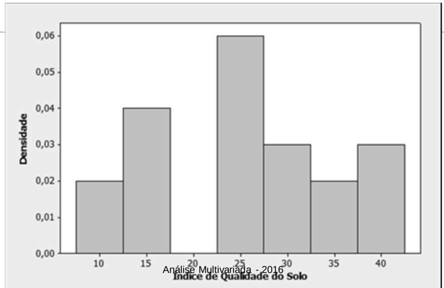
Variable N N* Mean SE Mean StDev Minimum Q1 Median Q3 Maximum
Scores Y1 20 0 24,86 2,08 9,31 11,57 14,77 24,29 32,43 39,70

- Scores da 1ª. Componente
  - √ Índice de qualidade do solo em termos de sedimentos e areia

    (material orgânico tem pouco participação)

Descriptive Statistics: Scores y1

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Scores Y1	20	0	24,86	2,08	9,31	11,57	14,77	24,29	32,43	39,70



- Diferença de escala e unidades da variáveis
  - √ Recomendável padronização para análise de componentes

- Componentes principais (p=4) – Matriz de correlação

Principal Component Analysis: sedimentos; argila; morganico; ph

Eigenanalysis of the Correlation Matrix

Eigenvalue	1,6757	1,1461	0,9601	0,2181
Proportion	0,419	0,287	0,240	0,055
Cumulative	0,419	0,705	0,945	1,000

Variable	PC1	PC2	PC3	PC4
sedimentos	0,710	0,182	-0,147	-0,664
argila	0,702	-0,241	0,111	0,661
morganico	0,025	0,836	-0,423	0,349
ph	0,042	0,459	0,887	-0,026

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.