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 Agrupamentos
8. Análise de Correlação Canônica
9. Referências

Análise de Correlação Canônica

Objetivo Principal

- Estudar as relações lineares existentes entre dois conjuntos de variáveis
- Pode ser usada para a análise de interdependência
- Exemplo:
  - √ Conjunto de variáveis medindo traços da personalidade de estudantes do ensino médio
  - √ Conjunto de variáveis medindo seus interesses vocacionais

Idéia Básica

- Resumir a informação de cada conjunto de variáveis em combinações lineares
  - √ Uma combinação do primeiro conjunto e uma combinação do segundo conjunto
- Critério para escolha dos coeficientes das combinações
  - √ Maximização da correlação entre os dois conjuntos de variáveis

5
Técnicas Multivariadas em Saúde - 2015

- Variáveis canônicas:
  - √ As combinações lineares dos conjuntos de variáveis
- Correlações canônicas:
  - √ Correlação entre as variáveis canônicas
  - √ Mede o grau de associação entre os dois conjuntos de variáveis

6
Técnicas Multivariadas em Saúde - 2015

- Às vezes somos capazes de distinguir entre os dois conjuntos:
  - √ Variáveis dependentes
    - – Conjunto de variáveis que estamos interessados em explicar
  - √ Variáveis independentes:
    - – Variáveis explanatórias

7
Técnicas Multivariadas em Saúde - 2015

- Análise de correlações canônicas difere da análise de componentes principais e análise fatorial
  - √ Essas tratam apenas das situações em que se têm um único conjunto de variáveis
  - √ Análise de correlações canônicas:
    - – Busca de pares de variáveis canônicas que podem explicar muito da interdependência entre dois conjuntos de variáveis
  - √ Análise de componentes principais:
    - – Busca de um pequeno número de componentes que possa explicar muito da variância

8
Técnicas Multivariadas em Saúde - 2015

Modelagem

Modelo Teórico

- Suponha dois vetores aleatórios:
  - √ $\mathbf{X}_{px1}$, com vetor de médias $\boldsymbol{\mu}_{\mathbf{X}}$ (px1) e matriz de covariâncias $\boldsymbol{\Sigma}_{\mathbf{X}}$ (pxp)
  - √ $\mathbf{Y}_{qx1}$, com vetor de médias $\boldsymbol{\mu}_{\mathbf{Y}}$ (qx1) e matriz de covariâncias $\boldsymbol{\Sigma}_{\mathbf{Y}}$ (qxq)

- Matriz de covariâncias dos dois vetores aleatórios

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y}) = \begin{matrix} p \\ q \end{matrix}\left[\begin{array}{c|c} \overset{p}{\boldsymbol{\Sigma}_{XX}} & \overset{q}{\boldsymbol{\Sigma}_{XY}} \\ \hline \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{array}\right]$$

  - √ $\boldsymbol{\Sigma}_{\mathbf{XY}}$: matriz das covariâncias entre as variáveis aleatórias de X e de Y
    - Dimensão pxq
  - √ $\boldsymbol{\Sigma}_{\mathbf{YX}} = \boldsymbol{\Sigma}_{\mathbf{XY}}'$
  - √ Mesmo procedimento é válido para a matriz de correlações
  - √ Quando p e q são grandes há uma maior dificuldade em interpretar todos os elementos da matriz $\boldsymbol{\Sigma}_{\mathbf{XY}}$ conjuntamente

Proposta

- Relações existentes entre os vetores **X** e **Y** são estudadas por meio da análise de combinações lineares desses vetores
  - √ Combinações são construídas de maneira a se tornarem fortemente correlacionadas entre si
  - √ Essas combinações são denominadas variáveis canônicas

Procedimento

- Em cada estágio são construídas duas combinações lineares
  - √ Uma em **X** e outra em **Y**
- A técnica assegura que as variáveis canônicas de um par são não correlacionadas com as de outro par
- Quantidade possível de variáveis canônicas:
  - √ k = min(p, q)

13

Variáveis Canônica Teóricas

- 1º. par de variáveis canônicas:

$$U_1 = \underset{(1\times p)(p\times 1)}{\mathbf{a}_1'}\ \mathbf{X} \text{ e } V_1 = \underset{(1\times q)(q\times 1)}{\mathbf{b}_1'}\ \mathbf{Y}$$

  - √ $\mathbf{a}_1$ e $\mathbf{b}_1$ são escolhidos para maximizar a correlação entre U_1 e V_1, com $Var(U_1) = Var(V_1) = 1$
- 2º. par de variáveis canônicas:

$$U_2 = \mathbf{a}_2'\mathbf{X} \text{ e } V_2 = \mathbf{b}_2'\mathbf{Y}$$

  - √ $\mathbf{a}_2$ e $\mathbf{b}_2$ são escolhidos de modo que a correlação entre U_2 e V_2, seja maximizada no conjunto de combinações lineares de **X** e **Y** que não são correlacionas com U_1 e V_1
    - $Var(U_1) = Var(V_1) = 1$

14

- 2º. par de variáveis canônicas:

$$U_k = \mathbf{a}_k'\mathbf{X} \text{ e } V_k = \mathbf{b}_k'\mathbf{Y}$$

  - √ $\mathbf{a}_k$ e $\mathbf{b}_k$ são escolhidos de maneira a maximizar a correlação entre U_k e V_k, as quais não são correlacionadas com as (k–1) primeiras variáveis canônicas
- Correlação canônica:
  - √ Correlação entre as combinações lineares U_i e V_i, i=1, 2, ..., min(p, q)

15

Vetores de Coeficientes

- $\mathbf{a}_k$ e $\mathbf{b}_k$ são soluções do sistema de equações:

$$\begin{cases} \left(\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{YX}} - \lambda_k\boldsymbol{\Sigma}_{\mathbf{XX}}\right)\mathbf{a}_k &= \mathbf{0} \\ \left(\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}} - \lambda_k\boldsymbol{\Sigma}_{\mathbf{YY}}\right)\mathbf{b}_k &= \mathbf{0} \end{cases}$$

  - √ λ_k é o k-ésimo maior autovalor da matriz:
    - $(\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{YX}})$ ou, equivalentemente de
    - $(\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}})$

  ou seja, satisfaz as seguintes equações características:

$$\begin{cases} \left|\boldsymbol{\Sigma}_{\mathbf{XY}}\boldsymbol{\Sigma}_{\mathbf{YY}}^{-1}\boldsymbol{\Sigma}_{\mathbf{YX}} - \lambda_k\boldsymbol{\Sigma}_{\mathbf{XX}}\right| &= 0 \\ \left|\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}} - \lambda_k\boldsymbol{\Sigma}_{\mathbf{YY}}\right| &= 0 \end{cases}$$

17

- Valor absoluto da correlação canônica entre U_k e V_k: $\sqrt{\lambda_k}$

$$\rho_k^* = \lambda_k = \mathrm{Corr}^2(U_k, V_k) = \frac{\mathbf{a}_k' \boldsymbol{\Sigma}_{\mathbf{XY}} \mathbf{b}_k}{\left(\mathbf{a}_k' \boldsymbol{\Sigma}_{\mathbf{XX}} \mathbf{a}_k\right)\left(\mathbf{b}_k' \boldsymbol{\Sigma}_{\mathbf{YY}} \mathbf{b}_k\right)}$$

19

- As variáveis canônicas podem ser construídas para as variáveis padronizadas:

$$\begin{cases} \left(\mathbf{P}_{\mathbf{XY}} \mathbf{P}_{\mathbf{YY}}^{-1} \mathbf{P}_{\mathbf{YX}} - \lambda_k \mathbf{P}_{\mathbf{XX}}\right) \mathbf{a}_k = \mathbf{0} \\ \left(\mathbf{P}_{\mathbf{YX}} \mathbf{P}_{\mathbf{XX}}^{-1} \mathbf{P}_{\mathbf{XY}} - \lambda_k \mathbf{P}_{\mathbf{YY}}\right) \mathbf{b}_k = \mathbf{0} \end{cases}$$

  √ com equações características:

$$\begin{cases} \left|\mathbf{P}_{\mathbf{XY}} \mathbf{P}_{\mathbf{YY}}^{-1} \mathbf{P}_{\mathbf{YX}} - \lambda_k \mathbf{P}_{\mathbf{XX}}\right| = 0 \\ \left|\mathbf{P}_{\mathbf{YX}} \mathbf{P}_{\mathbf{XX}}^{-1} \mathbf{P}_{\mathbf{XY}} - \lambda_k \mathbf{P}_{\mathbf{YY}}\right| = 0 \end{cases}$$

  √ $\mathbf{P_{XX}}$ e $\mathbf{P_{YY}}$: matrizes de correlações teóricas das variáveis dos vetores **X** e **Y**, respectivamente

  √ $\mathbf{P_{XY}}$: matriz de correlações entre as variáveis que estão no vetor **X** e aquelas que estão no vetor **Y**

21

Estimação das Variáveis Canônicas

- Sejam amostras aleatórias de tamanho n dos vetores **X** e **Y**
  - √ $\mathbf{\Sigma_{XX}}$, $\mathbf{\Sigma_{YY}}$, $\mathbf{\Sigma_{XY}}$ e $\mathbf{\Sigma_{YX}}$ são estimadas por $\mathbf{S_{XX}}$, $\mathbf{S_{YY}}$, $\mathbf{S_{XY}}$ e $\mathbf{S_{YX}}$, respectivamente
  - √ Os sistemas de equações são resolvidos utilizando as matrizes amostrais
  - √ $\mathbf{P_{XX}}$, $\mathbf{P_{YY}}$, $\mathbf{P_{XY}}$ e $\mathbf{P_{YX}}$ são estimadas pelas respectivas matrizes de correlações amostrais $\mathbf{R_{XX}}$, $\mathbf{R_{YY}}$, $\mathbf{R_{XY}}$ e $\mathbf{R_{YX}}$.

22

Exemplo

- (Mingoti, 2005)
  Estudo com 44 vendedores de empresa
  - √ Variáveis:
    - X_1: desempenho nas vendas
    - X_2: desempenho nos lucros
    - X_3: captação de novos clientes
    - Y_1: desempenho em teste de habilidade escrita
    - Y_2: esempenho em teste de habilidade lógica
    - Y_3: desempenho em teste de habilidade social
    - Y_4: desempenho em teste de habilidade matemática

23

- Matrizes de correlações amostrais:

$$\mathbf{R_{XX}} = \begin{bmatrix} 1 & 0,923 & 0,894 \\ 0,923 & 1 & 0,849 \\ 0,894 & 0,849 & 1 \end{bmatrix} \mathbf{R_{YY}} = \begin{bmatrix} 1 & 0,602 & 0,194 & 0,366 \\ 0,602 & 1 & 0,457 & 0,649 \\ 0,194 & 0,457 & 1 & 0,606 \\ 0,366 & 0,649 & 0,606 & 1 \end{bmatrix}_{4\times4}$$

$$\mathbf{R_{XY}} = \begin{bmatrix} 0,529 & 0,757 & 0,715 & 0,932 \\ 0,492 & 0,783 & 0,502 & 0,949 \\ 0,674 & 0,718 & 0,687 & 0,860 \end{bmatrix}_{3\times4}$$

- Autovalores de ($\mathbf{R_{XX}}^{-1}\mathbf{R_{XY}}\mathbf{R_{YY}}^{-1}\mathbf{R_{YX}}$)

$$\underset{3\times3}{\mathbf{R}_{\mathbf{XX}}^{-1}}\underset{3\times4}{\mathbf{R_{XY}}}\underset{4\times4}{\mathbf{R}_{\mathbf{YY}}^{-1}}\underset{4\times3}{\mathbf{R_{YX}}} = \begin{bmatrix} 0,517 & 0,100 & 0,471 \\ 0,167 & 0,797 & -0,075 \\ 0,318 & 0,087 & 0,581 \end{bmatrix}_{3\times3} \begin{cases} \hat{\lambda}_1 & = 0,9892 \\ \hat{\lambda}_2 & = 0,7531 \\ \hat{\lambda}_3 & = 0,1591 \end{cases}$$

- Pares de variáveis canônicas:

$$\begin{cases} U_1 & = 0,4371\,Z_1 + 0,2087\,Z_2 + 0,3909\,Z_3 \\ V_1 & = 0,2563\,W_1 + 0,1031\,W_2 + 0,1846\,W_3 + 0,6417\,W_4 \end{cases}$$

$$\begin{cases} U_2 & = -1,2564\,Z_1 + 2,4435\,Z_2 - 1,1139\,Z_3 \\ V_2 & = -0,6793\,W_1 + 0,6238\,W_2 - 1,1025\,W_3 + 0,7209\,W_4 \end{cases}$$

$$\begin{cases} U_3 & = -2,7968\,Z_1 + 0,9514\,Z_2 + 1,9194\,Z_3 \\ V_3 & = 1,0198\,W_1 - 0,4017\,W_2 - 0,5963\,W_3 + 0,1138\,W_4 \end{cases}$$

√ Z_i e W_j são as variáveis originais padronizadas

$$\begin{cases} Z_i & = \frac{X_i - \bar{X}_i}{s_{X_i}},\ i = 1,2,3, \text{ e} \\ W_j & = \frac{Y_j - \bar{Y}_j}{s_{Y_j}},\ j = 1,2,3,4. \end{cases}$$

- Variáveis originais:

	Variável	$\bar{x}_i$	s_i
X_1	Desempenho vendas	24,76	1,87
X_2	Desempenho lucros	26,69	2,61
X_3	Captação novos clientes	25,69	1,19

	Variável	$\bar{y}_j$	s_j
Y_1	Habilidade escrita	5,65	1,78
Y_2	Habilidade lógica	7,00	1,65
Y_3	Habilidade social	5,27	1,08
Y_4	Habilidade matemática	15,00	5,53

√ U_i: pode ser interpretada como índice de desempenho global do vendedor

√ V_i: pode ser interpretada como índice de desempenho global do vendedor nos testes

- Score de indivíduo:

√ Vendedor com:

$$\begin{cases} x_1 & = 26 \\ x_2 & = 29 \\ x_3 & = 27 \end{cases} \begin{cases} z_1 & = 0,662 \\ z_2 & = 0,885 \\ z_3 & = 1,094 \end{cases} \quad u_1 = 0,902.$$

$$\begin{cases} y_1 & = 7 \\ y_2 & = 9 \\ y_3 & = 7 \\ y_4 & = 20 \end{cases} \begin{cases} w_1 & = 0,758 \\ w_2 & = 1,212 \\ w_3 & = 1,606 \\ w_4 & = 0,904 \end{cases} \quad v_1 = 1,195.$$

- Correlações canônicas:

$$\text{Corr}(U_1, V_1) = \sqrt{0,9892} = 0,9945.$$

√ V_1 é a melhor combinação linear para predizer U_1 (ou vice-versa)
  - V_1 é a combinação linear mais correlacionada com U_1.

√ Corr(U_1, V_1) > 0
  - Candidatos com maiores scores V_1 seriam os melhores candidatos ao emprego

√ Outras correlações canônicas:

$$\begin{cases} \text{Corr}(U_2, V_2) & = \sqrt{0,7531} = 0,8678. \\ \text{Corr}(U_3, V_3) & = \sqrt{0,1591} = 0,3988. \end{cases}$$

28

- Conclusão:

√ Estratégia de contratação:
  - Submeter cada candidato aos quatro testes psicológicos e calcular seu escore V_1.
  - Candidatos classificados de acordo com os valores de V_1

29

Variáveis Canônicas e Originais

- *Canonical loadings*:

√ Correlação das variáveis canônicas com as variáveis originais

$$\mathbf{R}^*_{U_k,\mathbf{X}} = \underset{(p\times p)}{\mathbf{R}_{\mathbf{XX}}} \underset{(p\times 1)}{\mathbf{a}_k}.$$

$$\mathbf{R}^*_{V_k,\mathbf{Y}} = \underset{(q\times q)}{\mathbf{R}_{\mathbf{YY}}} \underset{(q\times 1)}{\mathbf{b}_k}.$$

$$\mathbf{R}^*_{U_k,\mathbf{Y}} = \underset{(q\times p)}{\mathbf{R}_{\mathbf{YX}}} \underset{(p\times 1)}{\mathbf{a}_k}.$$

$$\mathbf{R}^*_{V_k,\mathbf{X}} = \underset{(p\times q)}{\mathbf{R}_{\mathbf{XY}}} \underset{(q\times 1)}{\mathbf{b}_k}.$$

31

Exemplo

- (Mingoti, 2005)
  Estudo com 44 vendedores de empresa

$$\mathbf{a}_1' = \begin{bmatrix} 0,4371 & 0,2087 & 0,3909 \end{bmatrix}$$

$$\mathbf{R}_{\mathbf{XX}}\,\mathbf{a}_1 = \begin{bmatrix} 1 & 0,923 & 0,894 \\ 0,923 & 1 & 0,849 \\ 0,894 & 0,849 & 1 \end{bmatrix} \begin{bmatrix} 0,4371 \\ 0,2087 \\ 0,3909 \end{bmatrix} = \begin{bmatrix} 0,9792 \\ 0,9440 \\ 0,9589 \end{bmatrix}$$

$$\begin{cases} \text{Corr}(U_1, X_1) & = 0,9792. \\ \text{Corr}(U_1, X_2) & = 0,9440. \\ \text{Corr}(U_1, X_3) & = 0,9589. \end{cases}$$

32

√ Proporção da variância total explicada pelas variáveis canônicas (separadamente)

$$\text{PVTE}_{U_k} = \frac{\sum_{i=1}^{p} \text{Corr}(U_k, X_i)^2}{p} \times 100.$$

$$\text{PVTE}_{V_k} = \frac{\sum_{i=1}^{q} \text{Corr}(V_k, Y_i)^2}{q} \times 100.$$

35

Exemplo

- Estudo com 44 vendedores de empresa
  - √ Variáveis canônicas do vetor **X**:

Variável		*Canonical loadings*					
		U_1	U_2	U_3	V_1	V_2	V_3
X_1	Desempenho vendas	0,9793	–0,0049	–0,2023	0,9740	–0,0043	–0,0807
X_2	Desempenho lucros	0,9443	0,3291	–0,0013	0,9392	0,2856	–0,0005
X_3	Captação novos clientes	0,9589	–0,1703	0,2269	0,9537	–0,1478	0,0905
	PVTE (%)	92,34	4,58	3,08			

36

√ Variáveis canônicas do vetor **Y**:

Variável		*Canonical loadings*					
		V_1	V_2	V_3	U_1	U_2	U_3
Y_1	Habilidade escrita	0,6009	–0,2535	0,7047	0,5976	–0,2200	0,2811
Y_2	Habilidade lógica	0,7795	0,1797	0,0164	0,7753	0,1559	–0,0066
Y_3	Habilidade social	0,6899	–0,5124	–0,5104	0,6861	–0,4447	–0,2036
Y_4	Habilidade matemática	0,9468	0,2093	–0,1307	0,9417	0,1816	0,0521
	PVTE (%)	58,53	10,07	19,36			

√ De maneira geral, correlações com as variáveis originais são maiores para o 1º. Par de variáveis canônicas e decrescem para os outros dois pares

37

Inferência

39

Inferência para as Variáveis Canônicas

- Problema:
  - √ Se os vetores **X** e **Y** entre si (ou não correlacionados)
    - Análise canônica será inútil
    - $\mathbf{a}'_k\,\mathbf{X}$ e $\mathbf{b}'_k\,\mathbf{Y}$ terão correlação zero para qualquer escolha de vetores $\mathbf{a}'_k$ e $\mathbf{b}'_k$.
  - √ Importante:
    - Análise da matriz de covariâncias (correlações) cruzadas a fim de determinar se elas são próximas ou não da matriz nula

40

- Hipóteses:
  - √ H_0: $\mathbf{\Sigma_{XY}} = \mathbf{0}_{pxq}$ vs. H_1: $\mathbf{\Sigma_{XY}} \neq \mathbf{0}_{pxq}$.
- Estatística de teste:
  - Teste da razão de verossimilhança

$$\begin{aligned}\text{TRV} &= -2\ln(\Lambda) = n\ln\left(\frac{|\mathbf{S_{XX}}||\mathbf{S_{YY}}|}{\mathbf{S}}\right)\\ &= n\ln\left(\frac{|\mathbf{R_{XX}}||\mathbf{R_{YY}}|}{\mathbf{R}}\right)\\ &= -n\ln\left(\prod_{i=1}^{p}\left(1-\hat{\rho}_i^{*2}\right)\right)\\ &= -n\ln\left(\prod_{i=1}^{p}\left(1-\hat{\lambda}_i\right)\right).\end{aligned}$$

- Distribuição amostral: $\text{TRV} \overset{H_0}{\sim} \chi^2_{pq}$.

46

Teste de Hipóteses Aproximado

- Teste aproximado para as matrizes de covariâncias e de correlação
  - √ Válido apenas quando **X** e **Y** são normais multivariados
  - √ Nesse caso, também é possível avaliar a significância das variáveis canônicas obtidas
  - √ Teste assintótico
    - Válido para amostras suficientemente grandes

45

- Correção de Bartllet para amostras pequenas
  - √ H_0: $\mathbf{\Sigma_{XY}} = \mathbf{0}_{pxq}$ vs. H_1: $\mathbf{\Sigma_{XY}} \neq \mathbf{0}_{pxq}$.
  - √ Estatística de teste:
    - Teste da razão de verossimilhança

$$\begin{aligned}\text{TRV} &= -\left(n-1-\frac{p+q+1}{2}\right)\ln\left(\prod_{i=1}^{p}\left(1-\hat{\rho}_i^{*2}\right)\right)\\ &= -\underbrace{\left(n-1-\frac{p+q+1}{2}\right)}_{\text{correção de Bartllet}}\ln\left(\prod_{i=1}^{p}\left(1-\hat{\lambda}_i\right)\right).\end{aligned}$$

  - √ Distribuição amostral: $\text{TRV} \overset{H_0}{\sim} \chi^2_{pq}$.

47

Teste de Significância

- Teste para a significância das correlações canônicas:
  - √ Suposição:
    - **X** e **Y** são normais multivariadas e $\mathbf{\Sigma_{XY}} \neq \mathbf{0}$
- Objetivo:
  - √ Testar se as m primeiras correlações canônicas são significativas
    - As variáveis canônicas correspondentes seriam as mais importantes para a caracterização da informação dos dois conjuntos em estudo
    - m < k = min(p, q)

48

- Hipóteses:
  - √ $\mathrm{H}_0^{(m)} : \rho_1^{*2} \neq 0, \rho_2^{*2} \neq 0, \ldots, \rho_m^{*\,2} \neq 0, \rho_{m+1}^{*\,2} = 0, \ldots, \rho_k^{*2} = 0.$
  - √ $\mathrm{H}_1^{(m)} : \rho_i^{*2} \neq 0$, para algum $i \geq m+1$.
- Estatística de teste:
  - Teste da razão de verossimilhança

$$\mathrm{TRV} = -\left(n - 1 - \frac{p+q+1}{2}\right) \ln\left(\prod_{i=m+1}^{k}\left(1 - \hat{\rho}_i^{*2}\right)\right)$$
$$= -\left(n - 1 - \frac{p+q+1}{2}\right) \ln\left(\prod_{i=m+1}^{k}\left(1 - \hat{\lambda}_i\right)\right).$$

- Distribuição amostral: $\mathrm{TRV} \overset{\mathrm{H}_0}{\sim} \chi^2_{(p-m)(q-m)}$.

49

Exemplo

- Estudo com 44 vendedores de empresa
  - √ Supondo normalidade multivariada de **X** e **Y**.
- Teste para a matriz de correlações
  - √ H_0: $\mathbf{P_{XY}} = \mathbf{0}_{3x4}$ vs. $\mathbf{P_{XY}} \neq \mathbf{0}$.
  - √ Estatística de teste:

$$\mathrm{TRV} = -\left(n - 1 - \frac{p+q+1}{2}\right) \ln\left(\prod_{i=1}^{p}\left(1 - \hat{\lambda}_i\right)\right)$$
$$= -\left(44 - 1 - \frac{3+4+1}{2}\right) \ln\left[(1-0,982)(1-0,7531)(1-0,1591)\right]$$
$$= 237,91.$$

  - √ Ponto crítico: $\chi^2_{3\times4;\,0,05} = 21,06.$
  - √ Rejeita-se H_0.

50

- Conclusão do teste:
  - √ A matriz de correlações teóricas entre os vetores **X** e **Y** é diferente da matriz nula

51

- Teste de significância das correlações canônicas
  - √ $\mathbf{P_{XY}} \neq \mathbf{0}$.
- Hipóteses: $H_0^{(2)} : \rho_1^{*2} \neq 0, \rho_2^{*2} \neq 0, \rho_3^{*2} = 0$ vs. $H_1^{(2)} : \rho_3^{*2} \neq 0$.
- Estatística de teste

$$\begin{aligned} \text{TRV} &= -\left(n - 1 - \frac{p+q+1}{2}\right) \ln\left(\prod_{i=m+1}^{k} \left(1 - \hat{\lambda}_i\right)\right) \\ &= -\left(44 - 1 - \frac{3+4+1}{2}\right) \ln(1 - 0,1591) \\ &= 6,76. \end{aligned}$$

  - √ Ponto crítico:

$$\chi^2_{(p-m)(q-m);\,0,05} = \chi^2_{(3-2)(4-2);\,0,05} = \chi^2_{2;0,05} = 5,99.$$

  - √ Não há evidência amostral para se rejeitar H_0.

52

- Conclusão:
  - √ Há indicação para este problema que as duas correlações canônicas e, consequentemente os dois pares de variáveis canônicas são os mais importantes

53

Comentários Adicionais

54

Regressão Linear Múltipla

- Relação entre as correlações canônicas e a regressão linear múltipla:
  - √ Pode ser demonstrado que a regressão linear múltipla é um caso especial da análise de correlações canônicas

55

- Considere o vetor: $[Y, \ X_1, \ X_2, \ \ldots, \ X_p]'$
  - √ Y: variável resposta
  - √ X_i: variável explicativa, i = 1, 2, ..., p
  - √ n observações amostrais deste vetor
- Objetivo:
  - √ Encontrar a combinação linear de $[X_1, X_2, ..., X_p]'$ que tenha a maior correlação amostral com Y

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Tem a maior correlação com Y

- Solução por análise de correlações canônicas:

$$\begin{cases} U & = Y \\ V & = \hat{Y} \end{cases}$$

  - √ A solução é a mesma obtida pela metodologia de mínimos quadrados ordinários
  - √ Correlação canônica

$$\begin{aligned} \text{Corr}(Y, \hat{Y}) &= \frac{\sum_{j=1}^{n}(Y_j - \bar{Y})(\hat{Y}_j - \bar{Y})}{\left(\sum_{j=1}^{n}(Y_j - \bar{Y})^2\right)\left(\sum_{j=1}^{n}(\hat{Y}_j - \bar{Y})^2\right)} \\ &= \frac{\text{SQ}_{\text{modelo}}}{\text{SQ}_{\text{total}}} = R^2 \end{aligned}$$

Exemplo

- Estudo com 44 vendedores de empresa
- Modelo de regressão linear
  - √ Considerando as variáveis padronizadas

$$\begin{cases} \hat{Z}_1 & = 0,17W_1 + 0,13W_2 + 0,24W_3 + 0,65W_4 \\ R^2 & = 0,9552 \end{cases}$$

- Modelo de correlações canônicas

$$\begin{cases} \mathbf{X} & = X_1 \\ \mathbf{Y} & = \left[Y_1, \ Y_2, \ Y_3 \ Y_4\right]'. \end{cases}$$

  - √ Variáveis canônicas

$$\begin{cases} U_1 & = Z_1 \\ V_1 & = 0,17W_1 + 0,13W_2 + 0,24W_3 + 0,65W_4. \end{cases}$$

  - √ O melhor preditor de $U_1 = Z_1$ é a variável canônica V_1.

Casos Particulares

- São casos particulares da análise de correlações canônicas:
  - √ ANOVA univariada
  - √ Análise discriminante

Aplicações

- Variáveis psicológicas:
  - √ Controle: locus de controle

    (atribuição do indivíduo a responsabilidade por eventos em suas vidas)
  - √ Conceito: auto-conceito
  - √ Motivacao: motivação
- Variáveis acadêmicas:
  - √ Leitura:
  - √ Escrita:
  - √ Matematica:
  - √ Ciencia:
  - √ Sexo: indicadora de mulher (0 = homem, 1 = mulher)

Exemplo

- Fatores Psicológicos e Acadêmicos
  - √ Pesquisa com 3 variáveis psicológicas e 4 variáveis acadêmicas (escores padronizados de testes) e gênero, para 600 calouros universitários.
  - √ Fonte: http://www.ats.ucla.edu/stat/r/dae/canonical.htm
  - √ Dados: *mmreg.csv*

Objetivo

- Verificar como as variáveis psicológicas se relacionam com as variáveis acadêmicas e com sexo
  - √ Interessa-se também verificar quantas variáveis canônicas (dimensões) são necessárias para entender a associação entre os dois conjuntos de dados

- Métodos que podem ser utilizados:
  - √ Análise de correlações canônicas
  - √ Regressões separadas de mínimos quadrados ordinários
    - Cada variável em um conjunto de dados
    - Não serão produzidos resultados multivariados e não haverá informação sobre dimensionalidade
  - √ Regressão linear múltipla
    - Opção razoável se não interesse em dimensionalidade

66

- Comandos em R

```
> library(ggplot2)
> library(GGally)
> library(CCA)
>
> #  Carregamento de Dados
>
> calouros <- read.csv("mmreg.csv")
> colnames(calouros) <- c("Controle", "Conceito", "Motivacao", "Leitura",
+ "Escrita", "Matematica", "Ciencia", "Sexo")
```

  - √ Descritiva básica

```
> summary(calouros[,1:7])
    Controle          Conceito          Motivacao         Leitura
 Min.   :-2.23000   Min.   :-2.620000   Min.   :0.0000   Min.   :28.3
 1st Qu.:-0.37250   1st Qu.:-0.300000   1st Qu.:0.3300   1st Qu.:44.2
 Median : 0.21000   Median : 0.030000   Median :0.6700   Median :52.1
 Mean   : 0.09653   Mean   : 0.004917   Mean   :0.6608   Mean   :51.9
 3rd Qu.: 0.51000   3rd Qu.: 0.440000   3rd Qu.:1.0000   3rd Qu.:60.1
 Max.   : 1.36000   Max.   : 1.190000   Max.   :1.0000   Max.   :76.0
    Escrita         Matematica        Ciencia
 Min.   :25.50   Min.   :31.80   Min.   :26.00
 1st Qu.:44.30   1st Qu.:44.50   1st Qu.:44.40
 Median :54.10   Median :51.30   Median :52.60
 Mean   :52.38   Mean   :51.85   Mean   :51.76
 3rd Qu.:59.90   3rd Qu.:58.38   3rd Qu.:58.65
 Max.   :67.10   Max.   :75.50   Max.   :74.20
```

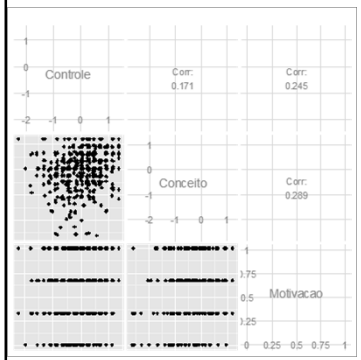
Sexo
Feminino: 327
Masculino: 273

67

- Conjunto das variáveis psicológicas

```
psic <- calouros[, 1:3]          # variáveis psicológicas

     # Scatter matrix plot - variáveis psicológicas
ggpairs(psic, axisLabels="internal", upper = list(params = list(size = 4)))
```

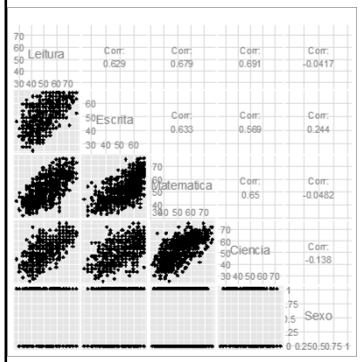


- Variável Motivação é discreta

68

- Conjunto das variáveis acadêmicas

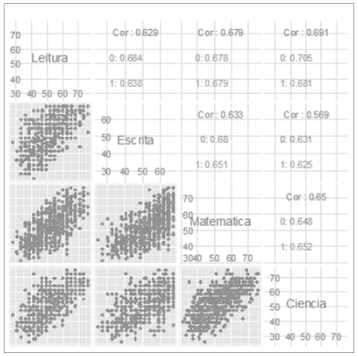
```
> acad <- calouros[, 4:8]# variáveis acadêmicas
>
> # Scatter matrix plot - variáveis acadêmicas
> ggpairs(acad, axisLabels="internal", upper = list(params = list(size = 4)))
```



- Correlações moderadas entre as variáveis acadêmicas
  - √ São as mesmas por sexo?

69

- Scatter matrix plot das variáveis acadêmicas
  - √ Estratificadas por Sexo



- Aparentemente, as correlações por sexo não diferem muito entre si.


```
> acad <- calouros[, 4:8]
> # correlações
> matcor(psic, acad)# library(CCA)
```

√ Matriz de correlações

```
$Xcor
          Controle  Conceito Motivacao
Controle  1.0000000 0.1711878 0.2451323
Conceito  0.1711878 1.0000000 0.2885707
Motivacao 0.2451323 0.2885707 1.0000000
```

R_{XX}

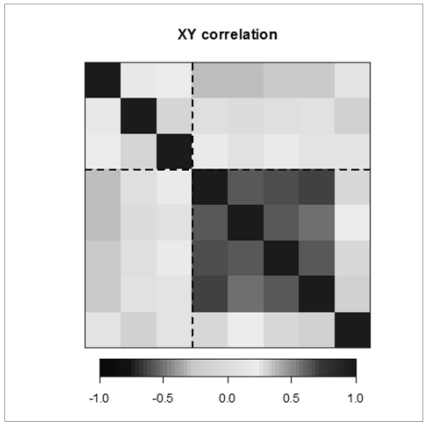
```
$Ycor
             Leitura   Escrita Matematica    Ciencia        Sexo
Leitura    1.00000000 0.6285909  0.6792757  0.6906929 -0.04174278
Escrita    0.62859089 1.0000000  0.6326664  0.5691498  0.24433183
Matematica 0.67927568 0.6326664  1.0000000  0.6495261 -0.04821830
Ciencia    0.69069291 0.5691498  0.6495261  1.0000000 -0.13818587
Sexo       -0.04174278 0.2443318 -0.0482183 -0.1381859  1.00000000
```

R_{YY}

```
$XYcor
             Controle    Conceito  Motivacao     Leitura     Escrita Matematica
Controle   1.0000000  0.17118778 0.24513227  0.37356505 0.35887684  0.3372690
Conceito   0.1711878  1.00000000 0.28857075  0.06065584 0.01944856  0.0535977
Motivacao  0.2451323  0.28857075 1.00000000  0.21060992 0.25424818  0.1950135
Leitura    0.3735650  0.06065584 0.21060992  1.00000000 0.62859089  0.6792757
Escrita    0.3588768  0.01944856 0.25424818  0.62859089 1.00000000  0.6326664
Matematica 0.3372690  0.05359770 0.19501347  0.67927568 0.63266640  1.0000000
Ciencia    0.3246269  0.06982633 0.11566948  0.69069291 0.56914983  0.6495261
Sexo       0.1134108 -0.12595132 0.09810277 -0.04174278 0.24433183 -0.0482183
               Ciencia        Sexo
Controle    0.32462694  0.11341075
Conceito    0.06982633 -0.12595132
Motivacao   0.11566948  0.09810277
Leitura     0.69069291 -0.04174278
Escrita     0.56914983  0.24433183
Matematica  0.64952612 -0.04821830
Ciencia     1.00000000 -0.13818587
Sexo       -0.13818587  1.00000000
```

R

- Visualização da matriz de correlações



- Análise de correlações canônicas:
  - √ Comandos em R:

```
> # Análise de correlações canônicas
>
> calouros.cc1 <- cc(psic, acad)
>
```

  - √ Correlações canônicas

```
> # Valores das correlações canônicas
> calouros.cc1$cor
[1] 0.4641 0.1675 0.1040
```

√ Coeficientes canônicos:

```
> # Coeficientes canônicos
> calouros.cc1[3:4]
$xcoef
             [,1]    [,2]    [,3]
Controle  -1.2538 -0.6215 -0.6617
Conceito   0.3513 -1.1877  0.8267
Motivacao -1.2624  2.0273  2.0002

$ycoef
                [,1]      [,2]      [,3]
Leitura    -0.044621 -0.004910  0.021381
Escrita    -0.035877  0.042071  0.091307
Matematica -0.023417  0.004229  0.009398
Ciencia    -0.005025 -0.085162 -0.109835
Sexo       -0.632119  1.084642 -1.794647
```

√ Interpretação:

- Uma unidade de aumento no escore da variável Leitura diminui 0,046 na 1ª. variável canônica do conjunto 2 (V_1), mantidas constantes todas as outras variáveis
- Ser mulher leva a um decréscimo na 1ª. dimensão canônica (V_1), mantidos constantes os outros preditores

√ Cargas canônicas:

```
> # Canonical loadings
> calouros.cc2 <- comput(psych, acad, calouros.cc1)
>
> # display canonical loadings
> calouros.cc2[3:6]
$corr.X.xscores
             [,1]    [,2]    [,3]
Controle  -0.90405 -0.3897 -0.1756
Conceito  -0.02084 -0.7087  0.7052
Motivacao -0.56715  0.3509  0.7451

$corr.Y.xscores
                [,1]     [,2]     [,3]
Leitura    -0.3900 -0.06011  0.01408
Escrita    -0.4068  0.01086  0.02647
Matematica -0.3545 -0.04991  0.01537
Ciencia    -0.3056 -0.11337 -0.02395
Sexo       -0.1690  0.12646 -0.05651

$corr.X.yscores
               [,1]     [,2]     [,3]
Controle  -0.419555 -0.06528 -0.01826
Conceito  -0.009673 -0.11872  0.07333
Motivacao -0.263207  0.05878  0.07749

$corr.Y.yscores
             [,1]     [,2]    [,3]
Leitura    -0.8404 -0.35883  0.1354
Escrita    -0.8765  0.06484  0.2546
Matematica -0.7639 -0.29795  0.1478
Ciencia    -0.6584 -0.67680 -0.2304
Sexo       -0.3641  0.75493 -0.5434
```


√ Correlações entre as variáveis psicológicas observadas e as variáveis canônicas:

Variável		*Canonical loadings*					
		U_1	U_2	U_3	V_1	V_2	V_3
X_1	Controle	–0,9041	–0,3897	–0,1756	–0,4192	–0,0653	–0,0183
X_2	Conceito	–0,0208	–0,7087	0,7052	–0,0097	–0,1187	0,0733
X_3	Motivação	–0,5672	0,3509	0,7451	–0,2632	0,0588	0,0775
	PVTE (%)	38,0	25,9	36,1			

√ Correlações entre as variáveis acadêmicas observadas e as variáveis canônicas:

Variável		*Canonical loadings*					
		V_1	V_2	V_3	U_1	U_2	U_3
Y_1	Leitura	–0,8404	–0,3588	0,7047	–0,3900	–0,0601	0,0141
Y_2	Escrita	–0,8765	0,0648	0,0164	–0,4068	0,0109	0,0265
Y_3	Matemática	–0,7639	–0,2980	–0,5104	–0,3545	–0,0499	0,0154
Y_4	Ciência	–0,6584	–0,6768	–0,1307	–0,3056	–0,1134	–0,0240
Y_5	Sexo	–0,3641	0,7549	–0,2304	–0,1690	0,1265	–0,0565
	PVTE (%)	52,5	25,0	9,1			

√ De maneira geral, correlações com as variáveis originais são maiores para o 1º. Par de variáveis canônicas e decrescem para os outros dois pares

Comentários

- Variável canônica é um tipo de variável latente
  - √ Análoga aos fatores obtidos em Análise Fatorial
- Quantidade de variáveis canônicas é igual ao número de variáveis do conjunto menor
  - √ O número de dimensões significantes pode ser menor
- No exemplo
  - √ Há 3 dimensões canônicas, mas apenas as 2 primeiras são estatisticamente significativas
    - O R não efetua diretamente teste de dimensões canônicas

78

- Resultado:

```
> ev <- (1 - calouros.cc1$cor^2)
>
> n <- dim(psic)[1]
> p <- length(psic)
> q <- length(acad)
> k <- min(p, q)
> m <- n - 3/2 - (p + q)/2
>
> w <- rev(cumprod(rev(ev)))
>
> # initialize
> d1 <- d2 <- f <- vector("numeric", k)
>
> for (i in 1:k) {
+     s <- sqrt((p^2 * q^2 - 4)/(p^2 + q^2 - 5))
+     si <- 1/s
+     d1[i] <- p * q
+     d2[i] <- m * s - p * q/2 + 1
+     r <- (1 - w[i]^si)/w[i]^si
+     f[i] <- r * d2[i]/d1[i]
+     p <- p - 1
+     q <- q - 1
+ }
>
> pv <- pf(f, d1, d2, lower.tail = FALSE)
> (dmat <- cbind(WilksL = w, F = f, df1 = d1, df2 = d2, p = pv))
     WilksL        F df1      df2          p
[1,] 0.7543611 11.715733  15 1634.653 7.497594e-28
[2,] 0.9614300  2.944459   8 1186.000 2.905057e-03
[3,] 0.9891858  2.164612   3  594.000 9.109218e-0
```

81

Teste de Dimensionalidade Canônica

- Teste de significância das dimensões canônicas
  - √ Hipótese nula dos testes:
    - Teste 1: Todas as dimensões são significantes?
    - Teste 2: As dimensões 2 e 3 são significantes?
    - Teste 3: A dimensão 3 é significante?

80

- Resultados:
  - √ Teste 1: Todas as dimensões são significantes?
    (F = 11,72)
  - √ Teste 2: As dimensões 2 e 3 são significantes?
    (F = 2,94)
  - √ Teste 3: A dimensão 3 é significante
    (F = 2,165)
- Conclusão:
  - √ As dimensões 1 e 2 são significantes

82

Coeficientes Canônicos Padronizados

- Se variáveis do modelo tem desvios padrão muito diferentes
  - √ Padronização de coeficiente permite comparações mais simples entre variáveis

- Padronização dos coeficientes canônicos:

```
> # Coeficientes Padronizados
>
> # conjunto psic - padronização pela matriz diagonal dos desvios-padrão
> s1 <- diag(sqrt(diag(cov(psic))))
> s1 %*% calouros.ccl$xcoef
        [,1]    [,2]    [,3]
[1,] -0.8404 -0.4166 -0.4435
[2,]  0.2479 -0.8379  0.5833
[3,] -0.4327  0.6948  0.6855
>
> # conjunto acad - padronização pela matriz diagonal dos desvios-padrão
> s2 <- diag(sqrt(diag(cov(acad))))
> s2 %*% calouros.ccl$ycoef
         [,1]     [,2]     [,3]
[1,] -0.45080 -0.04961  0.21601
[2,] -0.34896  0.40921  0.88810
[3,] -0.22047  0.03982  0.08848
[4,] -0.04878 -0.82660 -1.06608
[5,] -0.31504  0.54057 -0.89443
```

- √ Interpretação é análoga
- √ Aumento de 1 desvio-padrão de Leitura diminui 0,45 desvio-padrão no score da 1ª. variável canônica (V_1), mantidas constantes todas as outras variáveis

- Teste de dimensões canônicas

Tabela 1: Testes de Dimensões Canônicas

Dimension	Canonical Corr.	Mult. F	df1	df2	p
1	0.46	11.72	15	1634.7	0.0000
2	0.17	2.94	8	1186	0.0029
3	0.10	2.16	3	594	0.0911

- √ Indica que duas das três dimensões canônicas são estatisticamente significantes a um nível de 5%
- √ A dimensão 1 tem uma correlação canônica de 0,46 entre os dois conjuntos de variáveis
  - – Correlação canônica é 0,17 para a dimensão 2

- Tabela dos coeficientes canônicos padronizados:

Tabela 2: Coeficientes Canônicos Padronizados

	Dimensão 1	2
Variáveis psicológicas		
locus de controle	-0.84	-0.42
autoconceito	0.25	-0.84
motivação	-0.43	0.69
Variáveis Acadêmcias e Gênero		
leitura	-0.45	-0.05
escrita	-0.35	0.41
matemática	-0.22	0.04
ciência	-0.05	-0.83
gênero (feminino=1)	-0.32	0.54

- √ Variáveis psicológicas:
  - – 1ª. dimensão fortemente influenciada por controle (0,84)
  - – 2ª. dimensão: autoconceito (–0,84) e motivação (0,69)
- √ Variáveis acadêmicas e gênero
  - – 1ª. dimensão: leitura (0,45), escrita (0,35) e gênero (0,32)
  - – 2ª. dimensão: escrita (0,41), ciência (0,83) e gênero (0,54)

Precauções com a Análise

- Suposta a normalidade multivariada de ambos os conjuntos de variáveis
- Análise de correlações canônicas não é recomendada para pequenas amostras

88

Exemplo

- (Green, 1973) Solo e vegetação em Belize:
  - √ Estudo dos fatores que influenciaram a locação de lugares maias pré-históricos.
  - √ Amostra: 151 quadrados de 2,5 x 2,5 km, na região de Corozal, em Belize
  - √ Fonte: Manly, B. J. F. Métodos estatísticos multivariados: uma aplicação, 2008.
  - √ Dados: *belize.csv*

91

- Variáveis de solo:
  - √ X_1: % de solo com enriquecimento constante de calcário.
  - √ X_2: % de solo de prado com cálcio na água subterrânea
  - √ X_3: % de solo com matriz de coral sob condições de enriquecimento constante de calcário.
  - √ X_4: % de solos aluvial e orgânico adjacentes a rios e solo orgânico salino na costa.

92

- Variáveis de vegetação:
  - √ Y_1: % de floresta decídua estacional com ervas de folhas largas.
  - √ Y_2: % de floresta de locais altos e baixos coberta com água, plantas herbáceas em lugares úmidos e pântanos.
  - √ Y_3: % de floresta de palma de cohune (palmeira das Honduras).
  - √ Y_4: % de floresta mista.
- Obs.:
  - √ Os valores não somam 100 % para todos os quadrados (não há necessidade de remover variáveis)

93

Objetivo

- Verificar como as variáveis de solo se relacionam com as variáveis de vegetação
  - √ Interessa-se também verificar quantas variáveis canônicas (dimensões) são necessárias para entender a associação entre os dois conjuntos de dados

- Comandos em R

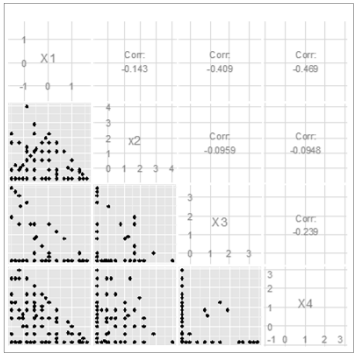
```
> library(ggplot2)
> library(GGally)
> library(CCA)
>
> #  Carregamento de Dados
>
> belize <- read.csv2("belize.csv", head = T, skip = 11)
> head(belize)
  quadrado X1 x2 X3 X4 Y1  Y2 Y3 Y4
1        1 40 30  0 30  0  25  0  0
2        2 20  0  0 10 10  90  0  0
3        3  5  0  0 50 20  50  0  0
4        4 30  0  0 30  0  60  0  0
5        5 40 20  0 20  0  95  0  0
6        6 60  0  0  5  0 100  0  0
> # padronização dos dados
> dados <- scale(belize[,-1])
> solo <- dados[,1:4]
> vegetacao <- dados[,-(1:4)]
```


√ Descritiva básica

```
> summary(belize[,-1])
       X1               x2               X3               X4
 Min.   :  0.00   Min.   : 0.00   Min.   :  0.000   Min.   : 0.00
 1st Qu.: 20.00   1st Qu.: 0.00   1st Qu.:  0.000   1st Qu.: 0.00
 Median : 50.00   Median : 0.00   Median :  0.000   Median :10.00
 Mean   : 47.71   Mean   :11.32   Mean   :  9.854   Mean   :20.15
 3rd Qu.: 75.00   3rd Qu.:20.00   3rd Qu.:  0.000   3rd Qu.:40.00
 Max.   :100.00   Max.   :80.00   Max.   :100.000   Max.   :90.00
       Y1               Y2               Y3               Y4
 Min.   :  0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.000
 1st Qu.:  7.50   1st Qu.: 10.00   1st Qu.: 0.000   1st Qu.: 0.000
 Median : 50.00   Median : 40.00   Median : 0.000   Median : 0.000
 Mean   : 43.31   Mean   : 43.05   Mean   : 1.026   Mean   : 2.351
 3rd Qu.: 75.00   3rd Qu.: 70.00   3rd Qu.: 0.000   3rd Qu.: 0.000
 Max.   :100.00   Max.   :100.00   Max.   :50.000   Max.   :90.000
```


- Conjunto das variáveis de solo:

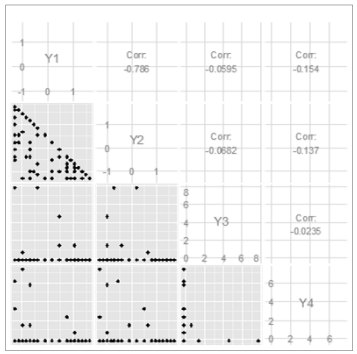
```
> dados <- scale(belize[,-1])
> solo <- dados[,1:4]
> vegetacao <- dados[,-(1:4)]
> ggpairs(solo, axisLabels="internal", upper = list(params = list(size = 4)))
```



- Comportamento da variável X_4 é pouco usual

- Conjunto das variáveis de vegetação:

```
> dados <- scale(belize[,-1])
> solo <- dados[,1:4]
> vegetacao <- dados[,-(1:4)]
> # Scatter matrix plot - variáveis de vegetação
> ggpairs(vegetacao, axisLabels="internal", upper = list(params = list(size = 4)))
```



- Correlações moderadas entre as variáveis de vegetação
  - √ Assumem poucos valores
- Dados claramente são não normais


```
> # matriz de correlações
> belize.cor <- matcor(solo, vegetacao)# library(CCA)
> belize.cor
$Xcor
           X1          x2          X3          X4
X1  1.0000000 -0.14326685 -0.40885658 -0.46922525
x2 -0.1432668  1.00000000 -0.09588203 -0.09483207
X3 -0.4088566 -0.09588203  1.00000000 -0.23873341
X4 -0.4692253 -0.09483207 -0.23873341  1.00000000

$Ycor
           Y1          Y2          Y3          Y4
Y1  1.00000000 -0.78605573 -0.05950747 -0.15400206
Y2 -0.78605573  1.00000000 -0.06817932 -0.13657853
Y3 -0.05950747 -0.06817932  1.00000000 -0.02351585
Y4 -0.15400206 -0.13657853 -0.02351585  1.00000000

$XYcor
           X1          x2          X3          X4          Y1          Y2
X1  1.00000000 -0.14326685 -0.40885658 -0.46922525  0.37826377 -0.26932914
x2 -0.14326685  1.00000000 -0.09588203 -0.09483207 -0.22909283  0.38308412
X3 -0.40885658 -0.09588203  1.00000000 -0.23873341  0.34875268 -0.22379544
X4 -0.46922525 -0.09483207 -0.23873341  1.00000000 -0.39405458  0.34748408
Y1  0.37826377 -0.22909283  0.34875268 -0.39405458  1.00000000 -0.78605573
Y2 -0.26932914  0.38308412 -0.22379544  0.34748408 -0.78605573  1.00000000
Y3 -0.02920585 -0.10448319 -0.01721720  0.20699455 -0.05950747 -0.06817932
Y4  0.14137377 -0.04942136 -0.07477096 -0.01281809 -0.15400206 -0.13657853
           Y3          Y4
X1 -0.02920585  0.14137377
x2 -0.10448319 -0.04942136
X3 -0.01721720 -0.07477096
X4  0.20699455 -0.01281809
Y1 -0.05950747 -0.15400206
Y2 -0.06817932 -0.13657853
Y3  1.00000000 -0.02351585
Y4 -0.02351585  1.00000000
```

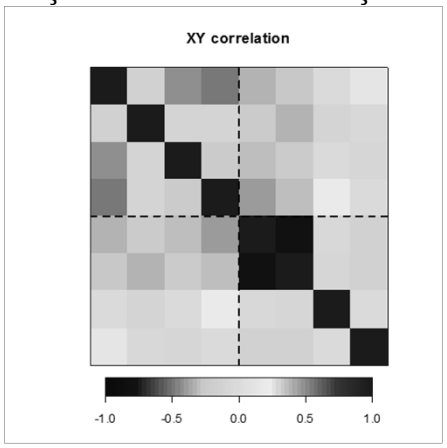
√ Matriz de correlações

R_{XX}

R_{YY}

R

- Visualização da matriz de correlações



- Análise de correlações canônicas:
  - √ Comandos em R:

```
> # Análise de correlação canônica
>
> belize.cc1 <- cc(solo, vegetacao)
```

  - √ Correlações canônicas

```
> # Valores das correlações canônicas
> belize.cc1$cor
[1] 0.7610692 0.5324149 0.2382218 0.1214940
```

√ Coeficientes canônicos:

```
> # Coeficientes canônicos
> belize.cc1[3:4]
$xcoef
           [,1]       [,2]       [,3]        [,4]
X1 -1.3255395 -0.4989341  0.3836752 -0.43805739
x2 -0.2870778 -0.8799652 -0.5703078 -0.01739138
X3 -1.1212504 -0.2916985  0.1427833  0.71860445
X4 -0.5561916 -0.9665043  0.8666097  0.15160040

$ycoef
           [,1]       [,2]       [,3]       [,4]
Y1 -1.6768275 -0.6967539 -0.2208609  0.12032174
Y2 -1.0018424 -1.5000802 -0.3046702  0.01098072
Y3 -0.2156676 -0.3172089  0.9165872  0.26264203
Y4 -0.5081984 -0.3062062  0.2004938 -0.93469975
```

√ Podem ser trocados os sinais em (U_1, V_1) e (U_2, V_2):

– Não altera a correlação canônica

Corr (U_1, V_1) = Corr ($-U_1$, $-V_1$)

- Pares de variáveis canônicas:

$$\begin{cases} U_1 &= 1,33\,Z_1 + 0,29\,Z_2 + 1,12\,Z_3 + 0,56\,Z_4 \\ V_1 &= 1,68\,W_1 + 1,00\,W_2 + 0,22\,W_3 + 0,51\,W_4 \end{cases}$$

$$\begin{cases} U_2 &= 0,50\,Z_1 + 0,88\,Z_2 + 0,29\,Z_3 + 0,97\,Z_4 \\ V_2 &= 0,70\,W_1 + 1,50\,W_2 + 0,32\,W_3 + 0,31\,W_4 \end{cases}$$

$$\begin{cases} U_3 &= 0,38\,Z_1 - 0,57\,Z_2 + 0,14\,Z_3 + 0,87\,Z_4 \\ V_3 &= -0,22\,W_1 - 0,30\,W_2 + 0,92\,W_3 + 0,20\,W_4 \end{cases}$$

$$\begin{cases} U_4 &= -0,44\,Z_1 - 0,01\,Z_2 + 0,72\,Z_3 + 0,15\,Z_4 \\ V_4 &= 0,12\,W_1 + 0,01\,W_2 + 0,26\,W_3 - 0,93\,W_4 \end{cases}$$

√ Z_i e W_j são as variáveis originais padronizadas

$$\begin{cases} Z_i &= \frac{X_i - \bar{X}_i}{s_{X_i}}, \text{ e} \\ W_j &= \frac{Y_j - \bar{Y}_j}{s_{Y_j}}, \; i,j = 1,2,3,4. \end{cases}$$

√ Cargas canônicas:

```
> # Canonical loadings
> belize.cc2 <- comput(solo, vegetacao, belize.cc1)
>
> # display canonical loadings
> belize.cc2[c(3,6)]
$corr.X.xscores
           [,1]       [,2]          [,3]        [,4]
X1 -0.56500103  0.1999069  0.0003682975 -0.80050667
x2  0.06308064 -0.6888603 -0.7211485117 -0.03791011
X3 -0.41898770  0.2274037 -0.1662912582  0.86318254
X4  0.36068909 -0.5793048  0.7065760097  0.18724236

$corr.Y.yscores
           [,1]       [,2]       [,3]       [,4]
Y1 -0.79822609  0.54842546 -0.06679338  0.24000680
Y2  0.40035051 -0.88894455 -0.22093666  0.02615429
Y3 -0.03562821 -0.16627171  0.94578747  0.27671359
Y4 -0.10806174  0.01343353  0.25456389 -0.96090552
```

√ A troca de sinal de (U_1, V_1) e (U_2, V_2) altera o sinal entre as variáveis canônicas e as variáveis originais

Exemplo

√ Correlações entre as variáveis de solo e de vegetação observadas e as variáveis canônicas:

Variável		*Canonical loadings*			
		U_1	U_2	U_3	U_4
X_1	Solo tipo 1	**0,57**	–0,20	0,00	**–0,81**
X_2	Solo tipo 2	–0,06	**0,69**	**–0,72**	–0,04
X_3	Solo tipo 3	**0,42**	–0,23	–0,17	**0,86**
X_4	Solo tipo 4	–0,36	**0,58**	**0,71**	0,19
	PVTE (%)	15,7	22,5	26,2	35,6
Variável		*Canonical loadings*			
		V_1	V_2	V_3	V_4
Y_1	Vegetação tipo 1	**0,80**	**–0,55**	–0,07	0,24
Y_2	Vegetação tipo 2	–0,40	**0,89**	–0,22	0,03
Y_3	Vegetação tipo 3	0,04	0,17	**0,95**	0,28
Y_4	Vegetação tipo 4	0,11	–0,01	0,25	**–0,96**
	PVTE (%)	20,3	28,0	25,3	26,5

Considerando as correlações no intervalo (–0,5; 0,5)

- 1º. Par de variáveis canônicas:

$$\begin{cases} U_1 &= 1,33\,Z_1 + 0,29\,Z_2 + 1,12\,Z_3 + 0,56\,Z_4 \\ V_1 &= 1,68\,W_1 + 1,00\,W_2 + 0,22\,W_3 + 0,51\,W_4 \end{cases}$$

X	U_1	Y	V_1
X_1	**0,57**	Y_1	**0,80**
X_2	–0,06	Y_2	–0,40
X_3	**0,42**	Y_3	0,04
X_4	–0,36	Y_4	0,11

√ U_1: presença de solos tipo 1 (solo com enriquecimento de calcário) e tipo 3 (solo com matriz de coral sob condições de enriquecimento constante).

√ V_1: presença de vegetação 1 (floresta decídua estacional com erros de folhas largas).

106

Técnicas Multivariadas em Saúde - 2015

- 2º. Par de variáveis canônicas:

$$\begin{cases} U_2 &= 0,50\,Z_1 + 0,88\,Z_2 + 0,29\,Z_3 + 0,97\,Z_4 \\ V_2 &= 0,70\,W_1 + 1,50\,W_2 + 0,32\,W_3 + 0,31\,W_4 \end{cases}$$

X	U_2	Y	V_2
X_1	–0,20	Y_1	**–0,55**
X_2	**0,69**	Y_2	**0,89**
X_3	–0,23	Y_3	0,17
X_4	**0,58**	Y_4	–0,01

√ U_2: presença de solos tipo 2 (solo de prado com cálcio na água subterrânea) e tipo 4 (solo aluvial e orgânico adjacente a rios e solo orgânico salino na costa).

√ V_2: presença de vegetação 2 (floresta de locais altos e baixos cobertas com água, plantas herbáceas em lugares úmidos e pântanos) e ausência de vegetação tipo 1.

107

Técnicas Multivariadas em Saúde - 2015

- 3º. Par de variáveis canônicas:

$$\begin{cases} U_3 &= 0,38\,Z_1 - 0,57\,Z_2 + 0,14\,Z_3 + 0,87\,Z_4 \\ V_3 &= -0,22\,W_1 - 0,30\,W_2 + 0,92\,W_3 + 0,20\,W_4 \end{cases}$$

X	U_3	Y	V_3
X_1	0,00	Y_1	–0,07
X_2	**–0,72**	Y_2	–0,22
X_3	–0,17	Y_3	**0,95**
X_4	**0,71**	Y_4	0,25

√ U_3: presença de solos tipo 4 e ausência de solo tipo 2.

√ V_3: presença de vegetação tipo 3 (floresta de palmeiras das Honduras).

108

Técnicas Multivariadas em Saúde - 2015

- 4º. Par de variáveis canônicas:

$$\begin{cases} U_4 &= -0,44\,Z_1 - 0,01\,Z_2 + 0,72\,Z_3 + 0,15\,Z_4 \\ V_4 &= 0,12\,W_1 + 0,01\,W_2 + 0,26\,W_3 - 0,93\,W_4 \end{cases}$$

X	U_4	Y	V_4
X_1	**–0,81**	Y_1	0,24
X_2	–0,04	Y_2	0,03
X_3	**0,86**	Y_3	0,28
X_4	0,19	Y_4	**–0,96**

√ U_4: presença de solos tipo 3 e ausência de solo tipo 1.

√ V_4: ausência de vegetação tipo 4 (floresta mista).

109

Técnicas Multivariadas em Saúde - 2015

- Relações mais importantes entre as variáveis:

  (descritos nos dois primeiros pares)

  √ Par (U_1, V_1):

X	U_1	Y	V_1
X_1	**0,57**	Y_1	**0,80**
X_2	–0,06	Y_2	–0,40
X_3	**0,42**	Y_3	0,04
X_4	–0,36	Y_4	0,11

  – Presença dos solos tipo 1 e 3 e ausência do solo tipo 4 são associados com a presença da vegetação 1.

  √ Par (U_2, V_2):

X	U_2	Y	V_2
X_1	–0,20	Y_1	**–0,55**
X_2	**0,69**	Y_2	**0,89**
X_3	–0,23	Y_3	0,17
X_4	**0,58**	Y_4	–0,01

  – Presença dos solos tipo 2 e 4 é associada com a presença de vegetação tipo 2 e ausência da vegetação tipo 1.

110

Técnicas Multivariadas em Saúde - 2015

Comentários

- Problema potencial não mencionado
  - √ Correlação espacial dos dados em quadrados próximos
  - √ Se correlação espacial existir:
    - Quadrados vizinhos tendem a ter o mesmo solo e vegetação
    - Dados não fornecem 151 observações independentes
    - Dados independentes corresponderão a um conjunto menor
  - √ Efeito de correlação aparecerá em teste de significância global das correlações canônicas
    - Tendência às correlações parecerem ser mais significantes do que realmente são.

113

Técnicas Multivariadas em Saúde - 2015

√ Problema potencial sempre que observações ocorrerem em diferentes lugares no espaço

- Solução:
  - √ Assegurar que observações estejam sufucientemente afastadas umas das outras
  - √ Há métodos que levam em conta a correlação espacial

114

Técnicas Multivariadas em Saúde - 2015

Conclusões

Técnicas Multivariadas em Saúde - 2015

116

Conclusões

- Em geral a análise de correlações canônica é utilizada quando o objetivo do estudo é de natureza exploratória
- Os resultados de análise de correlações canônicas têm a reputação de serem difíceis de serem interpretados
  - √ As variáveis devem ser conhecidas muito bem para se conseguir extrair explicações convincentes dos resultados

- A análise de correlações canônicas pode oferecer uma descrição útil e maior compreensão sobre a associação entre os dois conjuntos de variáveis.

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.