Análise Multivariada 1 - 01/2023
Prof. George von Borries
Listas de Exercícios

Importante: o aluno pode utilizar o `R`, `SAS` sempre que necessário e consultar o professor antes de utilizar outras linguagens de programação. A solução dos exercícios deve estar organizada e pronta para ser entregue quando solicitados. O aluno deve ser claro na resposta e incluir código de programação com explicações sempre que fizer uso do mesmo. As entregas serão sempre via Aprender.

Lista 1: Introdução

Data de Divulgação: 10/04/2023

1. **Exercício de Leitura:**

   Ler os seguintes artigos:

   - Lozano, T. (2016) *Data Science: an Emerging Discipline*, CITAS.
   - Kiliç, G.G.; Turanli, M.; Özden, Ü.H. (2019) *Data Mining and Statistics in Data Science*, International Social Science Studies Journal.

     Os artigos acima são encontrados em

     https://www.semanticscholar.org/paper/DATA-MINING-AND-STATISTICS%3A-WHAT"S-THE-CONNECTION-Friedman/6b3cb8325ca437b6b005027e2e2c64adb99609c8
   - Zola, A. (2019) *Machine Learning Engineer vs. Data Scientist*. Disponível em `https://www.springboard.com/blog/data-science/machine-learning-engineer-vs-data-scientist/`

   Fazer um resumo de três parágrafos e não mais de uma página indicando semelhanças e diferenças entre Estatística, Mineração de Dados e Ciência de Dados. Você pode incluir referência a textos mais recentes sobre estes temas.

2. Escolha uma área de pesquisa de interesse. Pesquise artigos publicados em revista indexadas e descreva (resumidamente) um exemplo incando o tipo de problema (ou problemas) entre os listados abaixo. Inclua referência bibliográfica e indique as características dos dados e estudo que relacionam ao tipo de problema (ou problemas) indicado.
   (a) Análise multivariada clássica ($n < p$).
   (b) Mineração de dados (*Data Mining*) (n elevado).
   (c) Aprendizado estatístico (*Statistical Learning*) (p elevado).
   (d) Reconhecimento de padrões.
   (e) Data Science.

3. Descreva cada um dos problemas encontrados na análise multivariada, encontre um exemplo de caso real e indique uma falha nas técnicas estatísticas tradicionais.
   (a) *Mining* (mineração, n muito elevado).
   (b) *Scalability* (escalabilidade).
   (c) *High Dimensional Data* (dados em alta dimensão, $n > p$).

(d) *Pequenas amostras.*

(e) *Curse of Dimensionality* (Problema de dimensionalidade).

4. Johnson e Wichern - Exercício 1.2
5. Johnson e Wichern - Exercício 1.6
6. Johnson e Wichern - Exercício 1.14
7. Johnson e Wichern - Exercício 1.22

Lista 2: Álgebra de Matrizes e Vetores Aleatórios

Data de Divulgação: 14/04/2023

8. Provar o seguinte teorema:

   Sejam $\mathbf{A}$ e $\mathbf{B}$ matrizes idempotentes. Então,

   (a) $\mathbf{A} + \mathbf{B}$ é idempotente somente quando $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$.

   (b) $\mathbf{C} = \mathbf{AB}$ é idempotente somente quando $\mathbf{AB} = \mathbf{BA}$.

   (c) $\mathbf{I} - \mathbf{A}$ é idempotente.

9. Provar o seguinte teorema:

   Seja $\mathbf{X}_{(n\times k)}$ tal que $\text{rank}(\mathbf{X}) = k < n$.
   Então, $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é idempotente e simétrica e consequentemente, uma matriz projeção ortogonal.

10. Utilizando o `R` verifique, através de exemplos, que uma matriz de projeção tem autovalores somente no conjunto $\{0, 1\}$. A demonstração pode ser feita utilizando a equação característica e lembrando que se $\mathbf{M}$ é uma matriz de projeção, então $\mathbf{M} = \mathbf{M}^2 = \mathbf{M}^{\mathrm{T}}$.

11. Seja $\mathbf{X}$ uma matriz de dados $(n \times p)$ com matriz de covariância $\mathbf{S}$. Sejam $\lambda_1, \ldots, \lambda_p$ os autovalores de $\mathbf{S}$.

    (a) Mostre que a soma das variâncias s_{ii} de $\mathbf{X}$ (variação amostral total) é dada por $\lambda_1 + \ldots + \lambda_p$.

    (b) Mostre que a variância amostral generalizada é dada por $\lambda_1 \times \ldots \times \lambda_p$.

    (c) Mostre que a variância amostral generalizada se anula se as colunas de $\mathbf{X}$ somarem zero.

12. Seja $\mathbf{A}$ uma matriz quadrada $(k \times k)$ positiva definida. Mostre que,

    (a) $(\mathbf{A}^{1/2})' = \mathbf{A}^{1/2}$.

    (b) $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$.

    (c) $(\mathbf{A}^{1/2})^{-1} = \mathbf{C}\mathbf{D}^{-1/2}\mathbf{C}'$, sendo $\mathbf{D}^{-1/2} = \text{diag}(1/\sqrt{\lambda_1}, \ldots, 1/\sqrt{\lambda_k})$.

(d) $\mathbf{A}^{1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1/2}\mathbf{A}^{1/2} = \mathbf{I}$.

(e) $\mathbf{A}^{-1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1}$.

13. Considere uma matriz de correlação $(r \times r)$ com a mesma correlação (ρ) em todas as células fora da diagonal. Encontre os autovalores e autovetores desta matriz quando $r = 2, 3, 4$. Generalize seus resultados para qualquer número r de variáveis. Como exemplo, faça $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$.

14. Considere a decomposição espectral de uma matrix $\mathbf{A}_{p\times p}$ positiva definida, isto é, $\mathbf{A}_{p\times p} = \mathbf{B}\,\mathbf{\Delta}\,\mathbf{B}^{\mathrm{T}}$. Seja

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & 3 & 2 \\ 2 & 5 & 1 & 1 \\ 3 & 1 & 8 & 2 \\ 2 & 1 & 2 & 3 \end{bmatrix}$$

(a) Obtenha $\mathbf{B}$ e $\mathbf{\Delta}$.

(b) Obtenha $\mathbf{A}^{1/2}$ e mostre que $(\mathbf{A}^{1/2})^2 = \mathbf{A}$.

(c) Obtenha $(\mathbf{A}^{1/2})^{-1}$ (descreva seus elementos) e mostre que $(\mathbf{A}^{1/2})^{-1}\mathbf{A}^{1/2} = \mathbf{I}$.

15. Johnson e Wichern - Exercício 2.1

16. Johnson e Wichern - Exercício 2.2

17. Johnson e Wichern - Exercício 2.6

18. Johnson e Wichern - Exercício 2.7

19. Johnson e Wichern - Exercício 2.8

20. Johnson e Wichern - Exercício 2.16

21. Johnson e Wichern - Exercício 2.18

22. Johnson e Wichern - Exercício 2.19

23. Johnson e Wichern - Exercício 2.21

24. Johnson e Wichern - Exercício 2.25

    Nota: $\boldsymbol{\rho}$ é a matriz de correlação populacional e $\mathbf{V}$ é a matriz diagonal de variâncias.

    Para calcular $\boldsymbol{\rho}$ é mais fácil fazer $\boldsymbol{\rho} = (\mathbf{V}^{1/2})^{-1}\mathbf{\Sigma}(\mathbf{V}^{1/2})^{-1}$, uma vez que $\mathbf{V}$ é uma matriz diagonal, i.e., $\mathbf{V} = \text{diag}\left[\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{11}}\right]$.

25. Johnson e Wichern - Exercício 2.27

26. Johnson e Wichern - Exercício 2.32

27. Johnson e Wichern - Exercício 2.41

28. Considere o seguinte conjunto de dados de Pacientes em Tratamento de Hemodiálise.
    (a) Represente graficamente e através de medidas descritivas.
    (b) Obtenha a decomposição espectral e verifique se existe indicação de uma possível redução da dimensão do estudo em questão. Justifique.

Idade	Proteína	Energia	Albumina	IMC
32	1.59	2738.86	4.2	24.1
61	0.49	824.26	3.9	29.8
51	1.14	1307.03	4.1	20.0
53	0.74	925.47	4.2	25.0
24	1.99	2787.46	3.8	21.5
65	1.00	1222.51	4.2	25.0
35	2.32	2038.28	4.1	18.7
45	0.93	1061.53	4.2	22.0
57	0.81	1657.73	4.2	31.2
32	1.23	1652.76	3.9	24.3
66	0.99	1636.25	4.1	27.7
27	1.40	1845.07	4.0	21.8
54	1.08	1542.30	3.9	29.0
55	1.22	1214.53	4.0	21.1
50	0.57	1451.17	4.0	27.1
48	0.83	1786.95	4.1	24.7
28	1.55	1975.26	3.5	18.8
66	1.10	1248.64	4.0	18.9
66	0.44	987.86	4.0	27.6
48	0.58	1067.10	4.3	26.4
60	0.43	968.62	4.0	35.9
59	0.66	836.94	3.9	25.3
50	1.81	1197.99	3.9	19.5
29	1.21	1818.31	4.2	21.8
40	0.98	1238.91	3.5	21.9
47	1.48	2153.47	3.5	17.3
52	0.98	1720.60	3.6	29.7
54	1.02	1906.30	4.5	31.9
53	0.82	981.85	3.9	26.2
47	0.46	1020.95	4.4	31.2
42	1.34	1028.10	3.6	18.1
79	1.48	1465.91	3.9	18.3
61	1.39	1456.12	3.9	24.9

Lista 3: Gráficos e Decomposição em Valores Singulares

Data de Divulgação: 18/04/2023

29. Encontrar um artigo com alguma representação gráfica multivariada. Discutir sobre a utilidade do gráfico e possível alternativa. Apresentar um exemplo do mesmo gráfico utilizando o `R`.

    Nota 1: O exemplo apresentado pode ser com os próprios dados do artigo, i.e., reproduzindo o gráfico apresentado no artigo.

    Nota 2: No dia da Prova 1 (data provável 09/05/2023) o estudante deverá entregar, via Moodle, uma cópia do artigo, até dois parágrafos com os comentários e o código `R` completo com dados e gráfico reproduzido. Esta atividade contará como parte da nota da Prova 1.

30. Utilize a decomposição espectral $\mathbf{\Sigma} = \mathbf{UDV}^{\mathrm{T}}$ para mostrar que $\sum_{i=1}^{p} \mathrm{Var}(\mathbf{X}_i) = \sum_{i=1}^{p} \lambda_i$, em que λ_i são os elementos (autovalores) da matriz diagonal $\mathbf{D}$.

31. Reproduza o estudo de redução de dimensão SVD de imagens, utilizando duas imagens. A primeira com poucos detalhes (abstrata, por exemplo) e a segunda com vários detalhes. Justifique sua escolha de dimensão na redução de cada imagem e compare os resultados. Você achou a redução compatível com as imagens utilizadas?

Lista 4: Análise de Componentes Principais

Data de Divulgação: 25/04/2023

32. Suponha que um pesquisador padronizou os dados de um estudo através da transformação de Mahalanobis ($\mathbf{Z} = \mathbf{XS}^{-1/2}$), em que $\mathbf{S}$ é a matriz de variância-covariâncias amostrais. Seria razoável aplicar componentes principais nos dados transformados? Justifique sua resposta.

33. Johnson e Wichern - Exercício 8.03.

34. Johnson e Wichern - Exercício 8.06.

35. Johnson e Wichern - Exercício 8.07.

36. Johnson e Wichern - Exercício 8.11. - Dados no arquivo Census-tract (T8-5.DAT).

    $\mathbf{X}_1$: Total Population (thousands); $\mathbf{X}_2$: Median School Years;

    $\mathbf{X}_3$: Total Employment (thousands); $\mathbf{X}_4$: Health Services Employment (hundreds);

    $\mathbf{X}_5$: Median Value Home ($10,000s).

37. **Johnson e Wichern - Exercício 8.12.** Dados no arquivo Air Pollution (T1-5.DAT). Os dados correspondem a 42 medidas de poluição do ar observadas na área de Los Angeles em um mesmo horário.

    X_1: Wind; $\mathbf{X}_2$: Solar Radiation; $\mathbf{X}_3$: CO; $\mathbf{X}_4$: NO;
    X_5: NO_2; $\mathbf{X}_6$: O_3; $\mathbf{X}_7$: HC.

38. Um estudo sobre dados nutricionais (Izenman, 2008) analisou 961 alimentos. Os componentes nutricionais observados foram: gordura (gramas), energia (calorias), carbohidratos (gramas), proteínas (gramas), colesterol (miligramas), peso (gramas) e gordura saturada (gramas). Os alimentos são listados em porções variadas e por isso as porções são divididas pelo peso correspondente de cada item.

    A matriz de variância dos dados é (colunas referentes a gordura, calorias, carbohidratos, proteínas, colesterol e gordura saturada),

$$\mathbf{S} = \begin{pmatrix} 0.037 & 0.306 & -0.007 & 0.003 & 0.023 & 0.010 \\ 0.306 & 3.747 & 0.157 & 0.045 & 0.178 & 0.082 \\ -0.007 & 0.157 & 0.062 & -0.002 & -0.027 & -0.002 \\ 0.003 & 0.045 & -0.002 & 0.008 & 0.020 & 0.001 \\ 0.023 & 0.178 & -0.027 & 0.020 & 0.456 & 0.014 \\ 0.010 & 0.082 & -0.002 & 0.001 & 0.014 & 0.004 \end{pmatrix}$$

    com autovalores $\lambda_1^v = 3.790$, $\lambda_2^v = 0.451$, $\lambda_3^v = 0.061$, $\lambda_4^v = 0.008$, $\lambda_5^v = 0.004$, $\lambda_6^v = 0.002$ e respectivos autovetores (v_i),

Item	v_1	v_2	v_3	v_4	v_5	v_6
Gord	-0.081	-0.019	0.364	-0.506	0.759	-0.167
Calo	-0.994	0.052	0.011	0.044	-0.083	-0.010
Carb	-0.041	0.092	-0.921	-0.186	0.325	0.037
Prot	-0.012	-0.040	0.028	0.827	0.554	0.080
Cole	-0.053	-0.993	-0.095	-0.035	-0.012	-0.019
GorS	-0.022	-0.022	0.093	-0.147	0.071	0.982

    A matriz de correlação,

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.816 & -0.140 & 0.157 & 0.175 & 0.747 \\ 0.816 & 1.000 & 0.324 & 0.261 & 0.136 & 0.643 \\ -0.140 & 0.324 & 1.000 & -0.087 & -0.163 & -0.141 \\ 0.157 & 0.261 & -0.087 & 1.000 & 0.328 & 0.142 \\ 0.175 & 0.136 & -0.163 & 0.328 & 1.000 & 0.311 \\ 0.747 & 0.643 & -0.141 & 0.142 & 0.311 & 1.000 \end{pmatrix}$$

    com autovalores $\lambda_1^c = 2.649$, $\lambda_2^c = 1.330$, $\lambda_3^c = 1.020$, $\lambda_4^c = 0.680$, $\lambda_5^c = 0.267$, $\lambda_6^c = 0.055$ e respectivos autovetores (c_i),

Item	c_1	c_2	c_3	c_4	c_5	c_6
Gord	-0.557	0.099	0.275	0.130	-0.455	0.617
Calo	-0.536	0.357	-0.137	0.075	-0.273	-0.697
Carb	0.025	0.672	-0.568	-0.286	0.157	0.344
Prot	-0.235	-0.374	-0.639	0.599	0.154	0.119
Cole	-0.253	-0.521	-0.326	-0.717	-0.210	-0.003
GorS	-0.531	-0.019	0.261	-0.150	0.791	0.022

Ainda são fornecidas as seguintes medidas:

- Variância amostral generalizada: VAG $= 5.011 \times 10^{-9}$
- Variância amostral total: VAT $= 4.314$

Utilizando os resultados apresentados, responda as seguintes questões:

(a) Sobre a matriz $\mathbf{S}$:
   1. Indique se esta matriz é positiva definida. Jusfique.
   2. Indique como você faria para obter a decomposição espectral da matriz $\mathbf{S}$?
   3. Indique um procedimento para obter $\mathbf{S}^{-1}$ e $\sqrt{\mathbf{S}}$ (se existir).

(b) Indique como obter a variância generalizada e interprete o resultado.

(c) Qual o motivo da grande diferença entre VAG e VAT? Isto fato é refletido em alguma outra medida apresentada para $\mathbf{S}$?

(d) Qual matriz, variância-covariância ou correlações você usaria para reduzir a dimensionalidade dos dados? Justifique.

(e) Com base em sua escolha para o item (d), obtenha as duas primeiras componentes principais para os dados e faça o gráfico *Scree Plot*.

(f) Quantas componentes principais você reteria no estudo? Justifique. Interprete (se possível) as duas primeiras componentes principais.

39. O pacote `kohonen` do `R` possui dados sobre análise química de 177 vinhos de três tipos de cultivares de uvas (Nebbiolo, Barberas e Grignolino) da região de Piemonte na Itália. O vinho da uva Nebbiolo é chamado de Barolo. A identificação dos vinhos está no objeto `vintages`. O artigo de origem destes dados é

    Forina, M.; Armanino, C.; Castino, M. e Ubigli, M. (1986) Multivariate data analysis as a discriminating method of the origin of wines. *Vitis*, 189-201. O arquivo pode ser obtido com os seguintes comandos `R`:

```
install.packages("kohonen")
library(kohonen)
data("wines")
```

Desconsidere a variável `OD ratio` e faça uma análise de componentes principais deste conjunto de dados.

Lista 5: Análise de Correspondências

Data de Divulgação: 05/05/2023

40. Pesquise sobre o pacote `ca` do `R`. Descreva os objetos da função `ca` (listados abaixo) e a relação com os resultados matriciais apresentados em sala. Reproduza estes objetos matricialmente.

    Objetos da função `ca`: sv, rowmass, rowdist, rowinertia, rowcoord, colmass, coldist, colinertia, colcoord, N.

41. O pacote `vcd` do `R` possui dados sobre sobre suicídio na Alemanha Ocidental obtidos de Heuer, J., 1979 (*Suicide by children and youth*). O arquivo possui 306 observações e 6 variáveis:

    - *Freq*: número de casos.
    - *sex*: sexo masculino (male) e feminino (female).
    - *method*: método de suicídio - poison (envenenamento), toxic gas (gás tóxico), cooking gas (gás de cozinha), hang (enforcamento), drown (afogamento), gun (arma de fogo), knife (faca), jump (pulo) e other (outros).
    - *age*: idade.
    - *age.group*: faixa etária.
    - *method2*: método de suícidio. Igual a *method*, mas com as categorias *toxic gas* e *cooking gas* unidas como *gas* (gás). Observação: neste caso você tem uma tabela com três variáveis, i.e., (sexo, método, grupo de idade).

    Realize uma análise de correspodências com as variáveis *method* e *age.group*.

    (a) Indique cada etapa de seu estudo e respectivas conclusões.

    (b) Calcule e interprete a estatística Qui-Quadrado e a Inércia Total.

    (c) Análise os resíduos em cada célula da tabela e interprete os resultados.

    (d) Apreente um mapa de correspondências e faça um breve resumo dos resultados gerais da análise.

42. Apresenta uma aplicação para análise de correspondências, indicando o objetivo do estudo e fonte de dados. Se possível, colete os dados e apresente a Tabela de Contigência. Observação: não precisa realizar a análise.

Lista 6: Análise de Fatorial Exploratória

Data de Divulgação: 25/05/2023

43. **Johnson e Wichern - Exercício 9.1.**
44. **Johnson e Wichern - Exercício 9.2.**
45. **Johnson e Wichern - Exercício 9.3.**
46. **Johnson e Wichern - Exercício 9.6.**
47. Reproduza dos resultados do **Exemplo 9.14 de Johnson e Wichern**.
48. **Johnson e Wichern - Exercício 9.10.**
49. **Johnson e Wichern - Exercício 9.12.**
50. **Johnson e Wichern - Exercícios 9.25.**
51. **Johnson e Wichern - Exercício 9.19.** - arquivo `table9_12-SalespeopleData.DAT` disponível no site do curso.
52. **Johnson e Wichern - Exercício 9.21.**
53. **Johnson e Wichern - Exercício 9.22.** - pesquise se o `R` tem os comandos necessários e aplique para obter a solução.
54. **Johnson e Wichern - Exercício 9.23.**

Lista 7: Normal Multivariada

Data de Divulgação: 13/06/2023

55. Assuma $\mathbf{Y} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I})$. Descreva e implemente um algoritmo em R para obter $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tal que

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & -1.5 \\ -1.5 & 4 \end{bmatrix}$$

56. Mostre que se $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, então

$$\mathrm{U} = (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p.$$

Este é o Resultado 5 das notas de aula.

57. Seja $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tal que

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 2 & a \\ a & 2 \end{bmatrix}$$

Represente a elipse para $a = 0; -1/2; 1/2$ e 1.

58. Considere $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tal que

$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

e os vetores $\mathbf{A} = [1 \;\; 1]$ e $\mathbf{B} = [1 \;\; -1]$. Mostre que $\mathbf{AX}$ é independente de $\mathbf{BX}$.

59. Seja $\mathbf{X} = \begin{bmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \end{bmatrix} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$.

Prove que $\Sigma_{12} = 0$ se, e somente se X_1 é independente de X_2.

60. **Johnson e Wichern - Exercício 4.26.**
61. **Johnson e Wichern - Exercício 4.27.**
62. **Johnson e Wichern - Exercício 4.35.**
63. **Rencher e Christensen - Problema 4.1**
64. **Rencher e Christensen - Problema 4.2**
65. **Rencher e Christensen - Problema 4.10**
66. **Rencher e Christensen - Problema 4.11**
67. **Rencher e Christensen - Problema 4.12**
68. **Rencher e Christensen - Problema 4.13**
69. **Rencher e Christensen - Problema 4.14**
70. **Rencher e Christensen - Problema 4.17**

Lista 8: Correlação Canônica

Data de Divulgação: 20/06/2023

71. **Johnson e Wichern - Exercício 10.1**
72. **Johnson e Wichern - Exercício 10.2**

73. **Johnson e Wichern - Exercício 10.9** - Itens (a) e (c)

74. **Johnson e Wichern - Exercício 10.10**

75. **Rencher e Christensen - Problema 11.8. (Diabetes Data)**

    Diabetes Data: $y_1 =$ *relative weight*; $y_2 =$ *fasting plasma glucose*; $x_1 =$ *glucose intolerance*; $x_2 =$ *insulin response to oral glucose*; $x_3 =$ *insulin resistance*.

    (a) Encontre as correlações canônicas entre (y_1, y_2) e (x_1, x_2, x_3).

    (b) Encontre os coeficientes padronizados das variáveis canônicas.

Lista 9: Análise de Agrupamentos

Data de Divulgação: 05/07/2023

76. **Johnson e Wichern - Exercício 12.3.**

77. **Johnson e Wichern - Exercício 12.5.**

78. **Johnson e Wichern - Exercício 12.6.**

79. **Johnson e Wichern - Exercício 12.7.**

    Importante: para os exercícios 12.11, 12.12 e 12.13 seguintes, é preciso fornecer o centro de cada grupo incial no comando `kmeans` do `R` em vez do número de grupos. Exemplo: se queremos dois grupos e o centro inicial são as linhas 1 e 4 da matrix x, então definimos `kc = data.frame(x[c(1,4),])` e indicamos este centro em `kmeans(x,centers=kc)`.

80. **Johnson e Wichern - Exercício 12.11.**

81. **Johnson e Wichern - Exercício 12.12.**

82. **Johnson e Wichern - Exercício 12.13.**

83. Considere o seguinte conjunto de 22 pontos (x, y).

Ponto	X	Y	Ponto	X	Y
1	1	9	2	2	10
3	2	9	4	2	8
5	3	9	6	7	14
7	12	9	8	13	10
9	13	8	10	14	10
11	14	8	12	15	9
13	7	7	14	6	3
15	7	3	16	8	3
17	6	2	18	7	2
19	8	2	20	6	1
21	7	1	22	8	1

(a) Plote o gráfico de dispersão para o conjunto de pontos. Quantos e quais grupos você indicaria através do gráfico.

(b) Construa a matriz de distâncias **D**, calculando as distâncias Euclidiana, de Manhattan e Mahalanobis. Compare os resultados. Explique a razão de possíveis diferenças nos resultados para as três distâncias.

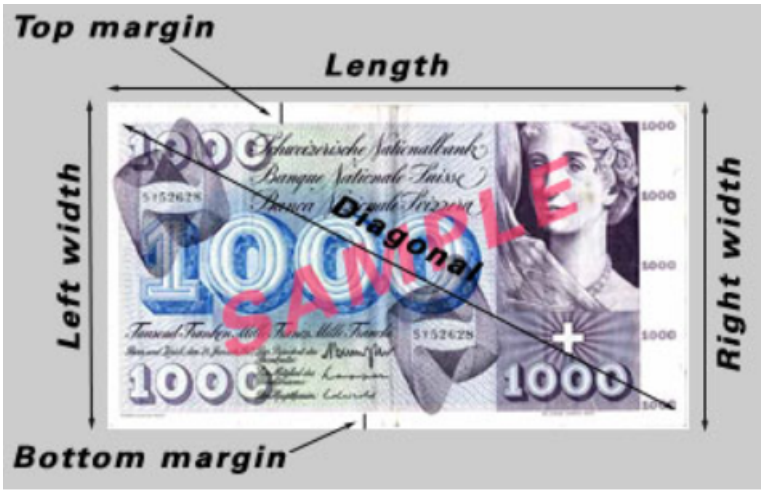
(c) Utilize os algoritmos de ligação simples e média, e obtenha os respectivos dendogramas. Em cada caso indique sua escolha para o número de grupos e liste os elementos de cada grupo. Os dendogramas obtidos são únicos? Justifique.

(d) Descreva e aplique o algoritmo de k-means para agrupar estes dados. Apresente os resultados e discuta.

84. Seis variáveis são medidas de 100 notas genuínas e 100 notas falsificadas (Flury and Riedwyl, 1988[1].

As colunas correspondem as seguintes variáveis:

1. Comprimento da nota (*Length*).
2. Altura da nota no lado esquerdo (*Left*).
3. Altura da nota no lado direito (*Righ*).
4. Distância da moldura interna a borda superior (*Bottom*).
4. Distância da moldura interna a borda inferior (*Top*).
6. Comprimento da diagonal (*Diagonal*).



As 100 primeiras notas são genuínas e as 100 últimas são falsificadas.

Importante: O conjunto de dados pode ser obtido através do comando `data(bank)` com o pacote `gclus`.

[1]Referência: Flury, B. and Riedwyl, H. (1988). Multivariate Statistics, A Practical Approach, Cambridge University Press.

(a) Calcule as médias de cada variável em cada população (notas verdadeiras e falsificadas). Obtenha também a matriz de variância-covariâncias Σ em cada população. Obtenha um *face plot* de cada população utilizando os valores médios. Compare os resultados númericos com os gráficos. Discuta.

Utilize o *face plot* para comparar os valores médios das variáveis para notas verdadeiras e notas falsas. Discuta os resultados.

(b) Utilize um método de agrupamento para verificar se é possível distiguir as notas falsas das verdadeiras, imaginando que você não soubesse se 100 últimas notas eram verdadeiras.

(c) Compare os resultados do agrupamento anterior com o agrupamento por mistura de normais.

(d) Utilize o Índice Ajustado de Rand para comparar seu agrupamento das 200 notas com a verdadeira classificação em genuínas e falsas. Avalie a qualidade do agrupamento.

Lista 10: Análise de Discriminantes e Classificação

Data de Divulgação: 11/07/2023

85. **Johnson e Wichern - Exercício 11.1.**

86. **Johnson e Wichern - Exercício 11.2.**

87. **Johnson e Wichern - Exercício 11.4.**

88. **Johnson e Wichern - Exercício 11.10.**

89. **Johnson e Wichern - Exercício 11.24.** - Dados: `T11-4-BankruptcyData.dat`.

90. **Johnson e Wichern - Exercício 11.32.** - Dados: `T11-8-Hemofilia.dat`.

91. Considere os dados sobre falsificação de notas do Exercício 84. Realize as análises Linear e Quadrática de discriminantes. É possível utilizar mistura de normais (função `mclust`) como análise de discriminantes? Se sim, faça o estudo e compare as estimativas `APER` para as três situações. Resuma os resultados e conclusões.