ME 731 - Métodos em Análise Multivariada
Segundo semestre de 2021
Lista de Exercícios VII

1. Resolva todos os exercícios deixados em sala de aula

2. Na análise de correlações canônicas considere um vetor aleatório $\boldsymbol{X}_{(p+q)}, p \leq q$, com as devidas partições do vetor de médias $\boldsymbol{\mu} = \mathcal{E}(\boldsymbol{X})$, da matriz de covariâncias $\boldsymbol{\Sigma} = Cov(\boldsymbol{X})$ e da matriz de correlações $\boldsymbol{\rho} = Corre(\boldsymbol{X})$. Considere o vetor aleatório padronizado $\boldsymbol{Z}$ com as partições correspondentes. Sejam $\boldsymbol{U}^{(\boldsymbol{Z})}_{(p\times 1)} = (U_1^{(\boldsymbol{Z})}, ..., U_p^{(\boldsymbol{Z})})'$ as $\boldsymbol{V}^{(\boldsymbol{Z})}_{(q\times 1)} = (V_1^{(\boldsymbol{Z})}, ..., V_q^{(\boldsymbol{Z})})'$ as variáveis canônicas obtidas a partir do vetor $\boldsymbol{Z}$. Responda os itens abaixo:

   a) Prove todos os resultados deixados como exercícios relativos a este assunto.
   b) Suponha que $\boldsymbol{X}_{(p+q)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Qual a distribuição conjunta do vetor $(\boldsymbol{U}^{(\boldsymbol{Z})}, \boldsymbol{V}^{(\boldsymbol{Z})})$? As componentes de $\boldsymbol{U}^{(\boldsymbol{Z})}$ são mutuamente independentes? O são também as componentes de $\boldsymbol{V}^{(\boldsymbol{Z})}$? E em relação aos vetores $\boldsymbol{U}^{(\boldsymbol{Z})}$ e $\boldsymbol{V}^{(\boldsymbol{Z})}$.

3. Considere os dados sobre "bitting fly" (Lista IV, Questão 10), com os dois seguintes grupos de variáveis: grupo 1 - comprimento da asa, largura da asa, grupo 2 - comprimento do terceiro palpo, largura do terceiro palpo, comprimento do quarto palpo, comprimento do 12^o segmento da antena e comprimento do 13^o segmento da antena. Realize uma análise de correlação canônica, à semelhança do que foi visto em sala com o intuito de: reduzir a dimensionalidade, estudar a estrutura de dependência entre as variáveis intra e entre grupos, identificar padrõe entre as variáveis e entre as unidades experimentais (e se existir, entre grupos definidos por variáveis categorizadas presentes no banco de dados) e para a utilização das variáveis canônicas para futuras análises (uni e multivariadas). Apresente comentários e conclusões pertinentes, verificação de normalidade (multivariada) etc.

4. Repita a Questão anterior para os dados da Questão 12 da Lista II. Considere os dois seguintes grupos de variáveis: grupo 1 - hurdles, run200m, run800m e grupo 2 - highjump, shot, longjump, javelin.

5. Repita a Questão 4 para os dados da Questão 8 da Lista 5. Considere os dois seguintes grupos de variáveis: grupo 1 - cut, color, clarity e grupo 2 - carat, x, y, z, depth e table. Com relação às variáveis do grupo 1, atribua o valor 1 para a pior categoria, o valor 2 para a segunda pior categoria, e assim por diante.

6. Os dados disponíveis em:

   bioData <– read.csv("http://msekce.karlin.mff.cuni.cz/~maciak/NMST539/bioData.csv", header = T)

   chemData <– read.csv("http://msekce.karlin.mff.cuni.cz/~maciak/NMST539/chemData.csv", header = T)

   correspondem, respectivamente, a características biológicas e químicas de vários rios na repúblcia Tcheca. Repita a Questão 4, considerando os dois grupos de variáveis acima.

7. Considere os dados disponíveis no pacote "heplots" sob o nome de "Rohwer", referentes a um experimento conduzido por William D. Rohwer em crianças do jardim de infância elaborado para examinar o quão bom o desempenho em um conjunto de tarefas "associadas aos pares" (PA) pode prever o desempenho em algumas medidas de aptidão e desempenho. Considere os seguinte grupos de variáveis: grupo 1 - SAT, PPVT, Raven; grupo 2 - n, s, ns, na, ss. Realize uma ACC mais completa possível. Além disso, ajuste (um) modelo(s) de regressão univariado(s), entre o(s) par(es) de variável(is) canônicas, para avaliar como as variáveis do grupo 2 afetam as do grupo 1.

8. Repita a Questão anterior, considerando os grupos definidos pela variável SES.

9. Repita as Questões 7 e 8 considerando as variáveis originais, ao invés das variáveis canônicas. Compare os resultados obtidos aqui com aqueles obtivos nas Questões mencionadas.

10. Para os dados da iris de Fisher, considere a ACC apresentada aqui. Ajuste um modelo de regressão univariado apropriado (com o primeiro par de variáveis canônicas), para avaliar como as variáveis do segundo grupo afetam as do primeiro.

11. Repita a Questão anterior considerando os tipo de iris.

12. Repita as Questões 10 e 11 considerando as variáveis originais, ao invés das variáveis canônicas. Compare os resultados obtidos aqui com aqueles obtivos nas Questões mencionadas.

13. Considere a metodologia de análise de discriminante apresentada em classe. Considere dois vetores aleatórios independentes $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$, tais que $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, i=1,2. Considere $\boldsymbol{x}$ um valor observado de algum desses dois vetores. Responda os itens:

    a) Prove que, se $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ então, $-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$.

b) Seja $y_i = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ e $m = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. Prove que $m = \frac{1}{2}(y_1 + y_2)$

c) Prove que, se $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ então, $-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) = -\frac{1}{2}\boldsymbol{x}'\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{x} + \left(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{x} - c$, em que
$c = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}\left(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right)$

14. Considere os dados sobre salmão ("dados salmon"), disponíveis no site do curso. Considere, nesta questão, a variável procedência como a definidora de grupo, enquanto que diâmetro das guelras durante a fase em água doce (em mm) e diâmetro das guelras durante a fase no mar (em mm), comom variáveis resposta. O objetivo é criar uma regra de classificação adequada conforme visto em sala. Além disso, faça uma análise descritiva apropriada. Apresente comentários e conclusões pertinentes.

15. Repita a questão anterior, considerando o sexo como variável definidor de grupo.

16. Repita a questão anterior, considerando tanto o sexo como procedência como variáveis definidoras de grupo.

17. O problema a ser resolvido diz respeito a características relativas à saúde de n = 768 mulheres indianas, portadoras de uma herança genética chamada prima (disponíveis no pacote " mlbench" sob o nome "PimaIndiansDiabetes2", elimine todas as linhas com pelo menos um NA). Devido a informações omissas, em relação a pela menos uma das variáveis, tem-se apenas n = 392. Foram medidas as seguintes variáveis (entre parênteses estão os nomes sugeridos para nos reportarmos a elas):

- Número de vezes que engravidou (gravidez).
- Concentração de glicose plasmática a 2 horas em um teste de tolerância à glicose oral (glicose).
- Pressão arterial diastólica (em mm Hg) (pressão).
- Espessura da dobra da pele do tríceps (em mm) (triceps).
- Insulina sérica de 2 horas (em mu U/ml) (insulina).
- Índice de massa corporal (peso em kg/(altura em m^2)) (IMC).
- Idade (em anos) (idade).
- Se tem ou não diabetes (grupo, neg: não tem, pos: tem).

O objetivo é criar uma regra de classificação adequada conforme visto em sala, para a variável grupo, em função das outras variáveis. Além disso, faça uma análise descritiva apropriada. Apresente comentários e conclusões pertinentes.

18. O conjunto de dados sobre vinhos (disponível no pacote do R "rattle.data" sob o nome de "wine") contém os resultados de uma análise química de vinhos cultivados em uma área específica da Itália. Três tipos de vinho (grupo) estão representados nas 178 amostras, com os resultados de 13 análises químicas registradas (variáveis de "Alcohol" a "Proline") para cada amostra (variáveis resposta). A variável Type (grupo) foi transformada em uma variável categorizada. O objetivo é criar uma regra de classificação adequada conforme visto em sala, utilizando todas as variáveis disponíveis. Além disso, faça uma análise descritiva apropriada. Apresente comentários e conclusões pertinentes.

19. Considere os dados sobre petróleo bruto (disponíveis no site do curso sob o mesmo nome, com a respectiva descrição). O objetivo é criar e testar uma regra de classificação para identificar a procedência (zonas de arenito) das quais as amostras de petróleo foram extraídas. Além disso, faça uma análise descritiva apropriada. Apresente comentários e conclusões pertinentes.