

ME - 731 Análise Multivariada
Segundo semestre de 2009
Lista de Exercícios III
Entrega: Exercícios 4 e 5 em 20/10/2009

Obs1: Não é necessário digitar a resolução da lista, os exercícios podem ser entregues feitos à mão.

Legenda do banco de dados da Iris.

Sepal.Length - comprimento da sépala, Sepal.Width - largura da sépala, Petal.Width - largura da pétala, Petal.Length - comprimento da pétala.

Legenda do banco de dados das companhias de cereal.

mfr - fabricante (G = General Mills, K = Kellogs, N = Nabisco , P = post, Q = Quaker Oats, R = Ralston Purine), calories - quantidade de calorias em uma porção, protein - quantidade de proteína (em gramas) em uma porção, fat - quantidade de gordura (em gramas) em uma porção, fibre - quantidade de fibra em uma porção.

Exercícios

1. Considere o banco de dados da Íris, especificamente as variáveis: *Petal.Length* e *Petal.Width*, e os 3 grupos de íris: setosa, versicolor e virginica. Responda os itens (sugestão: utilize as metodologias vistas em salas e as funções **mlnm** e **ecl.mlnm**, disponíveis na página do curso):
 - a) Faça uma análise descritiva por variável e por grupo. Você diria que a suposição de normalidade é razoável para os dados? Justifique adequadamente.
 - b) Teste a igualdade de matrizes de covariâncias entre os três grupos, utilizando o teste de Box. Qual a sua conclusão considerando $\alpha = 5\%$?
 - c) Utilizando o modelo linear normal multivariado, em sua forma vetorial, teste a igualdade dos vetores de médias utilizando $\alpha = 5\%$. Qual sua conclusão?
 - d) Utilize os resíduos studentizados e verifique se a suposição de normalidade é razoável para os dados?
 - e) Se você rejeitou a hipótese do item c), identifique as diferenças, em termos de grupos, para cada variável.

- f) Originalmente, tínhamos 6 médias (3 grupos e 2 variáveis). Os testes que você realizou o levaram a diminuir esse número? Apresente as estimativas das médias do modelo final que você obteve e seus respectivos erros-padrão.
- g) A metodologia que você utilizou para analisar os dados em questão é apropriada? Justifique adequadamente.
2. Considere $\mathbf{X} \sim D_2(\mathbf{0}, \Sigma)$, $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, $\rho \in (-1, 1)$. Responda os itens:
- Calcule os autovalores e autovetores (ortonormalizados) associados à Σ .
 - Obtenha as componentes principais associadas ao vetor \mathbf{X} .
 - Interprete as componentes principais que você encontrou no item b).
 - O que ocorre com as componentes principais (em termos de suas variâncias e dos autovetores associados à elas) quando $|\rho| \rightarrow 1$. Como as variáveis originais passam a estar relacionadas às componentes principais.
 - Repita os itens a), b), c), d) quando $\rho = 0$.
3. Considere $\mathbf{X} \sim D_p(\boldsymbol{\mu}, \Sigma)$, Σ uma matriz positiva definida, com autovalores - autovetores (ortonormalizados) $(\lambda_1, \mathbf{e}'_1)', (\lambda_2, \mathbf{e}'_2)', \dots, (\lambda_p, \mathbf{e}'_p)'$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. Prove que a variância generalizada de Σ é igual aos produto dos autovalores de Σ . Discuta as vantagens e desvantagens de se utilizar a variância generalizada ao invés da variância total (soma das variâncias) como um critério de escolha do número de componentes principais.
4. Com base no conjunto de dados sobre informações nutricionais de cereais (o arquivo está disponível na página do curso sob o nome de *cereal.txt*), e considerando a matriz de correlações, utilize a metodologia das componentes principais para responder os seguintes itens:
- Calcule os autovalores e autovetores (ortonormalizados) associados à Σ .
 - Faça o “screeplot” e construa uma tabela com a variância explicada por cada componente e acumulada. Quantas componentes você reteria? Justifique adequadamente sua resposta.
 - Calcule as correlações entre cada uma das quatro componentes e as variáveis originais (utilize o valores simétricos dos autovetores com relação aqueles obtidos pelo programa R).

- d) Interprete cada uma das quatro componentes que você obteve (utilize o valores simétricos dos autovetores com relação aqueles obtidos pelo programa R).
- d) Considere as duas primeiras componentes obtidas. A variância explicada por elas é razoável? Justifique sua resposta.
- e) Faça um gráfico de dispersão entre as componentes do item d). Você conseguiria identificar grupos distintos entre as fabricantes de cereal? Como se caracterizam estes grupos, em termos das componentes principais e em termos das variáveis originais? Sugestão: Calcule medidas descritivas para as componentes principais e para as variáveis originais em cada um desses grupos.
5. Considere os dados analisados na Questão 1) desta lista como todas as quatro variáveis. Responda os itens:
- a) Calcule a matriz de covariância \mathbf{S}_p^2 , utilizando as matrizes de covariância \mathbf{S}_i^2 , $i = 1, 2, 3$.
- b) Calcule a matriz de correlações \mathbf{R} a partir da matriz \mathbf{S}_p^2 .
- c) Calcule os autovalores e autovetores associados à \mathbf{R} e retenha apenas a primeira componente principal. Lembre-se de considerar o valor simétrico dos autovetores que você encontrar, se estiver utilizando o programa R. Interprete tal componente.
- d) Considere que primeira componente resume bem a informação dos grupos. Desejamos testar a hipótese $H_0 : \mu_1 = \mu_2 = \mu_3$ vs $H_1 : \text{pelo menos uma diferença}$, em que $\mu_i, i = 1, 2, 3$ é a média do grupo i com relação à primeira componente principal. Utilize a metodologia usual de modelos lineares univariados para testar a referida hipótese e verificar, caso você rejeite H_0 , onde residem as diferenças. Considere $\alpha = 5\%$.