ME - 731 Análise Multivariada
Segundo semestre de 2009
Lista de Exercícios VI
Entrega: Exercícios 2 e 4 até o dia 15/12 às 10h na sala do Professor (210 IMECC)

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

Obs2: Os exercícios sequem, a menos que mencionado o contrário, as notações definidas nas notas de aulas disponíveis no site e em classe

Exercícios

1. 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 $\boldsymbol{\rho} = Corre(\boldsymbol{X})$, definidas em classe. 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 intens:

   a) Calcule $\mathcal{E}(\boldsymbol{U}^{(\boldsymbol{Z})})$, $\mathcal{E}(\boldsymbol{V}^{(\boldsymbol{Z})})$, $Cov(\boldsymbol{U}^{(\boldsymbol{Z})})$, $Cov(\boldsymbol{V}^{(\boldsymbol{Z})})$, $Corr(\boldsymbol{U}^{(\boldsymbol{Z})})$ e $Corr(\boldsymbol{V}^{(\boldsymbol{Z})})$.

   b) Prove que: $Cov\big(\boldsymbol{Z}^{(1)}, \boldsymbol{U}^{(\boldsymbol{Z})}\big) = Corr\big(\boldsymbol{Z}^{(1)}, \boldsymbol{U}^{(\boldsymbol{Z})}\big) = \boldsymbol{A}^{(\boldsymbol{Z})}\boldsymbol{\rho}_{11}$, $Cov\big(\boldsymbol{Z}^{(2)}, \boldsymbol{V}^{(\boldsymbol{Z})}\big) = Corr\big(\boldsymbol{Z}^{(2)}, \boldsymbol{V}^{(\boldsymbol{Z})}\big) = \boldsymbol{B}^{(\boldsymbol{Z})}\boldsymbol{\rho}_{22}$.

   c) Prove que: $Cov(\boldsymbol{Z}^{(1)}) = \boldsymbol{A}^{(\boldsymbol{Z})-1}\boldsymbol{A}'^{(\boldsymbol{Z})-1}$, $Cov(\boldsymbol{Z}^{(2)}) = \boldsymbol{B}^{(\boldsymbol{Z})-1}\boldsymbol{B}'^{(\boldsymbol{Z})-1}$.

   d) 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 independentes? O são as componentes de $\boldsymbol{V}^{(\boldsymbol{Z})}$? Com relação às componentes de $\boldsymbol{U}^{(\boldsymbol{Z})}$ e $\boldsymbol{V}^{(\boldsymbol{Z})}$, o 0que podemos afirmar no que diz respeito à independência de suas componentes?

2. Considere os arquivo de dados da iris setosa (com as 4 variáveis) e, num primeiro momento, ignore a existência dos grupos (espécie). Considere que $\boldsymbol{X}^{(1)} = (X_1^{(1)}, X_2^{(1)})'$ e $\boldsymbol{X}^{(1)} = (X_1^{(2)}, X_2^{(2)})'$, em que $X_1^{(1)}$: comprimento da sépala, $X_2^{(1)}$: largura da sépala, $X_1^{(2)}$: comprimento da pétala e $X_2^{(2)}$: largura da pétala. Queremos estudar como as

variáveis relacionadas as dimensões das pétalas influenciam aquelas relacionadas às dimensões das sépalas, através das análise de correlações canônicas, utilizando as variáveis padronizadas. Sugestão: veja os slides inerentes à análise de correlação canônica.

Responda os intens:

a) Faça um diagrama de dispersão matricial entre as 4 variáveis. O que você pode dizer a respeito da estrutura de correlação delas?

b) Obtenha os dois pares de variáveis canônicas, apresentando as equações canônicas, as correlações entre cada variável canônica e as variáveis padronizadas e as correlações canônicas entre as variáveis canônicas. Interprete todos estes resultados, de modo adequado. Interprete, inclusive, cada uma das variáveis canônicas em termos das variáveis padronizadas.

c) Calcule o percentual da variabalidade de cada uma das variáveis canônicas explicadas pelas variáveis padronizadas. Calcule a proporção das somas das variâncias das variáveis originais explicadas pelas variáveis canônicas. Interprete todos estes resultados, de modo adequado.

d) Com base nos resultados dos itens anteriores, quantos pares de variáveis canônicas você escolheria para representar os dados? Jusitifique, adequadamente, sua resposta.

e) Utilizando somente o primeiro par de variáveis canônicas, ajuste um modelo de regressão normal linear da primeira variável canônica contra a segunda. Interprete este modelo da forma mais ampla possível e verifique se os coeficientes são significativos. Verifique se a suposição de normalidade é razoável para dos dados.

f) Faça um diagrama de dispersão entre as duas primeiras variáveis canônicas. Como as espécies (setosa, versicolor e virginca) se comportam com relação à estas variáveis? As espécies apresentam comportamentos diferntes? Como cada uma das espécies se caracteriza? Oo seja, descreva o comportamento das espécies. Sugestão: utilize símbolos e ou cores diferentes para cada grupo.

3. Considere a metodologia de análise de discrimante 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, $-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$.

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

d) Repita os itens de a) - c), considerando amostras aleatórias (independentes) dos vetores aleatórios $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$, ou seja, substituindo os parâmetros em questão pelos respectivos estimadores usualmente adotados.

4. Considere os dados da Tabela 11.9 nas páginas 666 e 667 do livro do Johnson & Wichern (sexta edição) sobre características nutricionais de cerais de fabricantes americanas (note que, embora se pareçam, não são exatamente os mesmos dados de ceral considerados anteriormente). Considere todas as variáveis em questão (calorias, proteína, gordura, sódio, fibra, carboidratos, açucar e potássio). Os grupos são os fabricantes (considere apenas dois) G e K. Responda aos itens:

a) Faça um diagrama de dispersão matricial entre as todas as variáveis. O que você pode dizer a respeito da estrutura de correlação delas?

b) Calcule a matriz de covariâncias ponderada $\boldsymbol{S}_p^2$ (considerando os dois grupos) bem como a respectiva matriz de correlações.

c) Assuma distribuição normal multivariada para cada um dos 3 grupos, com iguais matrizes de covariâncias. Assuma também custos iguais de classificação errada e iguais probabilidades a priori de classificação. Considere as 10 primeiras observações de cada um dos grupos e construa a regra de classificação de Fisher, dentro desse contexto. Interprete a função discriminante de modo adequado. Estime a PTCE bem como o TAP (utilizando os outros cereais que não entraram na regra de classificação). Critique, do modo mais completo possível, a regra de classificação que você obteve.

d) Existem indícios de que algum(ns) fabricantes estão relacionados à cereais mais "nutritivos" (alto teor de proteína, baixo teor de gordura, alto teor de fibra, baixo teor de açúcar e assim por diante)? Sugestão: faça um gráfico de dispersão, dos valores da função discriminante, para cada um dos grupos, plotando, no gráfico, as iniciais das fabricantes.