5. Análise discriminante

5.1. Considere que tem 3 grupos para as quais dispomos das seguintes observações para duas variáveis (Johnson e Wichern, 2007):

Grupo	V1	V2
1	-2	5
1	0	3
1	1	1
2	0	6
2	2	4
2	1	2
3	1	-2
3	0	0
3	-1	-4

a) Para cada grupo obtenha as médias amostrais.
b) Obtenha a matriz de variâncias-covariâncias conjunta $\left(S_{pooled}\right)$.
c) Obtenha as matrizes das somas de quadrados entre os grupos, $\mathbf{B}$, e dentro dos grupos, $\mathbf{W}$, ou seja, $\mathbf{B} = \sum_{i=1}^{g} n_i(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})'$ e $\mathbf{W} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{\boldsymbol{x}}_{\boldsymbol{i}}\right)\left(x_{ij} - \overline{\boldsymbol{x}}_{\boldsymbol{i}}\right)'$.
d) Determine os valores próprios da matriz $\mathbf{W^{-1}B}$, ou seja, resolva $\left|\mathbf{W^{-1}B} - \boldsymbol{\lambda}\mathbf{I}\right| = \mathbf{0}$, sendo $\mathbf{I}$ a matriz identidade.
e) Obtenha os valores próprios normalizados $\boldsymbol{v}_i$ associados aos valores próprios $\boldsymbol{\lambda}_i$ obtidos em d), ou seja, resolva $\left(\mathbf{W^{-1}B} - \boldsymbol{\lambda}_i\mathbf{I}\right)\boldsymbol{v}_i = \mathbf{0}$.

5.2. Foi realizado um estudo junto dos alunos de uma turma com apoio pedagógico recorrente a matemática. No ficheiro XXX estão disponíveis as seguintes informações para cada aluno:

- nota: resultado do aluno num teste de avaliação global;
- grupo: resultado final do aluno com base na nota do teste global: A = aprovado (nota superior a 12), AR = aprovado com restrição (nota entre 8 e 12), R = reprovado (nota inferior a 8);
- nhoras: número de horas de estudo semanal;
- autoconc: autoconceito a matemática de acordo com o grau de concordância (valor entre 0 e 10) obtido de um conjunto de itens do tipo "a matemática é uma das minhas disciplinas preferidas);
- dimensao: dimensão do agregado familiar;
- rendimento: rendimento do agregado;
- apoiofam: número de vezes por semana que recorreram aos pais ou irmãos mais velhos;
- apoioprof: número de vezes por semana que recorreram ao professor para esclarecer dúvidas.

a) Considere as variáveis: número de horas de estudo e autoconceito.
   i) Represente a nuvem de pontos, identificando o grupo a que pertence cada ponto. Comente.
   ii) Obtenha a contribuição das variáveis originais na função discriminante.

iii) Qual a importância relativa das variáveis originais na função discriminante?
iv) Escolha o número de funções discriminantes.
v) Construa o mapa territorial.
vi) Obtenha a matriz de confusão. Valide os resultados.

b) Considere todas as variáveis.
i) Valide os pressupostos da análise discriminante.
ii) Selecione as variáveis explicativas.
iii) Obtenha a contribuição das variáveis originais nas funções discriminantes.
iv) Qual a importância relativa das variáveis originais nas funções discriminantes?
v) Escolha o número de funções discriminantes.
vi) Construa o(s) mapa(s) territorial(is).
vii) Obtenha a matriz de confusão. Valide os resultados.

5.3. Foram analisadas algumas profissões segundo os níveis de prestígio, taxa de suicídio, rendimento e níveis de instrução, havendo a suspeita de que existem grupos homogéneos que foram previamente obtidos com o algoritmo *K*-means (ficheiro: *profissõesAD.csv*).

a) Realize uma análise discriminante para identificar as variáveis que melhor discriminam os grupos e construir as respetivas funções discriminantes. Valide os pressupostos e discuta a sua adequação.
b) Com base nas funções obtidas, qual o grupo mais provável para uma profissão caracterizada por um rendimento de 3000 u.m., prestígio 25 e taxa de suicídio 13,5?

5.4. Um gestor de crédito de um determinado banco pretende identificar as características que melhor distinguem os clientes com maior propensão para incumprimento no pagamento de empréstimos. Para isso pensou aplicar a análise discriminante a um conjunto de dados com informação sobre 850 clientes, atuais e potenciais. Os clientes nas primeiras 700 linhas já receberam empréstimos. Os dados estão disponíveis no ficheiro *empréstimos.sav*.

a) Utilize uma amostra aleatória composta por 75% dos 700 clientes que já receberam empréstimos, construa um modelo de análise discriminante. Reserve os restantes 25% para validar a análise. Na construção do modelo deve ter em atenção os pressupostos da análise discriminante. Em caso de violação, compare os resultados de:
   - um modelo de análise discriminante linear (LDA),
   - um o modelo de análise discriminante quadrática (QDA), e
   - um modelo de análise discriminante linear construído com base nas variáveis previamente transformadas para atenuar a violação dos pressupostos.
b) Com base no modelo selecionado, classifique os 150 clientes potenciais em função da sua propensão para incumprimento no pagamento dos empréstimos.