Análise Multivariada

Exame | Duração: 3 horas

1º semestre – 2009/10 | 10/02/2010 – 9 horas

Grupo I — 4.0 valores

1. Seja $\boldsymbol{X}$ um vector aleatório com distribuição $\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, com $\boldsymbol{\mu} = (-1, 0, 1)^t$ e

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 5 & 0 \\ 5 & 4 & 7 \\ 0 & 7 & 3 \end{pmatrix}.$$

   (a) Indique, justificando, duas componentes de $\boldsymbol{X}$ que sejam independentes. (0.5)

   (b) Determine a distribuição conjunta das variáveis aleatórias (1.0)

$$(X_1 + X_2 + X_3)/3, \quad (X_1 + X_3)/2.$$

   (c) Determine a probabilidade de o vector aleatório $\boldsymbol{X}$ estar a uma distância de Mahalanobis de $\boldsymbol{\mu}$ superior a 1.5. (1.0)

2. Seja $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$, onde $\boldsymbol{I}_p$ representa a matriz identidade de dimensão p e $\boldsymbol{C}$ uma matriz ortogonal de dimensão $(p \times p)$. Mostre que $\boldsymbol{C}\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{C}\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$. (1.5)

Grupo II — 5.0 valores

A um conjunto de 23 mulheres foram feitas 3 medições dos níveis de glucose no sangue. As recolhas de sangue foram realizadas em pacientes em jejum. Admita que o vector aleatório associado a esta experiência, $\boldsymbol{X} = (X_1, X_2, X_3)^t$, tem distribuição $\mathcal{N}_3(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$, onde X_i representa o nível de glucose no sangue medido no instante i, com $i = 1, 2, 3$. Os resultados da experiência podem ser sumariados por:

$$\bar{\boldsymbol{x}} = (68.7, 69.7, 73.5)^t,$$

$$\boldsymbol{S}_X = \begin{pmatrix} 876.9342 & 268.3158 & 143.3684 \\ 268.3158 & 621.6211 & -0.0316 \\ 143.3684 & -0.0316 & 293.0105 \end{pmatrix} \quad \text{e} \quad \boldsymbol{S}_X^{-1} = \begin{pmatrix} 0.0014 & -0.0006 & -0.0007 \\ -0.0006 & 0.0019 & 0.0003 \\ -0.0007 & 0.0003 & 0.0038 \end{pmatrix}.$$

Note que $\boldsymbol{S}_X$ representa o estimador centrado de $\boldsymbol{\Sigma}$.

1. Teste a hipótese $H_0 :\ \boldsymbol{\mu}_1 = (70, 70, 70)^t$ contra a alternativa $H_1 :\ \boldsymbol{\mu}_1 \neq (70, 70, 70)^t$, ao nível de significância 5%. (2.5)

2. Admita que a outras 19 mulheres em jejum se pediu para ingerirem uma bebida com elevado teor de açúcar e que em 3 momentos distintos foram feitas medições do nível de glucose no sangue das referidas mulheres. Seja $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^t$ o vector aleatório associado a esta nova experiência que se admite ter distribuição $\mathcal{N}_3(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$, onde Y_i representa o nível de glucose no sangue medido no instante i, com $i = 1, 2, 3$. Considere os seguintes resultados: (2.5)

$$\bar{\boldsymbol{y}} = (106.75, 100.6, 115.8)^t,$$

$$\boldsymbol{A} = \begin{pmatrix} 498.8476 & 141.61335 & 74.3440 \\ 141.6134 & 352.77715 & -1.1633 \\ 74.3440 & -1.163286 & 187.0544 \end{pmatrix} \quad \text{e} \quad \boldsymbol{A}^{-1} = \begin{pmatrix} 0.0024 & -0.0010 & -0.0010 \\ -0.0010 & 0.0032 & 0.0004 \\ -0.0010 & 0.0004 & 0.0057 \end{pmatrix},$$

Onde $\boldsymbol{A}$ representa a matriz de covariâncias combinada.

Os médicos afirmam que os níveis médios antes e depois de ingerir a bebida com elevado teor de açúcar são distintos. Formule o teste de hipóteses que achar apropriado para confirmar a conjectura dos médicos e teste-a ao nível de significância de 5%. Comente o modo como este estudo foi conduzido e sugira um delineamento experimental que a seu ver possa ser mais adequado.

Grupo III — 6.0 valores

1. Mediu-se o comprimento, largura e altura da carapaças de 24 tartarugas fêmeas.

   Os dois maiores valores próprios, $\hat{\lambda}_j$, e correspondentes vectores próprios, $\hat{\boldsymbol{\gamma}}_j$, $j = 1, 2$, da matriz de correlações das observações são os seguintes:

	$\hat{\boldsymbol{\gamma}}_1$	$\hat{\boldsymbol{\gamma}}_2$
comprimento	0.578	-0.141
largura	0.577	0.626
altura	0.577	0.767
$\hat{\lambda}_j$	2.943	0.033

   (a) Qual a percentagem da variação total das observações explicada por cada uma e pelas duas componentes principais? (1.5)

   (b) Interprete as duas componentes principais. (1.5)

   (c) Mostre que o coeficiente de correlação entre a i-variável estandardizada, Z_i, e a j-ésima componente principal, Y_i é dado por: (1.0)

   $$\text{Cor}(Z_i, Y_j) = \sqrt{\lambda_j}\gamma_{ij},$$

   onde λ_j é o j-ésimo maior valor próprio da matriz de correlações das variáveis em estudo e $\boldsymbol{\gamma}_j = (\gamma_{1j}, \ldots, \gamma_{pj})^t$ é o vector próprio associado a λ_j.

   (d) Estime os coeficientes de correlações entre cada uma das componentes principais com Z_1, Z_2 e Z_3. Será que estes valores substanciam a interpretação formulada na alínea (1b)? Acha este processo de interpretação mais vantajoso? Justifique. (1.5)

   (e) Considere que se mediram 2 tartarugas fêmeas e se obtiveram os seguintes valores, já estandardizados: (0.5)

	comprimento	largura	altura
T_1	-0.118	-0.332	-0.056
T_2	-1.410	-1.443	-1.756

   Indique um critério de ordenação das tartarugas e ordene-as tendo em conta as medições apresentadas na tabela anterior.

Grupo IV — 5.0 valores

1. Mostre que a distância de Mahalanobis é invariante para transformações lineares da forma $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$, onde $\mathbf{A}$ é uma matrix não singular. (1.0)

2. Dada uma matriz de semelhanças, $\boldsymbol{C} = [c_{ij}]$, uma transformação habitual para obter, a partir de $\boldsymbol{C}$, uma matriz de dissemelhanças $\boldsymbol{D} = [d_{ij}]$ é $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}$. Mostre que de facto $\boldsymbol{D}$ é uma matriz de dissemelhanças. (1.0)

3. Considere a matriz de semelhanças: (1.0)

$$\boldsymbol{C} = \begin{bmatrix} 5 & 0 & 3 & 5 & 1 & 3 \\ & 5 & 2 & 0 & 4 & 2 \\ & & 5 & 3 & 1 & 3 \\ & & & 5 & 1 & 3 \\ & & & & 5 & 3 \\ & & & & & 5 \end{bmatrix},$$

que dá a semelhança entre seis locais arqueológicos. A semelhança é medida pelo número de tipos de cerâmica que dois locais têm em comum. Nota alguma regularidade que lhe permita destacar algum agrupamento dos locais?

4. Sabendo que a matriz de dissemelhanças obtida a partir de $\boldsymbol{C}$ pela transformação referida em (2) é:

$$\boldsymbol{D} = \begin{bmatrix} 0 & 3.16 & 2 & 0 & 2.83 & 2 \\ & 0 & 2.45 & 3.16 & 1.41 & 2.45 \\ & & 0 & 2 & 2.83 & 2 \\ & & & 0 & 2.83 & 2 \\ & & & & 0 & 2 \\ & & & & & 0 \end{bmatrix},$$

aplique o método da ligação completa e desenhe o respectivo dendrograma. Interprete os resultados obtidos sugerindo uma partição adequada dos locais arqueológicos. Compare os resultados obtidos com a sua resposta à pergunta 3. (2.0)

FORMULÁRIO

- $T^2(p,n) = \dfrac{np}{n-p+1} F(p, n-p+1).$
- $(n-1)(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^t \boldsymbol{S}_n^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim T^2(p, n-1)$
- Se $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ então

$$\frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^t \boldsymbol{S}_p^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) \sim T^2(p, n_1 + n_2 - 2)$$

.