1ª prova – 2º/2012

1. Os autovalores e autovetores de uma matriz de correlações amostral são

$$\widehat{\lambda}_1 = 1,96, \quad \widehat{\boldsymbol{e}}_1 = (0,625;\ 0,593;\ 0,507)',$$
$$\widehat{\lambda}_2 = 0,68, \quad \widehat{\boldsymbol{e}}_2 = (-0,219;\ -0,491;\ 0,843)',$$
$$\widehat{\lambda}_3 = 0,36 \quad \text{e} \quad \widehat{\boldsymbol{e}}_3 = (0,749;\ -0,638;\ -0,177)'.$$

   (a) Adotando um modelo com um único fator comum, apresente estimativas das cargas fatoriais, das comunalidades e das variâncias específicas.

   Solução. Aplicando o método dos componentes principais a um modelo com $k = 1$ fator, a estimativa da matriz (neste caso, vetor $p \times 1$) de cargas fatoriais é dada por

$$\widehat{\boldsymbol{\Lambda}} = \widehat{\lambda}_1^{1/2}\,\widehat{\boldsymbol{e}}_1 = \sqrt{1,96} \times \begin{pmatrix} 0,625 \\ 0,593 \\ 0,507 \end{pmatrix} = \begin{pmatrix} 0,875 \\ 0,830 \\ 0,710 \end{pmatrix}. \tag{1}$$

   A partir de (1) obtemos as estimativas das comunalidades, dadas por $\widehat{h}_1^2 = 0,875^2 = 0,766$, $\widehat{h}_2^2 = 0,830^2 = 0,689$ e $\widehat{h}_3^2 = 0,710^2 = 0,504$. Uma vez que foi utilizada a matriz de correlações amostrais, as estimativas das variâncias específicas são $\widehat{\psi}_1 = 1 - \widehat{h}_1^2 = 0,234$, $\widehat{\psi}_2 = 1 - \widehat{h}_2^2 = 0,311$ e $\widehat{\psi}_3 = 1 - \widehat{h}_3^2 = 0,496$.

   (b) Apresente uma forma de avaliar se o ajuste do modelo é satisfatório e comente.

   Solução. De acordo com o item (a), as estimativas das três comunalidades são maiores do que $0,5$, sendo que $\widehat{h}_3^2 \cong 0,5$. Notar que a percentagem da variância total (das variáveis padronizadas) explicada pelo primeiro componente principal é $100 \times 1,96/3 \cong 65\%$, que não é muito alta.

2. A Figura 1 contém o gráfico de dispersão de uma amostra de n pares das variáveis X_1 e X_2.

   (a) As observações A, B e C são aberrantes?

   Solução. Analisando as $n - 3$ observações (X_1, X_2) identificadas pelo símbolo "$\bullet$", concluímos que em relação a estas, as observações A, B e C são aberrantes bidimensionais.

   (b) Considere a amostra de $n - 3$ observações identificadas pelo símbolo "$\bullet$". Compare os resultados das análises de componentes principais nas seguintes condições: (i) com estas observações e (ii) com a amostra de $n-2$ observações obtida pela inclusão do ponto A.

Solução. Na Figura 1 apresentamos a localização aproximada dos eixos dos escores dos componentes principais (Y_1^*, Y_2^*) correspondentes à amostra de $n-3$ observações identificadas pelo símbolo "$\bullet$". Pela sua posição, notamos que a inclusão da observação A não provoca grande mudança nos eixos Y_1^* e Y_2^*, mas contribui para aumentar a variância dos escores Y_1^*. Sendo assim, ao incluir a observação A, há um aumento da proporção da variância total explicada pelo primeiro componente principal (e consquentemente há uma redução da proporção da variância total explicada pelo segundo componente principal).

(c) Considere a amostra de $n-2$ observações identificadas pelo símbolo "$\bullet$" e mais a observação A. Compare os resultados das análises de componentes principais nas seguintes condições: (i) com estas observações e (ii) com a amostra completa de n observações.

Solução. Pela localização dos pontos B e C, notamos que a inclusão destas observações não provoca grande mudança em relação á solução comentada no item 2b. Entretanto, a inclusão das observações B e C contribui para aumentar a variância dos escores Y_2^*. Sendo assim, ao incluir as observações B e C, há uma redução da proporção da variância total explicada pelo primeiro componente principal (e consquentemente há um aumento da proporção da variância total explicada pelo segundo componente principal).

3. O vetor $\boldsymbol{X}$ com distribuição $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ é dividido como

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{Z} \end{pmatrix}, \text{ com dimensões } q \times 1 \ (\boldsymbol{Y}) \text{ e } (p-q) \times 1 \ (\boldsymbol{Z}),\ p > 1 \text{ e } 1 \leq q < p.$$

O vetor de médias e a matriz de covariâncias são divididos como

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{pmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}'_{YZ} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix}.$$

$\boldsymbol{A}$ e $\boldsymbol{B}$ são matrizes $q \times q$ e $q \times (p-q)$, respectivamente. Determine a distribuição de $\boldsymbol{AY} + \boldsymbol{BZ}$.

Solução. Escrevemos

$$\boldsymbol{AY} + \boldsymbol{BZ} = \begin{bmatrix} \boldsymbol{A}, & \boldsymbol{B} \end{bmatrix} \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{Z} \end{pmatrix} = \boldsymbol{HX},$$

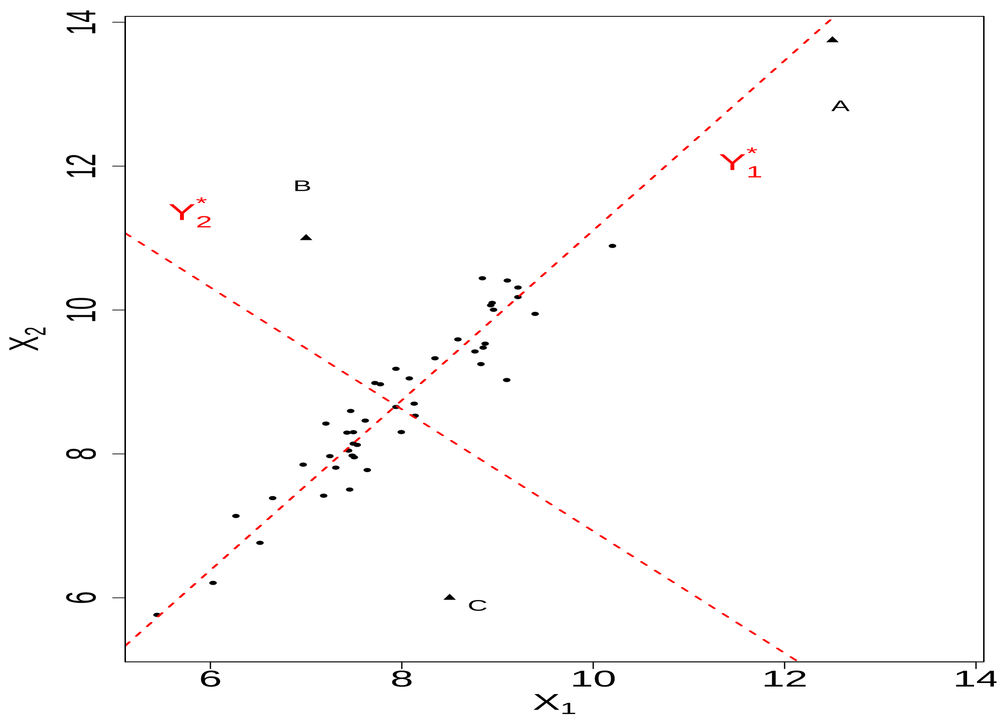

Figura 1: Gráfico de dispersão com escores dos componentes principais.

em que $\boldsymbol{H} = [\boldsymbol{A}, \boldsymbol{B}]$ é uma matriz $q \times p$. Portanto, $\boldsymbol{AY} + \boldsymbol{BZ}$ é um vetor de q combinações lineares de $\boldsymbol{X}$. Logo, a distribuição de $\boldsymbol{AY} + \boldsymbol{BZ}$ é normal q-variada com vetor de médias

$$\boldsymbol{H\mu} = \begin{bmatrix} \boldsymbol{A}, & \boldsymbol{B} \end{bmatrix} \begin{pmatrix} \boldsymbol{\mu}_Y \\ \boldsymbol{\mu}_Z \end{pmatrix} = \boldsymbol{A\mu}_Y + \boldsymbol{B\mu}_Z$$

e matriz de covariâncias

$$\begin{aligned} \boldsymbol{H\Sigma H}' &= \begin{bmatrix} \boldsymbol{A}, & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YZ} \\ \boldsymbol{\Sigma}'_{YZ} & \boldsymbol{\Sigma}_{ZZ} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}' \\ \boldsymbol{B}' \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A\Sigma}_{YY} + \boldsymbol{B\Sigma}'_{YZ}, & \boldsymbol{A\Sigma}_{YZ} + \boldsymbol{B\Sigma}_{ZZ} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}' \\ \boldsymbol{B}' \end{bmatrix} \\ &= \boldsymbol{A\Sigma}_{YY}\boldsymbol{A}' + \boldsymbol{B\Sigma}'_{YZ}\boldsymbol{A}' + \boldsymbol{A\Sigma}_{YZ}\boldsymbol{B}' + \boldsymbol{B\Sigma}_{ZZ}\boldsymbol{B}'. \end{aligned}$$

4. (a) O vetor $\boldsymbol{X} = (X_1, X_2)'$ tem função densidade $f(x_1, x_2) = C$, se $x_1^2 + x_2^2 \leq k^2$, e $f(x_1, x_2) = 0$, caso contrário. Determine o valor de C.
Solução. Definimos $A = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq k^2\}$. Como $f(x_1, x_2)$ é uma

função densidade, devemos ter

$$1 = \iint\limits_{\mathbb{R}^2} f(x_1, x_2) dx_1 dx_2 = \iint\limits_{A} C dx_1 dx_2 = C \iint\limits_{A} dx_1 dx_2$$
$$= C \times \text{área(A)} = C\pi k^2.$$

Logo, $C = \frac{1}{\pi k^2}$.

(b) Afirma-se que

$$f(\boldsymbol{x}) = \frac{1}{4\sqrt{3\pi^3}} \exp\left\{ -\left[\frac{(x_1-1)^2}{6} + \frac{(x_2-2)^2}{4} + \frac{x_3^2}{2} \right] \right\}, \ \boldsymbol{x} \in \mathbb{R}^3,$$

representa a função densidade de um vetor aleatório $\boldsymbol{X} = (X_1, X_2, X_3)'$ com distribuição $\mathrm{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Determine $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$. As variáveis são independentes?
Solução. Como $p = 3$, $f(\boldsymbol{x})$ deve ser escrita como

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{3/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right].$$

Iniciamos simplificando as parcelas entre colchetes na expressão de $f(\boldsymbol{x})$ do enunciado, obtendo

$$f(\boldsymbol{x}) = \frac{1}{4\sqrt{3\pi^3}} \exp\left\{ -\frac{1}{2} \left[\frac{(x_1-1)^2}{3} + \frac{(x_2-2)^2}{2} + x_3^2 \right] \right\}, \ \boldsymbol{x} \in \mathbb{R}^3. \qquad (2)$$

A expressão entre colchetes na eq. (2) corresponde a $(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$. Notando que na eq. (2) os coeficientes de $x_j x_l$, para $j \neq l$ são nulos, temos que $\boldsymbol{\mu} = (1, 2, 0)'$, $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(1/3, 1/2, 1)$, de modo que $\boldsymbol{\Sigma} = \mathrm{diag}(3, 2, 1)$, $|\boldsymbol{\Sigma}| = 6$ e $(2\pi)^{3/2} |\boldsymbol{\Sigma}|^{1/2} = 4\sqrt{3\pi^3}$. As variáveis são independentes, pois a distribuição é normal e a matriz de covariâncias $\boldsymbol{\Sigma}$ é diagonal.