1.3. The data matrix

$$X_{n\times p} = \begin{bmatrix} \underset{\sim}{x}_1^T \\ \vdots \\ \underset{\sim}{x}_n^T \end{bmatrix} = \begin{bmatrix} \underset{\sim}{x}_{(1)} & \cdots & \underset{\sim}{x}_{(p)} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

1.4. Summary statistics

$$\bar{x}_i = \frac{1}{n}\sum_{r=1}^{n} x_{ri}$$ sample mean (variable i)

$$\bar{\underset{\sim}{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$ sample mean vector

$\underset{\sim}{x}_r = \begin{bmatrix} x_{r1} & \cdots & x_{rp} \end{bmatrix}^T$

$$= \frac{1}{n}\sum_{r=1}^{n} \underset{\sim}{x}_r = \frac{1}{n} X^T \underset{\sim}{1}$$

$$= \frac{1}{n} \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{np} \end{bmatrix}}_{p\times n} \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{n\times 1}$$

$$= \frac{1}{n}\left(\sum_{r=1}^{n} x_{r1} + \cdots + \sum_{r=1}^{n} x_{rp}\right)$$

$$\sigma_{ii} = \frac{1}{n} \sum_{n=1}^{n} (x_{ni} - \bar{x}_i) = \sigma_i^2$$ SAMPLE VARIANCE

$$\sigma_{ij} = \frac{1}{n} \sum_{n=1}^{n} (x_{ni} - \bar{x}_i)(x_{nj} - \bar{x}_j)$$ SAMPLE COVARIANCE

$$= \frac{1}{n} \sum_{n=1}^{n} \left(x_{ni} x_{nj} - x_{ni} \cancel{\bar{x}_j}^{0} - \cancel{\bar{x}_i}^{0} x_{nj} + \bar{x}_i \bar{x}_j \right)$$

$$S = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{np} \end{bmatrix}$$

SAMPLE VARIANCE-COVARIANCE MATRIX

↳ possui efeitos de escala

$$S = \frac{1}{n} \sum_{n=1}^{n} (x_n - \bar{x})(x_n - \bar{x})$$

$$= \frac{1}{n} X^T X - \bar{x}\bar{x}^T = \frac{1}{n} \left(X^T X - \frac{1}{n} X^T \mathbf{1}_n \mathbf{1}_n^T X \right)$$

$$\underset{p \times m}{X^T} \underset{m \times p}{X} = \begin{bmatrix} \sum_{n=1}^{n} x_{n1}^2 & \cdots & \sum_{n=1}^{n} x_{n1} x_{np} \\ \vdots & & \\ \sum_{n=1}^{n} x_{n1} x_{np} & \cdots & \sum_{n=1}^{n} x_{np}^2 \end{bmatrix}_{p \times p}$$

$$\underset{n \times 1}{\mathbf{1}_n} \underset{1 \times n}{\mathbf{1}_n^T} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

$$\underset{p \times n}{X^T} \underset{n \times n}{\mathbf{1}_n \mathbf{1}_n^T} = \begin{bmatrix} \sum x_{n1} & \cdots & \sum x_{n1} \\ \vdots & \ddots & \vdots \\ \sum x_{np} & \cdots & \sum x_{np} \end{bmatrix}$$

$$X^T \frac{1}{n} \mathbf{1}^T X = \begin{bmatrix} \sum x_{n0} & \cdots & \sum x_{n0} \\ \vdots & \ddots & \vdots \\ \sum x_{np} & \cdots & \sum x_{np} \end{bmatrix} \begin{bmatrix} x_{01} & \cdots & x_{0p} \\ \vdots & \ddots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \left(\sum x_{n0}\right)^2 & \cdots & \left(\sum x_{n1}\right)\left(\sum x_{np}\right) \\ \vdots & \ddots & \vdots \\ \left(\sum x_{np}\right)\left(\sum x_{n0}\right) & \cdots & \left(\sum x_{np}\right)^2 \end{bmatrix}$$

$$= \frac{1}{n} X^T H X \quad \text{(Quadratic form)}$$

$$H = I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \qquad \text{CENTERING MATRIX}$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & & \ddots & \vdots \\ \vdots & & & \\ 1/n & 1/n & \cdots & 1/n \end{bmatrix} = \begin{bmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \frac{n-1}{n} \\ -\frac{1}{n} & \cdots & & \end{bmatrix}$$

$H = H^T \Rightarrow H$ is IDEMPOTENT

- SINCE H is idempotent, it follows that, for any p-vector a

$$a^T S a = \frac{1}{n} a^T X^T H^T H X a = \frac{1}{n} y^T y \geq 0$$

$$y = HXa$$

↳ prove that if $n > p \Rightarrow S > 0$

$\therefore S \geq 0$ (positive semi-definite)

$$M = \sum_{n=1}^{n} x_n x_n^T = X^T X$$

matrix of sums of squares and products

$$r_{ij} = s_{ij} / (s_i s_j)$$

$$R = (r_{ij}) \geq 0 \; ?$$

$$= D^{-1} S D^{-1}$$

$D = \text{DIAG}(s_i)$ and $S = DRD$

- distância de Mahalanobis
- distância de Pearson
- correlation coefficient
- distância euclidiana (padronizadas)
- sample correlation matrix ("padroniza" a covariância)
  1. Pearson
  2. Spearman
  3. Correl. Parciais
  4. matriz de importância
  5. cópulas
  6. distância entre observações

$$\text{trace}(S) = \text{trace}(\Delta) = \text{trace}\left(\text{diag}(\lambda_i)\right) = \sum_{i=1}^{p} \lambda_i$$ VARIÂNCIA TOTAL

$$(b) - |\Delta| = \prod_{i=1}^{p} \lambda_i$$ VARIÂNCIA GENERALIZADA

1.5 Linear Combinations

- Consider the following linear combination:

$$y_n = a_1 x_{n1} + \cdots + a_p x_{np} \qquad n = 1, \ldots, N$$

where a_i, $i = 1, \ldots, p$, are given.

- The mean of y is given by:

$$\bar{y} = \frac{1}{N} \underset{\sim}{a}^T \sum_{n=1}^{N} \underset{\sim}{x}_n = \underset{\sim}{a}^T \underset{\sim}{\bar{x}}$$

$$\begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \sum_{i=1}^{p} a_i \bar{x}_i$$ linear combination of the mean of each column

and the variance

$$\sigma^2_n = \frac{1}{N} \sum_{n=1}^{N} (y_n - \bar{y})^2 = \frac{1}{N} \sum_{n=1}^{N} \underset{\sim}{a}^T (\underset{\sim}{x}_n - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_n - \underset{\sim}{\bar{x}})$$

- In general, we may be interested in q-dimensional ($q \ll p$) linear transformation

$$\underset{\sim}{y}_n = A \underset{\sim}{x}_n + \underset{\sim}{b}, \qquad n = 1, \ldots, N$$

$$Y = X A^T + \mathbf{1} b^T, \qquad A_{q \times p}, \; b_{q \times 1}$$

- THE MEAN VECTOR AND COVARIANCE MATRIX FOR $\underset{\sim}{y_n}$ IS GIVEN BY:

$$\bar{\underset{\sim}{y}} = A\bar{\underset{\sim}{x}} + \underset{\sim}{b}$$

$$S_y = \frac{1}{m}\sum_{n=1}^{m}(\underset{\sim}{y_n} - \bar{\underset{\sim}{y}})(\underset{\sim}{y_n} - \bar{\underset{\sim}{y}})^T = ASA^T$$

- IF A IS NON-SINGULAR, THEN

$$S = A^{-1}S_y(A^T)^{-1}$$

- SOME EXAMPLES OF LINEAR TRANSFORMATIONS ARE:

1. SCALING TRANSFORMATION: SCALES EACH VARIABLE TO HAVE UNIT VARIANCE

$$* \quad y_n = D^{-1}(\underset{\sim}{x_n} - \bar{\underset{\sim}{x}}), \quad n = 1, \ldots, m$$

$$D = \text{DIAG}(s_i)$$

2. MAHALANOBIS TRANSFORMATION: ELIMINATES THE CORRELATION BETWEEN VARIABLES AND STANDARDIZE ITS VARIANCE.

IF $S > 0 \Rightarrow S^{-1}$ HAS UNIQUE POSITIVE SQUARE ROOT $S^{-1/2}$ AND

$$* \quad \underset{\sim}{z_n} = S^{-1/2}(\underset{\sim}{x_n} - \bar{\underset{\sim}{x}}), \quad n = 1, \ldots, m$$

3. Principal component transformation: By the Spectral Decomposition theorem the covariance matrix may be written as:

spectral decomposition

$$S = G L G^T, \quad GG^T = G^T G = I \text{ (identity)}$$

G: an orthogonal matrix of eigenvectors

L: diagonal matrix of the eigen values of S

$(l_1 \geq l_2 \geq \dots \geq l_p \geq 0)$

* $W_n = G^T (x_n - \bar{x})$

Since $S_w = G^T S G = L$ is diagonal, the columns of W, called principal components, represent uncorrelated linear combination of the variables.

sample matrix

Characteristic roots of S (eigenvalues)

$$|S - \lambda I_p| = 0 \Rightarrow \lambda = [\lambda_1 \lambda_2 \dots \lambda_p]$$

$$\Rightarrow L = \text{DIAG}(\lambda_i)$$

Exercise

$$S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}, \quad s_{11} = s_{22}$$

Compute the eigenvalues and eigenvectors

NOTE
- If $n > p \Rightarrow S$ has $\text{RANK}(S) \leq p \Rightarrow$
  $\Rightarrow$ At most p non-zero eigenvalues
- If $n < p \Rightarrow S$ has $n-p$ null eigenvalues

EXERCISE

State the relationship between:

1. n and p dimensions of a matrix S
2. Rank(S)
3. Eigenvalues of S

ESTATÍSTICAS DAS UNIDADES AMOSTRAIS

Estatísticas descritivas das unidades amostrais
- R-técnicas
- Q-técnicas

$D_{n \times n} = (d^2_{Eik})$ distância euclidiana amostral

$d^2_{Eik} = d^2_{ik} = (x_i - x_k)^T (x_i - x_k)$

$: \mathbb{R}^p \to \mathbb{R}$

$$d^2_{P_{ik}} = (x_i^* - x_k^*)^T (x_i^* - x_k^*) = (x_i - x_k)^T D_{S_{jj}}^{-1} (x_i - x_k)$$

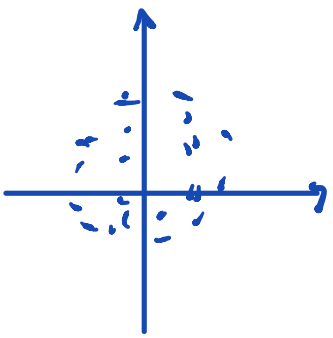
∴ DISTÂNCIA EUCLIDIANA PADRONIZADA (PEARSON)

$$= 2(1 - r_{ik})$$

↳ COEFICIENTE DE CORRELAÇÃO DE PEARSON

$$d^2_{M_{ik}} = z_i^T z_i = (y_i - \bar{y})^T S^{-1} (y_i - \bar{y})$$ DISTÂNCIA DE MAHALANOBIS

INTERPRETAÇÃO GEOMÉTRICA



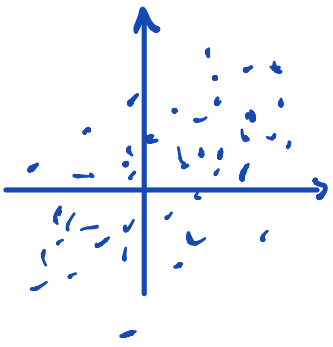
VARIÁVEIS:

a) INDEPENDENTES
b) HOMOCEDÁSTICAS

$$d^2_E = (x - \bar{x})^T (x - \bar{x})$$

a) INDEPENDENTES
b) HETEROCEDÁSTICAS

$$d^2_P = (x - \bar{x})^T D_{S_{ii}}^{-1} (x - \bar{x})$$



a) DEPENDENTES
b) HETEROCEDÁSTICAS

$$d^2_M = (y - \bar{y})^T S^{-1} (y - \bar{y})$$