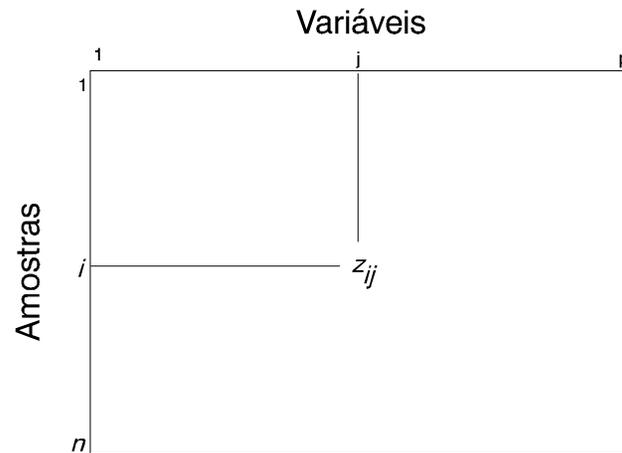


QUADRO DE DADOS GENÉRICO EM ANÁLISE DE DADOS

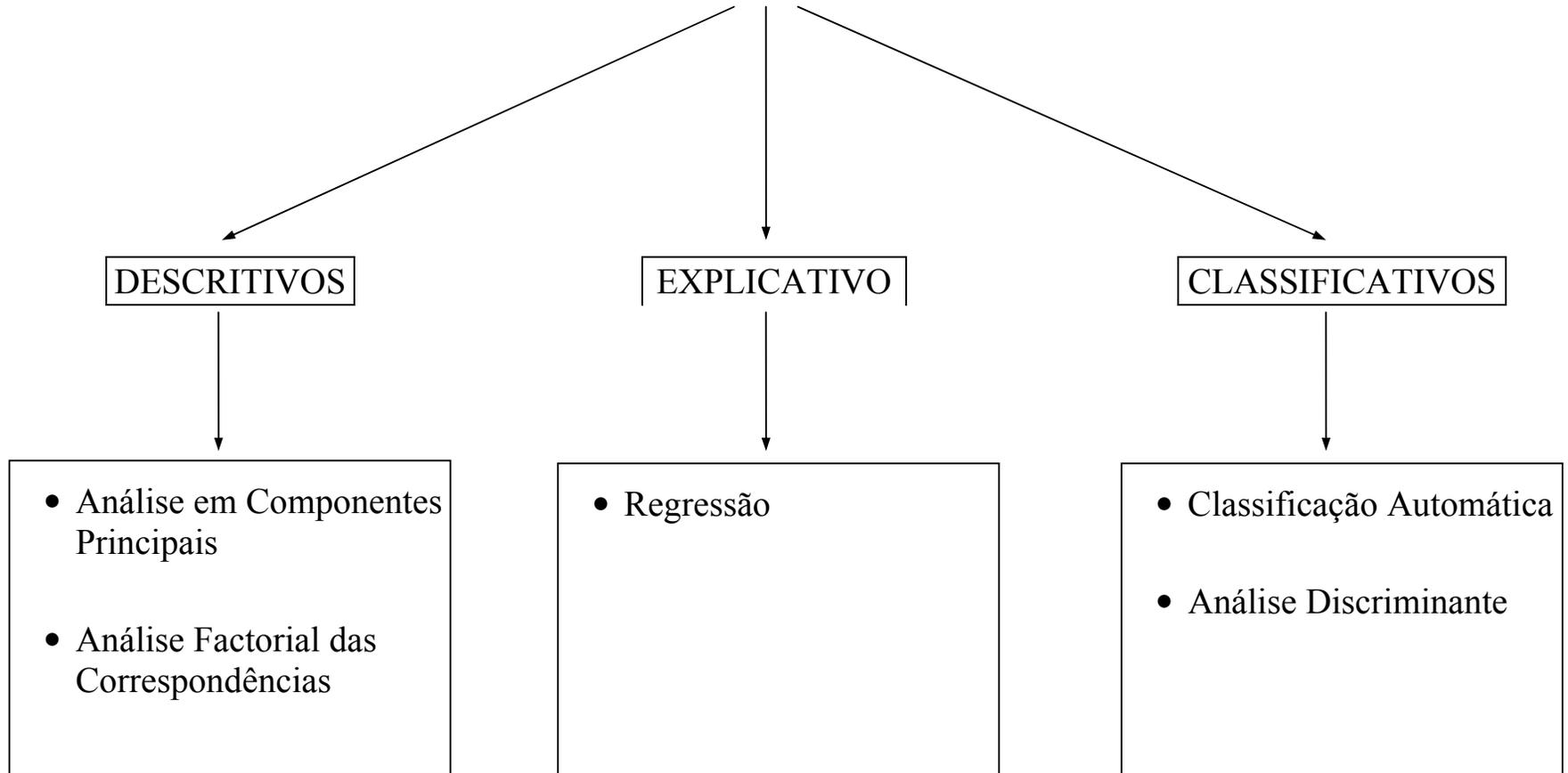


O elemento genérico z_{ij} corresponde ao valor do atributo (variável) j medido na amostra i .

Tipos de variáveis

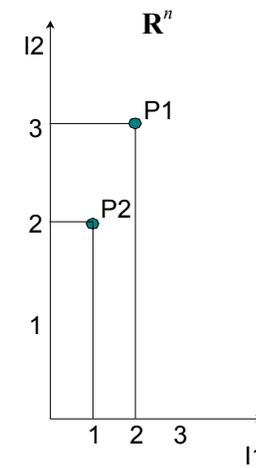
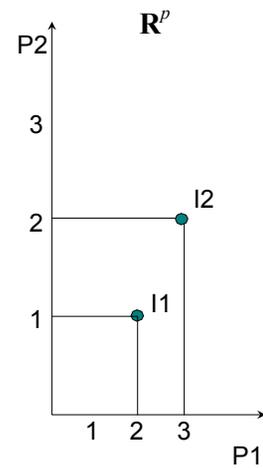
- Quantitativas (contínuas)
- Qualitativas (discretas)
 - Binárias
 - Lógicas

MÉTODOS DE ANÁLISE DE DADOS



INTERPRETAÇÃO GEOMÉTRICA DE UM QUADRO DE DADOS

	P1	P2
I1	2	1
I2	3	2



Cada linha do quadro pode ser tomada como um vector (ou um ponto) que representa a posição de uma amostra no espaço das variáveis (\mathbf{R}^p)

Cada coluna do quadro pode ser tomada como um vector (ou um ponto) que representa a posição de uma variável no espaço das amostras (\mathbf{R}^n)

ALGUNS CONCEITOS GEOMÉTRICOS

Seja uma nuvem constituída por n pontos x_i com coordenadas x_{ij} e massa m_i .

CENTRO DE GRAVIDADE

$$\mathbf{g} = \sum_{i=1}^n m_i \mathbf{t}_i$$

$$m_i = \frac{1}{n} \Rightarrow g_j = \frac{1}{n} \sum_{i=1}^n t_{ij} = \bar{t}_j \text{ (média da variável } j\text{)}$$

INÉRCIA EM RELAÇÃO AO CENTRO DE GRAVIDADE

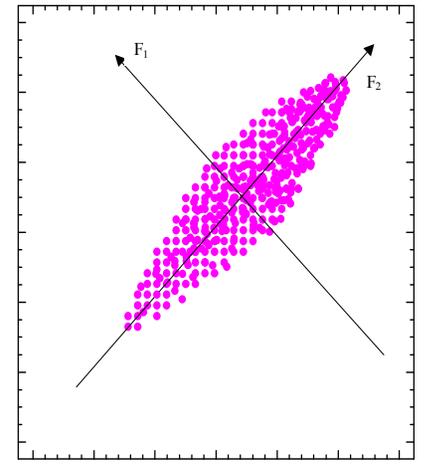
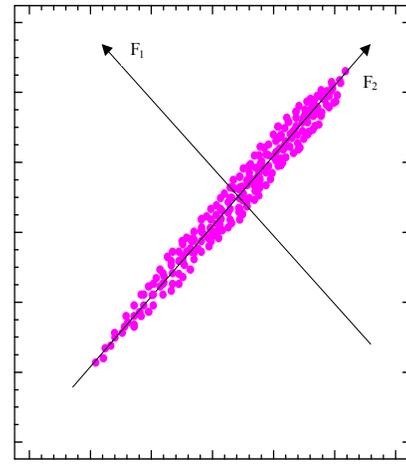
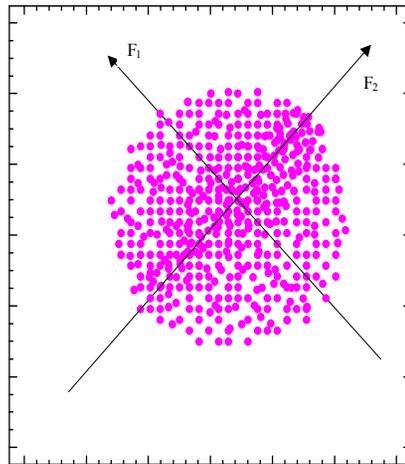
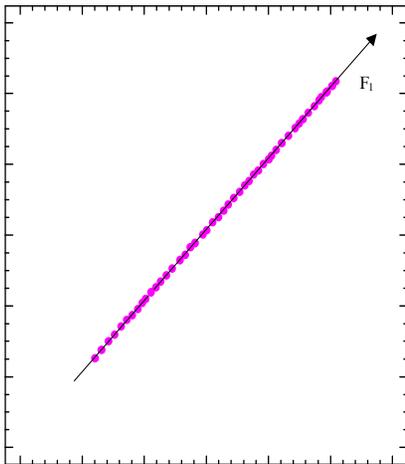
$$\mathbf{I}_{\mathbf{g}} = \sum_{i=1}^n m_i (\mathbf{t}_i - \mathbf{g})^2$$

$$m_i = \frac{1}{n} \Rightarrow I_{g_j} = \frac{1}{n} \sum_{i=1}^n (t_{ij} - g_j)^2 \text{ (variância da variável } j\text{)}$$

INTRODUÇÃO AOS MÉTODOS FACTORIAIS

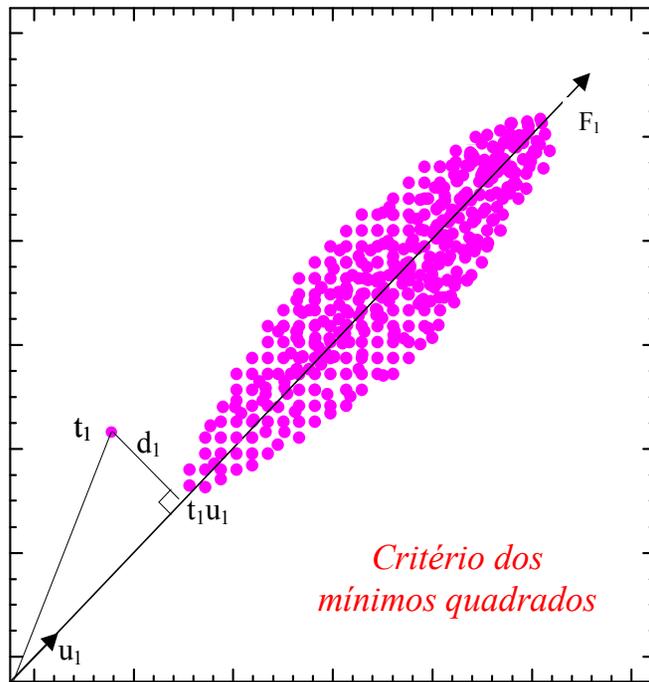
INTRODUÇÃO

Encontrar o sistema de eixos que melhor ajusta a nuvem de pontos, permitindo diminuir a dimensão do espaço com perda mínima de informação.



INTRODUÇÃO AOS MÉTODOS FACTORIAIS

Análise em R^p



Maximizando

$$(\mathbf{T}\mathbf{u}_1)' \mathbf{T}\mathbf{u}_1 = \mathbf{u}_1' \mathbf{T}' \mathbf{T}\mathbf{u}_1$$

Matriz
de
Inércia

sujeita aos constrangimentos

$$\mathbf{u}_1' \mathbf{u}_1 = 1$$

obtém-se

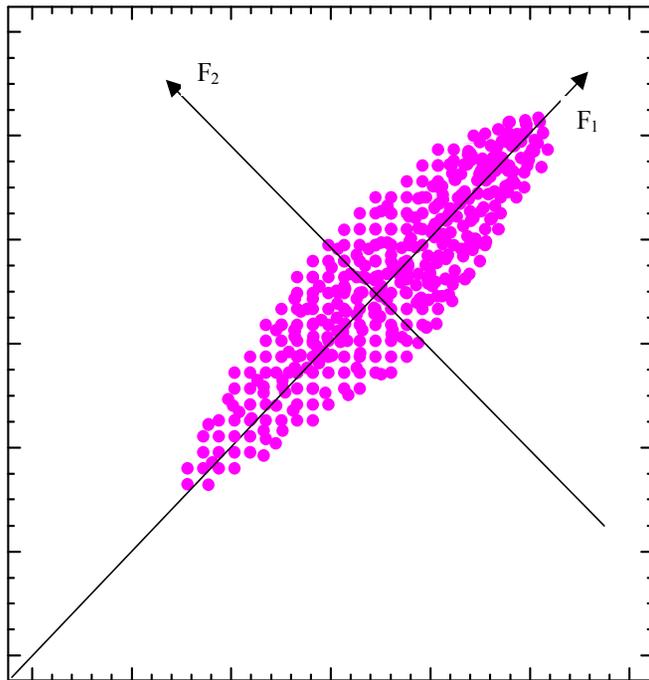
$$\mathbf{T}' \mathbf{T}\mathbf{u}_1 - \lambda_1 \mathbf{u}_1 = 0$$

1º
valor próprio de
 $\mathbf{T}' \mathbf{T}$

1º
vector próprio de
 $\mathbf{T}' \mathbf{T}$

INTRODUÇÃO AOS MÉTODOS FACTORIAIS

Análise em R^p



Maximizando

$$(\mathbf{T}\mathbf{u}_2)' \mathbf{T}\mathbf{u}_2 = \mathbf{u}_2' \mathbf{T}' \mathbf{T} \mathbf{u}_2$$

Matriz
de
Inércia

sujeita aos constrangimentos

$$\mathbf{u}_1' \mathbf{u}_2 = 0$$

$$\mathbf{u}_2' \mathbf{u}_2 = 1$$

obtém-se

$$\mathbf{T}' \mathbf{T} \mathbf{u}_2 - \lambda_2 \mathbf{u}_2 = 0$$

2º
valor próprio de
 $\mathbf{T}' \mathbf{T}$

2º
vector próprio de
 $\mathbf{T}' \mathbf{T}$

INTRODUÇÃO AOS MÉTODOS FACTORIAIS

Análise em R^p

No caso geral os p eixos factoriais obtêm-se resolvendo a equação matricial:

$$\mathbf{T}'\mathbf{T}\mathbf{u}_\alpha - \lambda_\alpha\mathbf{u}_\alpha = 0$$

Valores próprios de $\mathbf{T}'\mathbf{T}$

Vectores próprios de $\mathbf{T}'\mathbf{T}$

INTRODUÇÃO AOS MÉTODOS FACTORIAIS

Análise em R^n

No caso da análise da nuvem em R^n os eixos factoriais obtêm-se resolvendo a equação factorial seguinte:

$$\mathbf{T}\mathbf{T}'\mathbf{v}_\alpha - \mu_\alpha\mathbf{v}_\alpha = 0$$

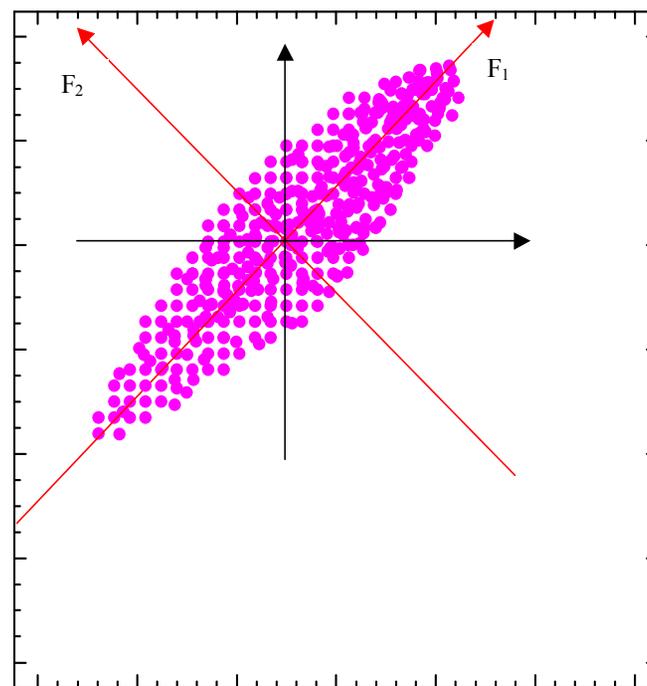
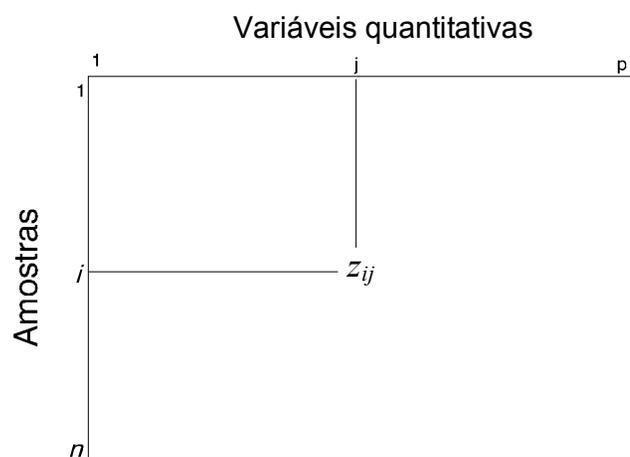
The diagram illustrates the derivation of the factorial equation. It features two red-bordered boxes: 'Valores próprios de $\mathbf{T}\mathbf{T}'$ ' on the left and 'Vectores próprios de $\mathbf{T}\mathbf{T}'$ ' on the right. Red arrows point from each box to the corresponding terms in the equation $\mathbf{T}\mathbf{T}'\mathbf{v}_\alpha - \mu_\alpha\mathbf{v}_\alpha = 0$ above them.

FÓRMULAS DE TRANSIÇÃO

$$\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{T} \mathbf{u}_\alpha$$
$$\mathbf{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{T}' \mathbf{v}_\alpha$$

ANÁLISE EM COMPONENTES PRINCIPAIS

A Análise em Componentes Principais é um método factorial de Análise de Dados, particularmente adaptado à descrição de variáveis quantitativas.



ANÁLISE EM COMPONENTES PRINCIPAIS

CODIFICAÇÃO DOS DADOS

1ª Codificação

$$t_{ij} = \frac{1}{\sqrt{n}} (z_{ij} - \bar{z}_j)$$

\bar{z}_j é a média aritmética dos valores tomados pela variável j .

A matriz de inércia $\mathbf{T}' \mathbf{T}$ é a matriz variância-covariância.

2ª Codificação

$$t_{ij} = \frac{1}{\sqrt{n}} \frac{z_{ij} - \bar{z}_j}{s_j}$$

s_j é o desvio padrão dos valores tomados pela variável j .

A matriz de inércia $\mathbf{T}' \mathbf{T}$ é a matriz de correlação.

ANÁLISE EM COMPONENTES PRINCIPAIS

Efeito em R^p da codificação dos dados

- A transformação $t_{ij} = z_{ij} - \bar{z}_j$ traduz-se numa translação da nuvem, de modo a fazer coincidir o centro de gravidade da nuvem com a origem do sistema de eixos.
- O coeficiente $\frac{1}{\sqrt{n}}$ tem por objectivo fazer coincidir a matriz de inércia com a matriz de variância-covariância (ou correlação).
- O quociente pelo desvio padrão s_j provoca a redução do efeito das variáveis muito dispersas sobre as distâncias entre indivíduos.

$$d^2(i, i') = \sum_{j=1}^p (t_{ij} - t_{i'j})^2 = \frac{1}{\sqrt{n}} \sum_{j=1}^p \left(\frac{z_{ij} - \bar{z}_j}{s_j} \right)^2$$

ANÁLISE EM COMPONENTES PRINCIPAIS

EFEITO EM R^N DA CODIFICAÇÃO DOS DADOS

Distância de cada variável j à origem o

$$d^2(j, o) = \frac{1}{n} \sum_{i=1}^n \left(\frac{z_{ij} - \bar{z}_j}{s_j} \right)^2 = 1$$

Distância entre dois pontos j e j'

$$d^2(j, j') = \frac{1}{n} \sum_{i=1}^n \left(\frac{z_{ij} - \bar{z}_j}{s_j} - \frac{z_{ij'} - \bar{z}_{j'}}{s_{j'}} \right)^2$$

$$d^2(j, j') = 2(1 - r_{jj'})$$

$r_{jj'}$ é o coeficiente de correlação entre as variáveis j e j' .

ANÁLISE EM COMPONENTES PRINCIPAIS

COORDENADAS DOS INDIVÍDUOS NOS EIXOS

$$\mathbf{W} = \mathbf{T}\mathbf{u}$$

Atendendo às fórmulas de transição, as coordenadas dos indivíduos no eixo α são dadas por

$$\mathbf{W}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{T}\mathbf{F}_\alpha$$

COORDENADAS DAS VARIÁVEIS NOS EIXOS

$$\mathbf{F} = \mathbf{T}'\mathbf{v}$$

Atendendo às fórmulas de transição, as coordenadas das variáveis no eixo α são dadas por

$$\mathbf{F}_\alpha = \sqrt{\lambda_\alpha} \mathbf{u}_\alpha$$

$$f_{\alpha j} = r_{\alpha j}$$

$r_{\alpha j}$ é o coeficiente de correlação entre a variável j e o eixo α .

ANÁLISE EM COMPONENTES PRINCIPAIS

REGRAS DE INTERPRETAÇÃO

Inércia total

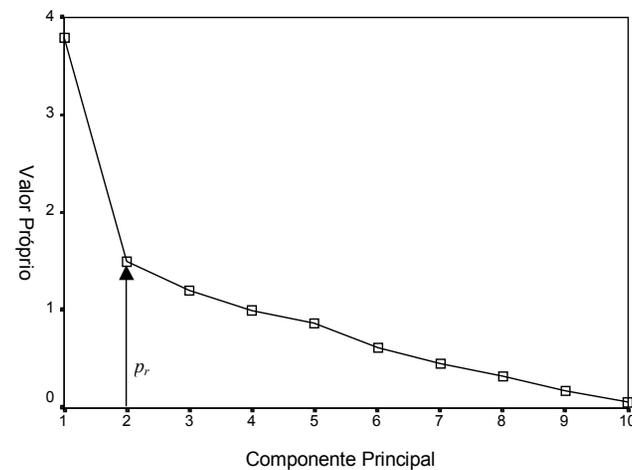
$$I_g = \text{tr}(\mathbf{T}'\mathbf{T}) = \sum_{\alpha=1}^p \lambda_{\alpha} = p$$

Inércia (em %) explicada por cada eixo α

$$I_{\alpha} = 100 \frac{\lambda_{\alpha}}{I_g}$$

CrITÉRIOS para selecção do número de eixos

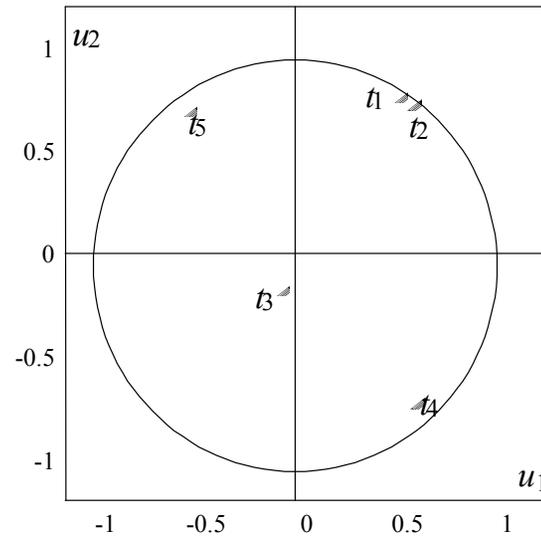
- Escolher os primeiros eixos que expliquem uma percentagem razoável da inércia da nuvem (>70%).
- Desprezar os eixos cujos valores próprios são menores do que 1.
- Desprezar os eixos com números de ordem superiores àquele que inicia a estabilização dos valores próprios do *scree plot*.



ANÁLISE EM COMPONENTES PRINCIPAIS

REGRAS DE INTERPRETAÇÃO

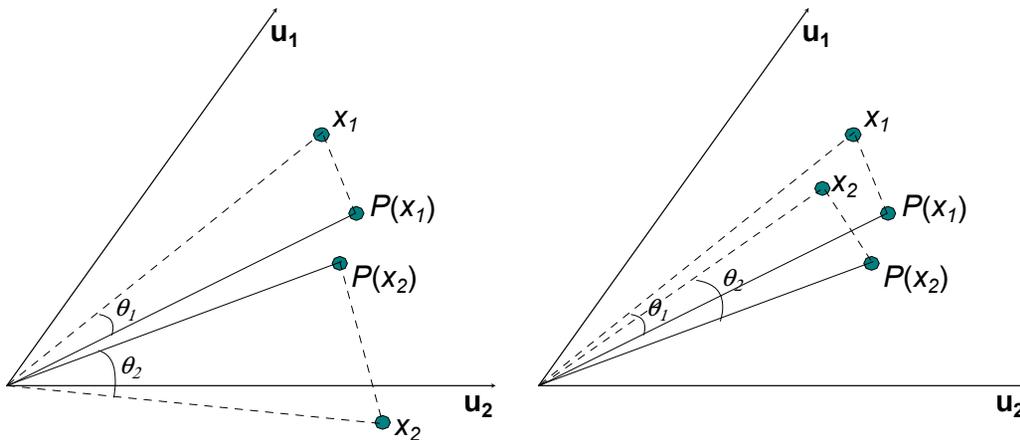
Análise em R^n



ANÁLISE EM COMPONENTES PRINCIPAIS

REGRAS DE INTERPRETAÇÃO

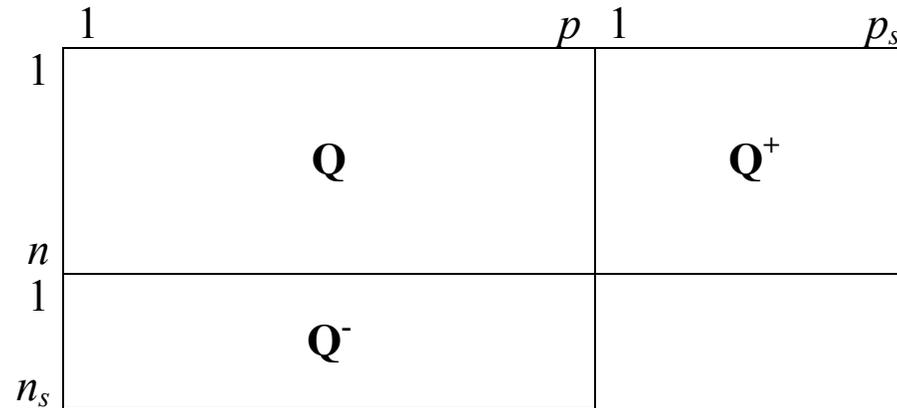
Análise em R^p



O coseno de θ é uma medida da qualidade de representação das amostras no plano factorial

ANÁLISE EM COMPONENTES PRINCIPAIS

INDIVÍDUOS E VARIÁVEIS SUPLEMENTARES



As novas variáveis do quadro \mathbf{Q}^+ ficarão posicionadas sobre a esfera de raio 1 de R^n após a transformação.

$$t_{ij}^+ = \frac{1}{\sqrt{n}} \frac{z_{ij} - \bar{z}_j^+}{s_j^+}$$

As coordenadas destes pontos num eixo α são dadas por:

$$(\mathbf{T}^+)' \mathbf{v}_\alpha$$

As novas linhas do quadro \mathbf{Q}^- devem ser comparáveis às linhas do quadro analisado:

$$t_{ij}^- = \frac{1}{\sqrt{n}} \frac{z_{ij} - \bar{z}_j}{s_j}$$

As coordenadas dos novos pontos linha são dadas por:

$$\mathbf{T}^- \mathbf{u}_\alpha$$

SÍNTESE DO ALGORITMO DE ACP

