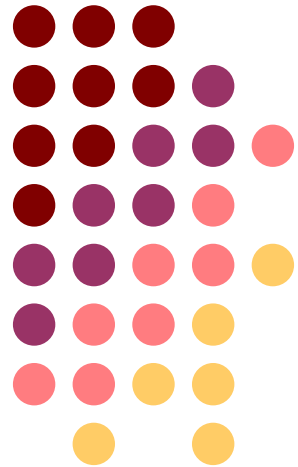
Análise de Componentes Principais Simbólicas

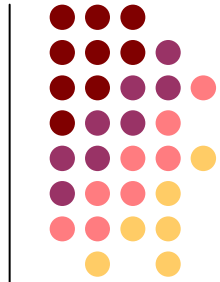
Universidade Federal de Pernambuco

CIn.ufpe.br



Centro de Informática U·F·P·E

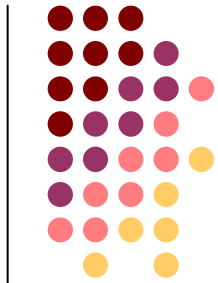
Análise de Componentes Principais



- O objetivo da análise de componentes principais é explicar a estrutura de variância-covariância de um conjunto de variáveis através de um número menor de combinações lineares não-correlacionadas dessas variáveis.

Centro de Informática U·F·P·E

Análise de Componentes Principais



- O objetivo da análise de componentes principais é explicar a estrutura variância-covariância de um conjunto de variáveis através de um número menor de combinações lineares não-correlacionadas dessas variáveis.

ACP para dados clássicos

- Na análise de componentes principais clássicos, temos n pontos $x_1, \ldots, x_n \in \mathbb{R}^n$ no espaço Euclidiano p-dimensional $\Re^n$.
- Ou seja, temos $x_k = (x_{k1}, \ldots, x_{kp})'$ um vetor coluna que descreve as propriedades de um objeto $k \in \Omega = \{1, \ldots, n\}$ em termos de valores $x_{kj} = Y_j(k)$, que foram observados para p variáveis quantitativas Y_j com domínios

$$y_j = \mathbb{R}(j = 1, \ldots, p).$$

Centro de Informática U·F·P·E

ACP para dados clássicos

- Esses dados são agrupados numa tabela de dados clássica $X = (x_{kj})_{n \times p}$
- Para dimensões pequenas (1,2 ou 3) esses pontos podem ser visualizados facilmente na reta real, no plano cartesiano ou no espaço, respectivamente.
- Mas para dimensões maiores nós enfrentamos o problema de como visualizar pontos de dimensões maiores com uma configuração de pontos de baixa dimensão s = 2 ou 3.

ACP para dados clássicos

- A análise de componentes principais clássica resolve esse problema assim:
- 1 – Seleciona-se uma dimensão adequada tal que $s << p$ (usualmente $s = 2$ ou 3)
- 2 – Considere um hiperplano s-dimensional H e, então, os pontos $x_1, \ldots, x_n$ são projetados nesse hiperplano ortogonalmente, sendo: $z_1 = \pi_H(x_1), \ldots, z_n = \pi_H(x_n) \in H$

Centro de Informática U·F·P·E

ACP para dados clássicos

- 3 – Seleciona-se o hiperplano H de maneira **ótima**, ou seja, minimizando a medida de aproximação ou distorção:

$$Q(H) := \sum_{k=1}^{n} \|x_k - \pi_H(x_k)\|^2 \rightarrow \min_H$$

com respeito a todas as escolhas do plano s-dimensional H.

ACP para dados clássicos

- 4 – Seja H* o hiperplano escolhido. Então os pontos projetados pertencerão a esse plano e serão uma representação ótima de dimensão menor dos dados originais.
- 5 – Essa configuração de dimensão maior é visualizada exatamente pelos pontos correspondentes no espaço de dimensão menor, que é chamado de espaço de fatores.

Otimização

- A solução do problema de otimização do passo 3 é encontrada em 4 outros passos:
  - Determinar os centróides $\bar{x} := (\sum_{i=1}^{n} x_i)/n$ e a matriz de espalhamento *p x p*:

$$S := \sum_{k=1}^{n} (x_{kj} - \bar{x}_j)(x_{kj'} - \bar{x}_{j'})$$

dos n pontos, que contém na soma de sua diagonal principal a inércia $\sum_{k=1}^{n} \|x_k - \bar{x}\|^2$ dos dados.

Centro de Informática U·F·P·E

Otimização

- Calcular os auto-valores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ e seus auto-vetores $v_1, v_2, \ldots, v_n \in \mathbb{R}^p$ correspondentes ortonormalizados da matriz S.
- Calcular os valores dos *s* componentes principais:

$$z^*_{kv} = v'_v(x_k - \bar{x}) para\, v = 1, \ldots, s$$

para cada ponto $x_k \in \mathbb{R}^p$.

Otimização

- Então, a visualização ótima $z_1^*, \ldots, z_n^* \in \mathbb{R}^s$ é dada pelos pontos:

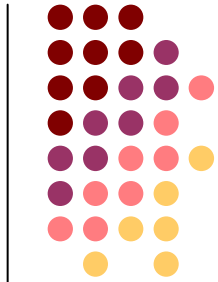
$$z_k^* := \begin{pmatrix} z_{k1}^* \\ \vdots \\ z_{ks}^* \end{pmatrix} = V_s'(x_k - \bar{x}) \quad para\ k = 1, \ldots, n$$

onde $V_s := (v_1, \ldots, v_s)$ é a matriz p x s com colunas

$$v_1, \ldots, v_s \in \mathbb{R}^p$$

Centro de Informática U·F·P·E

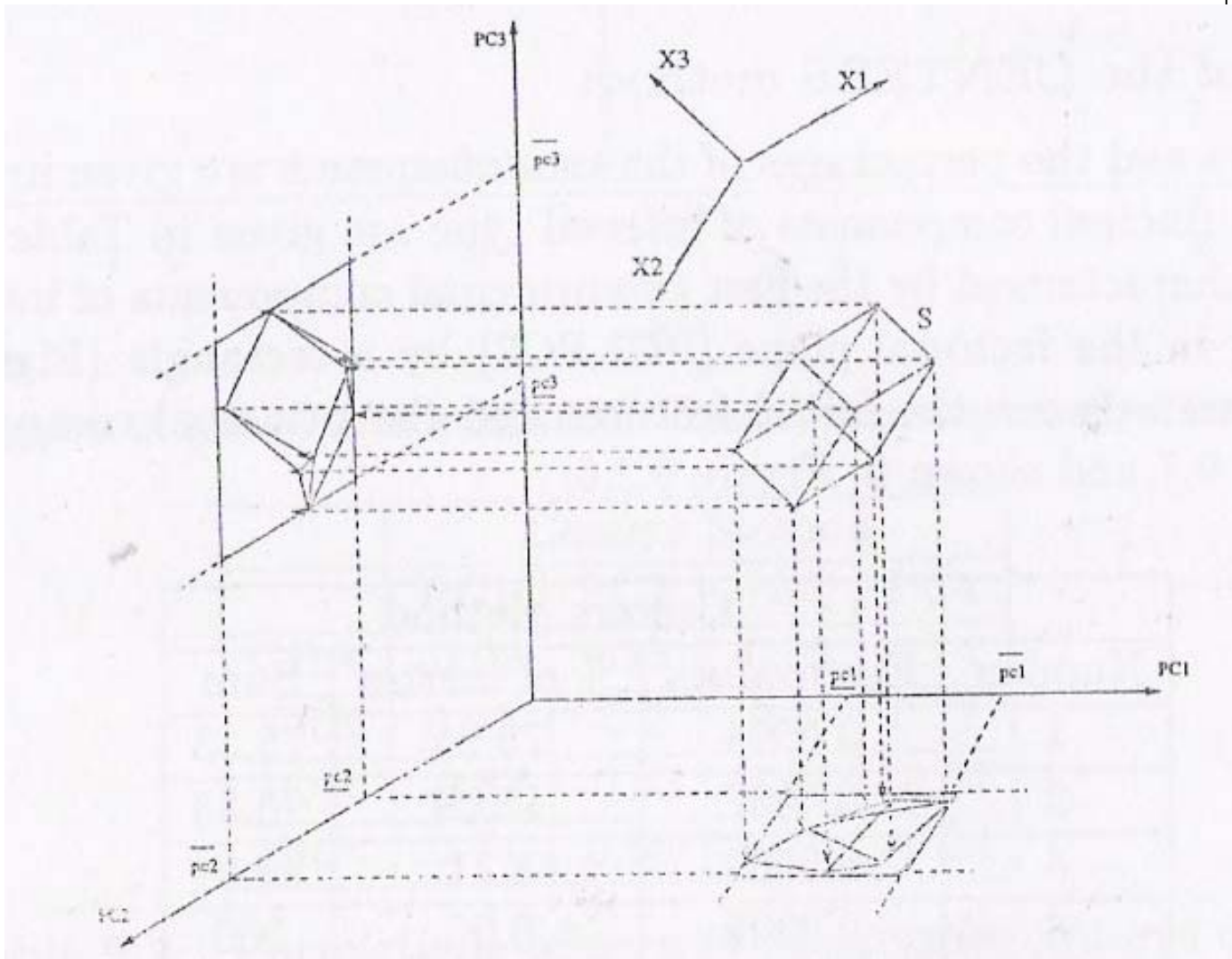
Análise de Componentes Principais Simbólicos



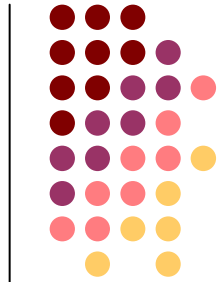
- A análise de componentes principais simbólicos visa descrever objetos i e os dados x_i para um numero reduzido $s < p$ de novas características intervalares, chamados principais componentes intervalares.
- Vamos observar agora a extensão de ACP para dados simbólicos de natureza intervalar
- Dois métodos serão apresentados:
  - Método dos Vértices
  - Método dos Centros

Centro de Informática U·F·P·E

Exemplo gráfico



Análise de Componentes Principais Simbólicos



- No caso intervalar temos n objetos, também descritos por *p* características $Y_1, \ldots, Y_p$ de tipo intervalar.
- Então Y_j têm seus valores no domínio $\mathcal{B}_j = \Im$, o conjunto de todos os intervalos fechados do espaço de observação $y_j = \mathbb{R}$. Se $\xi_{ij} = [x_{ij}^l, x_{ij}^u]$ é o intervalo dos possíveis valores da característica *j* para o objeto *i*.

Centro de Informática U·F·P·E

Análise de Componentes Principais Simbólicos

- Resultando na matriz dada por:

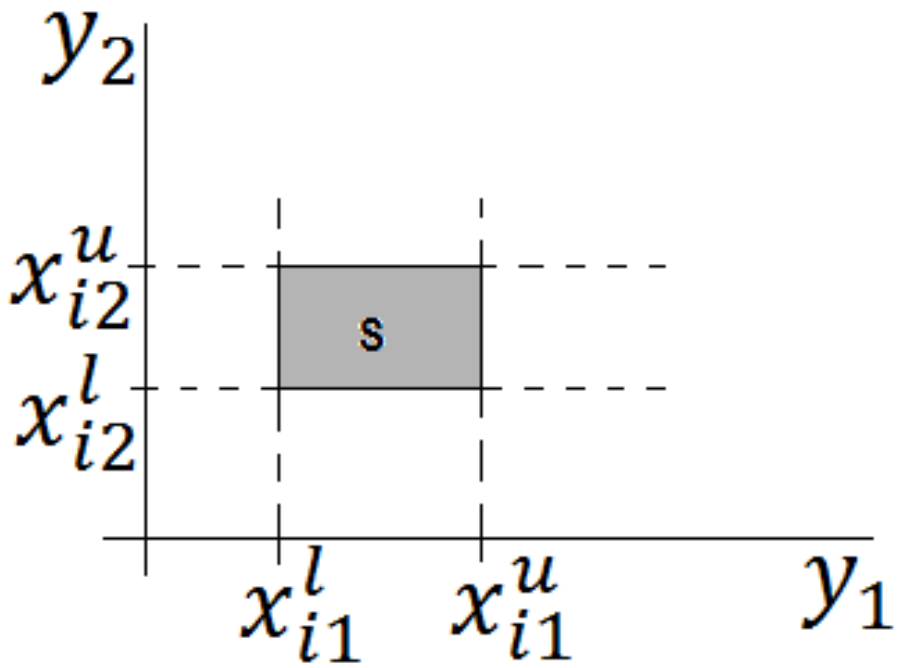
$$\underline{X} = \begin{pmatrix} x_1' \\ \vdots \\ x_m' \end{pmatrix} = \begin{pmatrix} \xi_{11} & \cdots & \xi_{1p} \\ \vdots & \ddots & \vdots \\ \xi_{n1} & \cdots & \xi_{np} \end{pmatrix}$$

Análise de Componentes Principais Simbólicos

- Seja $x_i' = (\xi_{i1}, \ldots, \xi_{ip}) = ([x_{i1}^l, \ldots, x_{i1}^u], \ldots, [x_{ip}^l, \ldots, x_{ip}^u])$ denota o vetor de dados simbólicos para o objeto *i*.
- Esse ponto pode ser visualizado no espaço de descrições $\mathbb{R}^p$ por um hiper-retângulo R_i com 2^p vértices.

Centro de Informática U·F·P·E

Exemplo de visualização para o caso p=2



Descrição

- Um hiper-retângulo no espaço p-dimensional pode ser descrito por uma matriz com 2^p linhas e p colunas onde cada linha contém as coordenadas de um vértice do hiper-retângulo no $\mathbb{R}^p$.

- Por exemplo, para p=2:

$$M_i := \begin{bmatrix} x_{i1}^l & x_{i2}^l \\ x_{i1}^l & x_{i2}^u \\ x_{i1}^u & x_{i2}^l \\ x_{i1}^u & x_{i2}^u \end{bmatrix}$$

O método dos vértices (algoritmo)

- 1 – Descreve-se cada vetor de dados de tipo intervalo x_i por uma matriz de dados numéricos M_i com 2^p linhas e p colunas, contendo os vértices de cada hiper-retângulo.
- 2 – Todas as matrizes são agrupadas numa nova matriz M com $n \times 2^p$ linhas e p colunas dadas por:

$$\underline{X} = \begin{pmatrix} x_i \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} [x_{11}^l, x_{11}^u] & \cdots & [x_{1p}^l, x_{1p}^u] \\ \vdots & \ddots & \vdots \\ [x_{n1}^l, x_{n1}^u] & \cdots & [x_{np}^l, x_{np}^u] \end{pmatrix}$$

O método dos vértices (algoritmo)

onde cada componente é um intervalo, na seguinte matriz numérica M:

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_n \end{pmatrix} = \left(\begin{bmatrix} x_{11}^l & \cdots & x_{1p}^l \\ \vdots & \ddots & \vdots \\ x_{11}^u & \cdots & x_{1p}^u \end{bmatrix} \right.$$
$$\vdots$$
$$\left. \begin{bmatrix} x_{n1}^l & \cdots & x_{n1}^l \\ \vdots & \ddots & \vdots \\ x_{n1}^u & \cdots & x_{n1}^u \end{bmatrix} \right)$$

O método dos vértices (algoritmo)

- 3 – Aplica-se o método clássico de ACP em todas as linhas da nova matriz M, com a escolha de uma dimensão aceitável $s \leq p$ do espaço de visualização $\mathbb{R}^s$. Sendo $Y_1^*, \ldots, Y_s^*$ os *s* primeiros componentes principais "numéricos" e $\lambda_1 \geq \cdots \geq \lambda_s \geq 0$ seus auto-valores associados.

Centro de Informática U·F·P·E

O método dos vértices (algoritmo)

- 4 – Os componentes principais de tipo intervalar $Y_1^I, \ldots, Y_s^I$ são construídos através dos componentes principais $Y_1^*, \ldots, Y_s^*$ "numéricos" como a seguir:
  - Seja L_i o conjunto de índices de linha k na matriz M que se referem aos vértices do hipercubo R_i correspondendo ao i-ésimo vetor de dados simbólicos x_i.

O método dos vértices (algoritmo)

- Para $k \in L_i$, seja y_{kv} o valor do componente principal numérico Y_v^* para o vértice de R_i com índice de linha k.
- O valor do componente principal de tipo intervalo Y_v^I para o i-ésimo objeto é caracterizado por:

$$y_{iv} = [y_{iv}^l, y_{iv}^u]$$

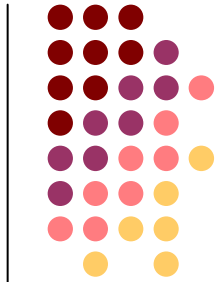
onde :

$$y_{iv}^l = \min_{k \in L_i} (y_{kv})$$

$$y_{iv}^u = \max_{k \in L_i} (y_{kv})$$

Centro de Informática U·F·P·E

Parâmetros de interpretação



- A visualização que é retornada por uma ACP clássica é normalmente justificada pelo cálculo de vários coeficientes que medem a qualidade da representação e a contribuição de cada fator para o diagrama resultado.
- Esses parâmetros de interpretação são facilmente estendidos para o caso simbólico.

Qualidade da representação

- Para medir a qualidade d representação do vetor x_i com respeito ao j-ésimo eixo fatorial $v_j \in \mathbb{R}^p$, são propostos os seguintes coeficientes:

$$COR_I^1(i, v_j) = \frac{\sum_{k \in L_i} q_k \cdot y_{kj}^2}{\sum_{k \in L_i} q_k . d^2(k, G)} = \frac{\sum_{k \in L_i} y_{kj}^2}{\sum_{k \in L_i} d^2(k, G)}$$

- Onde $G \in \mathbb{R}^p$ é o centróide de todas as linhas da matriz M e *d(k,G)* é a distância euclidiana entre a linha $k \in L_i$ de M e G.

Centro de Informática U·F·P·E

Medidas de contribuição de x_i

- A contribuição de x_i para a variância λ_j do j-ésimo componente:

$$CTR_I(i, v_j) = \frac{\sum_{k \in L_i} q_k \cdot y_{kj}^2}{\lambda_j} = \frac{p_i}{2^p . \lambda_j} \cdot \sum_{k \in L_i} y_{kj}^2$$

- Indica a contribuição dos 2^p vértices pertencentes ao hiper-retângulo R_i à soma total dos quadrados.
- A contribuição de x_i para a soma total dos quadrados de todos o vértices representando n hiper-retângulos

$$INR_I(i) = \frac{\sum_{k \in L_i} q_k . d^2(k, G)}{I_T} = \frac{p_i}{2^p} \cdot \frac{\sum_{k \in L_i} d^2(k, G)}{\sum_{j=1}^{p} \lambda_j}$$

O método dos centros

- O método dos vértices envolve muitos cálculos quando o número de características é grande.
- Nesse caso, será proposto outro método que aplica ACP clássica aos centros dos *n* hiper-retângulos R_i para encontrar os eixos fatoriais.
- A variação da imprecisão não poderá ser visualizada do resultado da ACP, mas deverá ser estimada da variabilidade variação ou imprecisão das características descritivas.

O método dos centros

- A matriz n x p contendo os centros dos hiper-retângulos é dada por:

$$\underset{\sim}{X} = \begin{pmatrix} x_{11}^{c} & \cdots & x_{1p}^{c} \\ \vdots & \ddots & \vdots \\ x_{np}^{c} & \cdots & x_{np}^{c} \end{pmatrix}$$

- Onde as coordenadas do i-ésimo centro $c_i = (x_{i1}^{c}, \ldots, x_{ip}^{c})'$ são resultado de:

$$x_{ij}^{c} = \frac{x_{ij}^{u} + x_{ij}^{l}}{2}, j = 1, \ldots, p; i = 1, \ldots, n$$

Centro de Informática U·F·P·E

O método dos centros (algoritmo)

- 1 – Transforma-se a matriz de dados $\underset{\sim}{X}$ na matriz $\underset{\sim}{X}$, encontrando os centros através do cálculo mostrado no slide anterior. Denota-se por $Y_1^c, \ldots, Y_p^c$ os novos valores reais das características descritoras.
- 2 – Aplica-se o método ACP clássico para a nova matriz $\underset{\sim}{X}$ dos centros c_i obtidos no primeiro passo.

Centro de Informática U·F·P·E

O método dos centros (algoritmo)

- 3 – Determine para cada objeto *i* o seus valores de componentes principais intervalares como o seguinte:
  - Se $\overline{x_j^c} := (\sum_{i=1}^{n} x_{ij}^c)/n$ é a média da característica Y_j^c (os valores da j-ésima coluna da matriz $\underset{\sim}{X}$), o v-ésimo componente principal do centro c_i é dado por:

$$y_{iv}^c = \sum_{j=1}^{p} \left(x_{ij}^c - \overline{x_j^c}\right) \cdot v_{jv}$$

$$v_{jv} = (v_{1v}, \dots, v_{pv})$$

  - Onde é o v-ésimo auto-vetor de S.

O método dos centros

- Dado que as coordenadas do i-ésimo centro x_{ij}^{c} estão localizadas entre os seus limites inferior e superior (x_{ij}^{l} e x_{ij}^{u}), é possível encontrar um intervalo $[y_{ik}^{l}, y_{ik}^{u}]$ em que possíveis valores do v-ésimo componente principal y_{iv}^{c} de c_i devem ser localizados.
- Dado que os componentes principais são funções lineares do dado central x_{ij}^{c}, nós obtemos os limites para os v-ésimos principal componentes para o objeto i.

Centro de Informática U·F·P·E

Limites dos v-ésimos componentes principais do objeto i

$$y_{iv}^{l} = \sum_{j=1}^{p} min_{x_{ij}^{l} \leq x_{ij}^{r} \leq x_{ij}^{u}} \left(x_{ij}^{r} - \overline{x_{j}^{c}} \right) \cdot v_{jv}$$

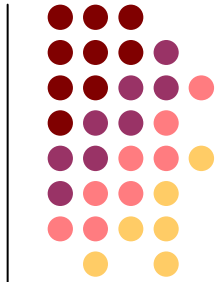
$$y_{iv}^{u} = \sum_{j=1}^{p} max_{x_{ij}^{l} \leq x_{ij}^{r} \leq x_{ij}^{u}} \left(x_{ij}^{r} - \overline{x_{j}^{c}} \right) \cdot v_{jv}$$

Centro de Informática U·F·P·E

Exemplo de óleos e gordura

- Para ilustrar os métodos propostos vamos utilizar o conjunto de dados de Ichino, reproduzido na tabela do próximo slide, ele consiste de uma classe de óleos descrita por p=4 características quantitativas de tipo intervalar: “Gravidade Específica”, “Ponto de Congelamento”, “Valor de Iodo” e “Saponificação”.

Tabela do conjunto de dados



Nome	GRA	CONG	IOD	SAP
Linseed	[0.93,0.94]	[-27.00,-18.00]	[170.00,204.00]	[118.00,196.00]
Perilla	[0.93,0.94]	[-5.00,-4.00]	[192.00,208.00]	[188.00,197.00]
Cotton	[0.92,0.92]	[-6.00,-1.00]	[99.00,113.00]	[189.00,198.00]
Sesame	[0.92,0.93]	[-6.00,-4.00]	[104.00,116.00]	[187.00,193.00]
Camellia	[0.92,0.92]	[-21.00,-15.00]	[80.00,82.00]	[189.00,193.00]
Olive	[0.91,0.92]	[0.00,6.00]	[79.00,90.00]	[187.00,196.00]
Beef	[0.86,0.87]	[30.00,38.00]	[40.00,48.00]	[190.00,199.00]
Hog	[0.86,0.86]	[22.00,32.00]	[53.00,77.00]	[190.00,202.0

Resultados do método dos vértices

- Auto-valores e inércia

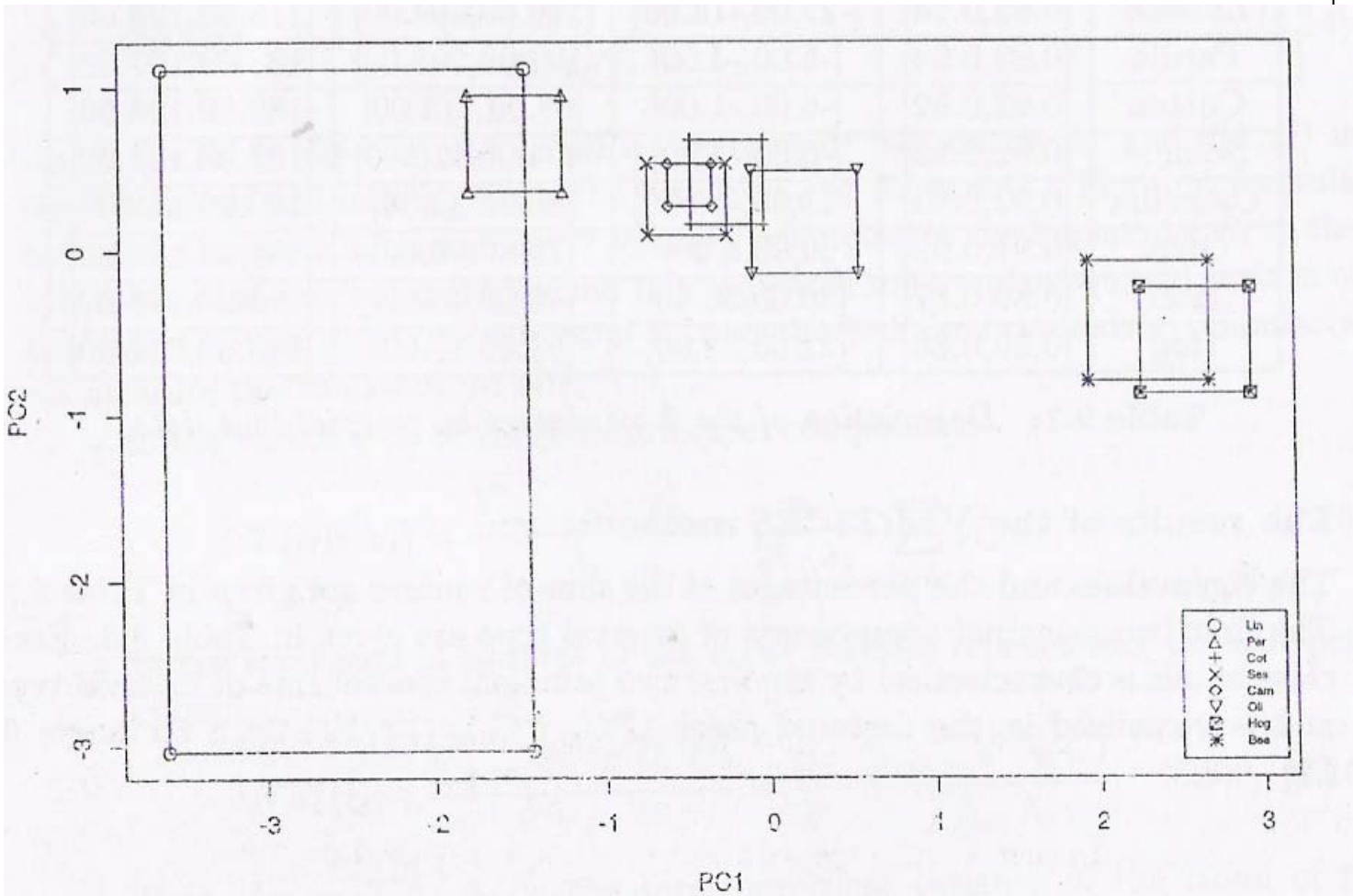
Método dos Vértices			
Número	Auto-valores	% de inércia	Somatório
1	2.7316	68.29	68.29
2	0.8093	20.23	88.52
3	0.3801	9.5	98.02
4	0.0790	1.98	100

Resultados do método dos vértices

- Os dois primeiros componentes principais de tipo intervalo

	Método dos Vértices	
Rótulo	PC1	PC2
L	[-3.58,-1.43]	[-3.04,1.10]
P	[-1.76,1.22]	[0.36,0.95]
Co	[-0.45,-0.01]	[0.16,0.67]
S	[-0.71,-0.23]	[0.09,0.53]
Ca	[-0.58,-0.32]	[0.27,0.53]
O	[-0.09,0.56]	[-0.14,0.49]
B	[2.26,2.93]	[-0.87,-0.23]
H	[1.95,2.68]	[-0.80,-0.07]

Resultado do método dos vértices (Representação retangular)



Centro de Informática U·F·P·E

Resultados do método dos centros

- Auto-valores e inércia

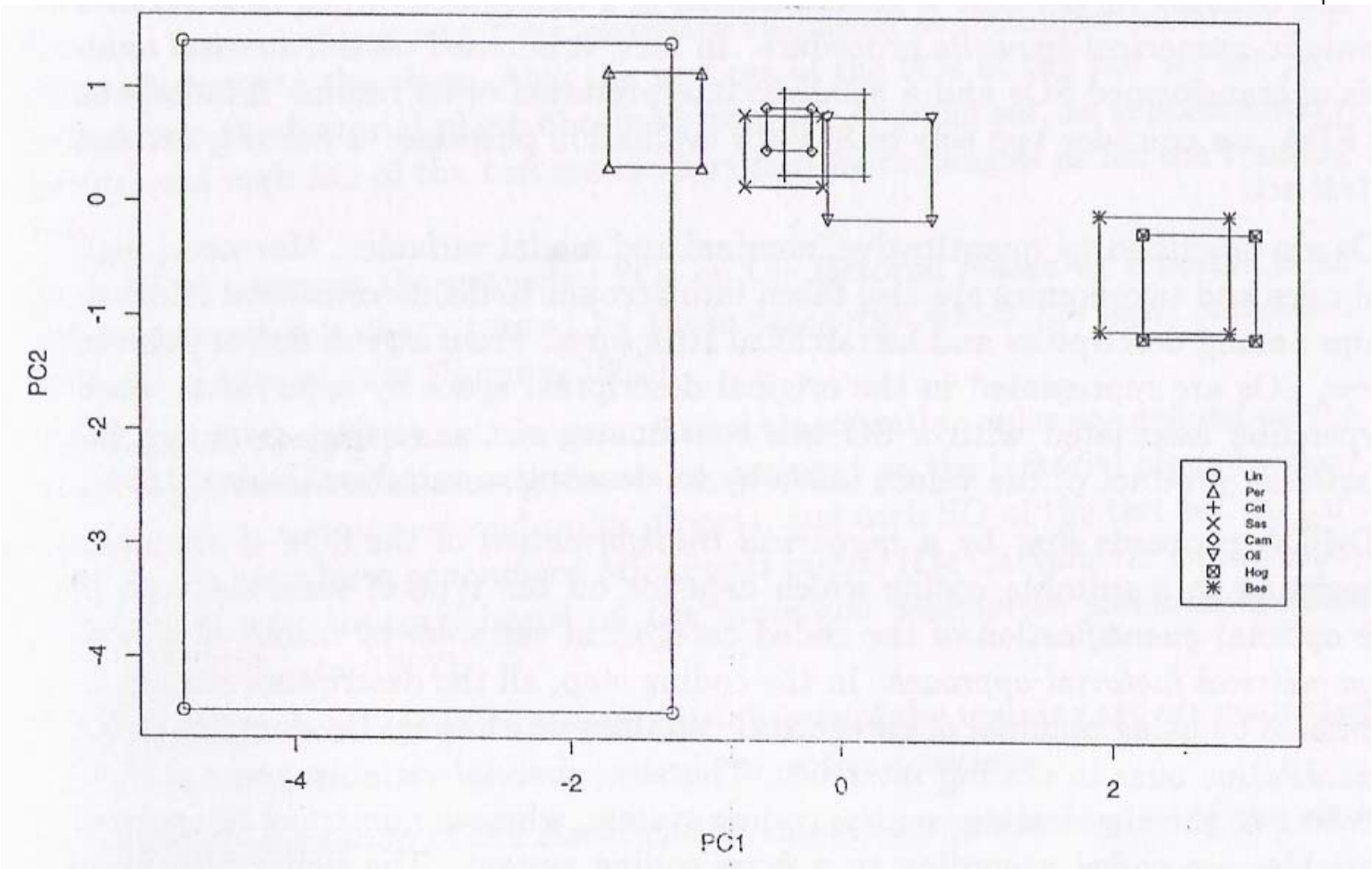
Método dos Centros			
Número	Auto-valores	% de inércia	Somatório
1	3.0094	75.24	75.24
2	0.6037	15.09	90.33
3	0.3483	8.71	99.04
4	0.0386	0.96	100

Resultados do método dos centros

- Os dois primeiros componentes principais de tipo intervalo

	Método dos Centros	
Rótulo	PC1	PC2
L	[-4.80,-1.25]	[-4.46,1.40]
P	[-1.72,-1.03]	[0.32,1.15]
Co	[-0.42,0.18]	[0.26,0.98]
S	[-0.70,-0.13]	[0.15,0.78]
Ca	[-0.55,-0.21]	[0.48,0.85]
O	[-0.09,0.69]	[-0.13,0.77]
B	[2.23,3.04]	[-1.15,-0.23]
H	[1.91,2.85]	[-1.09,-0.07]

Resultado do método dos centros (Representação retangular)



Centro de Informática U·F·P·E