MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2019 - IME

Análise Multivariada

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}$$

Já vimos ☺

Matriz de Dados: Estatísticas descritivas multivariadas em $\Re^{p}$, $\Re^{p\times p}$ e $\Re^{n\times n}$

Episóides de Concentração, Boxplot Bivariado, Espaços Duais

Matriz Aleatória: Distribuição Normal Multivariada, Distribuições Amostrais

Testes de Hipóteses Multivariadas para μ e Σ:

Caso de Uma, Duas e Muitas Populações N_p (MANOVA)

Regiões de Confiança, I.C. Simultâneos, Correções para Múltiplos testes

Decomposições de $Y_{n\times p}$

Técnicas de Redução de Dimensionalidade: $\Re^{p} \rightarrow \Re^{m}; \quad m \leq p$

Caso n>p, observações *iid* (soluções clássicas)

Caso n>p, observações estruturadas (correlacionadas)

Caso n<<p (Big-p): redução de dimensionalidade, integração e predição

Caso n>>p (Big-n): $\Re^{n} = \cup_{g=1}^{G} \Re^{n_g}$

Redução de Dimensionalidade em $\Re^p$

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$$Y_{n\times p};\quad n > p$$

$$Y_{ig_{p\times 1}} \sim \left(\mu_g;\Sigma_g\right)\ \ \Re^p \rightarrow \Re^m, m < p$$

Como veremos, a Redução de Dimensionalidade depende da estrutura dos dados!

Estrutura dos Dados:

$\mu_{g_{p\times 1}}$?

$\mu_g = \mu$

$\Sigma_{g_{p\times p}}$?

$\Sigma_g = \Sigma$

$i = 1,...,n_g;\ \ g = 1,...,G$?

iid

Redução de Dimensionalidade em $\Re^p$
Quociente de Rayleigh

Seja M uma matriz simétrica em $\Re^{pxp}$, com autovalores $\lambda_1 \geq \lambda_2 \geq ... \lambda_p$ e os correspondentes autovetores $V_1, V_2, ..., V_p$. Então:

$$\max_{\|a\|=1} a'Ma = \max_{a \neq 0} \frac{a'Ma}{a'a} = \lambda_1; \quad a = V_1 \in \Re^p$$

$$\min_{\|a\|=1} a'Ma = \min_{a \neq 0} \frac{a'Ma}{a'a} = \lambda_p; \quad a = V_p \in \Re^p$$

Muitos problemas de redução de dimensionalidade recaem na otimização de formas quadráticas (em $\Re^{p \times p}$) cuja solução está na teoria de decomposição espectral de matrizes simétricas. Veremos algumas destas soluções no espaço Dual: $\Re^{p \times p}$; $\Re^{n \times n}$; $\Re^{n \times p}$

Ténicas Multivariadas de Redução de Dimensionalidade

Como obter vetores reducionistas de dados?

Estrutura dos Dados
Objetivo da análise
Soluções (e Restrições impostas)
Representação Gráfica – BiPlot, Dendrograma, HeatMap

- Análise de Componentes Principais: $Y_{n\times p} \Rightarrow \Re^{p\times p}$
- Escalonamento Multidimensional: $(\Re^{n\times n})$
- Análise de Correspondência: $Y^{I\times J} \Rightarrow [0,1]^{(I-1)\times(J-1)}; D^{(I-1)\times(I-1)}; D^{(J-1)\times(J-1)}$
- Análise Fatorial: $Y_{n\times p} \Rightarrow \Re^{p\times p}$
- Análise Discriminante $Y_{n\times(p+1)} \Rightarrow \Re^{p\times p}$
- Análise de Agrupamento: $Y_{n\times p} \Rightarrow \Re^{n\times n}$
- Análise de Correlação Canônica: $Y_{n\times(p+q)} \Rightarrow \Re^{p\times q}(\Re^{p\times p}, \Re^{q\times q})$

✓ MANOVA: $Y_{n\times p} \Rightarrow \Re^{p\times p}$

Análise de Componentes Principais

Análise Clássica

n > p
Observações iid
(respostas quantitativas)

Análise de Componentes Principais

(Pearson, 1901)

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

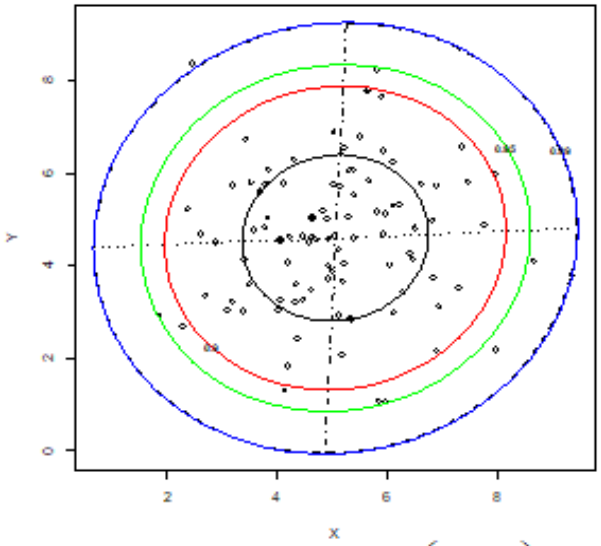
$$Y_{n\times p};\quad n > p \Rightarrow \boxed{Y_{i_{p\times 1}} \overset{iid}{\sim} (\mu;\Sigma)}$$

Premissa: Dados de Uma única População
Observações *iid*
Matriz de covariâncias "válida" ($\Sigma \in \Re^{pxp}$)

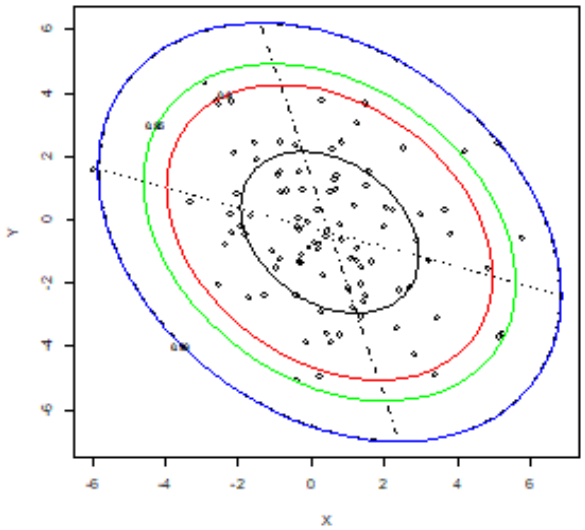
- A variável Y_j pode ser eliminada da análise?
- Como as variáveis podem ser ordenadas segundo sua "importância" na análise?

Considerar a estrutura de Σ

$$\mu' = (5,5) \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

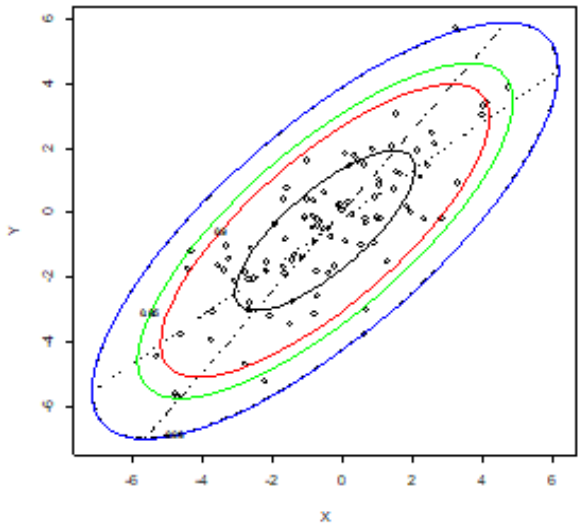


$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$

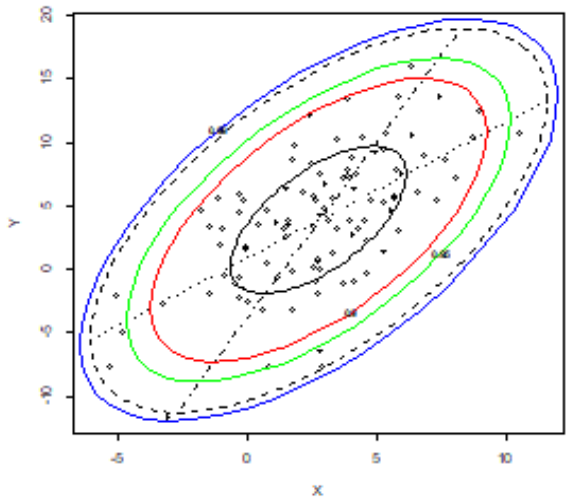


BoxPlot Bivariado

$$\mu' = (0,0) \quad \Sigma = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$$



$$\mu' = (3,4) \quad \Sigma = \begin{pmatrix} 9 & 10 \\ 10 & 25 \end{pmatrix}$$



Como são as correspondentes elipses para as variáveis padronizadas?

$$\Sigma \Leftrightarrow R$$

Análise de Componentes Principais
Estruturas de Σ e R

Como proceder com a redução de dimensionalidade nos seguintes casos?

Estrutura apropriada para a redução: ordenar as variáveis de acordo com a variância e calcular a contribuição para a variância total.

$$\Sigma_1 = \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \dots & 0 & \dots & \dots \\ 0 & 0 & 0 & \sigma_{pp} \end{pmatrix}; \qquad R_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Não há como reduzir a dimensionalidade de espaços formados por variáveis não correlacionadas e homocedásticas

$$\Sigma_2 = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \dots & 0 & \dots & \dots \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{pmatrix}; \qquad R_2 = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{pmatrix} = (1-\rho) I_p + \rho 1_p 1_p'$$

Correlação uniforme.
Se ρ for alto, um único CP deve explicar bem a (co)variância dos dados e ele é uma média ponderada que atribui pesos iguais à todas as variáveis.

$$\Sigma_3 = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ & \sigma_{22} & \dots & \sigma_{2p} \\ \sim & & \dots & \dots \\ & & \dots & \sigma_{pp} \end{pmatrix}; \qquad R_3 = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ & 1 & \dots & \rho_{2p} \\ \sim & & \dots & \dots \\ & & \dots & 1 \end{pmatrix}$$

Análise de Componentes Principais
Motivação

Arquivo HATCO (Hair et al., 2005)

$Y_{100\times 7}$

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45	4,4	0	1	1	2
…														
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1	0	60	5,2	1	0	1	3
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4	1	38	3,7	0	1	0	1
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4	1	42	4,3	0	1	0	1
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0	0	33	4,4	1	0	0	1

Análise de Componentes Principais
Motivação

Arquivo HATCO (Hair et al., 2005)

```
> colMeans(hatco)
   X1    X2    X3    X4    X5    X6   X7
3.515 2.364 7.894 5.248 2.916 2.665 6.971

> cov(hatco)
         X1          X2          X3          X4          X5          X6          X7
X1 1.74431818 -0.5514747  0.93261616  0.07533333  0.60713131  0.07851010 -1.01046970
X2 -0.55147475  1.4296000 -0.80769293  0.36821010  0.46078384  0.17165657  0.89035960
X3  0.93261616 -0.8076929  1.92238788 -0.18213333  0.06938990 -0.03667677 -0.98492323
X4  0.07533333  0.3682101 -0.18213333  1.28009697  0.25387071  0.68745455  0.35867879
X5  0.60713131  0.4607838  0.06938990  0.25387071  0.56438788  0.13945455 -0.06569293
X6  0.07851010  0.1716566 -0.03667677  0.68745455  0.13945455  0.59421717  0.21665152
X7 -1.01046970  0.8903596 -0.98492323  0.35867879 -0.06569293  0.21665152  2.51298889

> cor(hatco)
         X1          X2          X3          X4          X5          X6          X7
X1  1.00000000 -0.3492251  0.50929519  0.0504142  0.61190069  0.07711522 -0.4826309
X2 -0.34922515  1.0000000 -0.48721259  0.2721868  0.51298082  0.18624325  0.4697458
X3  0.50929519 -0.4872126  1.00000000 -0.1161041  0.06661728 -0.03431610 -0.4481120
X4  0.05041420  0.2721868 -0.11610408  1.0000000  0.29867737  0.78822454  0.1999811
X5  0.61190069  0.5129808  0.06661728  0.2986774  1.00000000  0.24080818 -0.0551613
X6  0.07711522  0.1862433 -0.03431610  0.7882245  0.24080818  1.00000000  0.1772939
X7 -0.48263094  0.4697458 -0.44811201  0.1999811 -0.05516130  0.17729392  1.0000000
```

Sugere um padrão não estruturado: p(p+1)/2 parâmetros

Análise de Componentes Principais
Motivação

$Y_{7\times6}$

Cães pré-históricos da Tailândia (Manly, 2005).

Grupo	X1	X2	X3	X4	X5	X6
G1	9.7	21.0	19.4	7.7	32.0	36.5
G2	8.1	16.7	18.3	7	30.3	32.9
G3	13.5	27.3	26.8	10.6	41.9	48.1
G4	11.5	24.3	24.5	9.3	40.0	44.6
G5	10.7	23.5	21.4	8.5	28.8	37.6
G6	9.6	22.6	21.1	8.3	34.4	43.1
Cão Pré-h	10.3	22.1	19.1	8.1	32.2	35.0

Dados dos Cães Pré-históricos

```
> colMeans(caes)
      X1       X2       X3       X4       X5       X6
10.48571 22.50000 21.51429  8.50000 34.22857 39.68571

> cov(caes)
         X1          X2          X3          X4          X5          X6
X1  2.881429  5.251667  4.846905 1.933333  6.527143  7.739762
X2  5.251667 10.556667  8.895000 3.593333 11.456667 15.583333
X3  4.846905  8.895000  9.611429 3.508333 13.427857 16.305238
X4  1.933333  3.593333  3.508333 1.356667  4.863333  5.920000
X5  6.527143 11.456667 13.427857 4.863333 24.362381 24.680476
X6  7.739762 15.583333 16.305238 5.920000 24.680476 31.518095

> cor(caes)
         X1          X2          X3          X4          X5          X6
X1 1.0000000 0.9522036 0.9210148 0.9778365 0.7790392 0.8121639
X2 0.9522036 1.0000000 0.8830567 0.9495056 0.7143894 0.8543129
X3 0.9210148 0.8830567 1.0000000 0.9715615 0.8775116 0.9368136
X4 0.9778365 0.9495056 0.9715615 1.0000000 0.8459362 0.9053263
X5 0.7790392 0.7143894 0.8775116 0.8459362 1.0000000 0.8906636
X6 0.8121639 0.8543129 0.9368136 0.9053263 0.8906636 1.0000000
```

Sugere um padrão de correlação uniforme

Análise de Componentes Principais

$$Y_i \in \Re^p \quad \rightarrow \quad Z_i = A_{p\times p}\, Y_{i_{p\times 1}} \in \Re^p$$

$$Cov(Y_i) = \Sigma_{p\times p} \qquad Cov(Z_i) = \Lambda = Diag(\lambda_j)$$

$$tr\ \Sigma = tr\Lambda$$

Transformação que preserva a variância total

Realizar uma transformação linear de Y com boas propriedades, como a de preservar a variância total

$$\left|\Sigma - \lambda I_p\right| = 0; \quad \Sigma P_j = \lambda_j P_j \qquad \Sigma = P\Lambda P' \quad ; \quad PP' = P'P = I \quad \Lambda = diag(\lambda_j)$$

$$tr\ \Sigma = tr\left(P\Lambda P'\right) = \sum_{j=1}^{p} \lambda_j P_j{}' P_j = \sum_{j=1}^{p} \lambda_j = tr\Lambda$$

$$\rightarrow \quad A = P'; \quad Z_{ij} = P_j'\, Y_{i_{p\times 1}} \in \Re$$

Aproximação para Σ ($\in \Re^{pxp}$) em $\Re^{mxm}$ (p<m)

$$\Sigma \ = \sum_{j=1}^{p} \lambda_j P_j P_j{}' \cong \sum_{j=1}^{m} \lambda_j P_j P_j{}';$$

$$Z_i = \begin{pmatrix} Z_{i1} \\ \dots \\ Z_{im} \end{pmatrix} = P'_{m\times p}\, Y_{i_{p\times 1}} \in \Re^m$$

Var(a'Y)

$$\arg\max_{\|a\|=1} \frac{a\ '\Sigma a}{a\ 'a} = P_1; \quad \max \frac{P_1{}'\Sigma P_1}{P_1{}'P_1} = \lambda_1; \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq \dots \geq \lambda_p$$

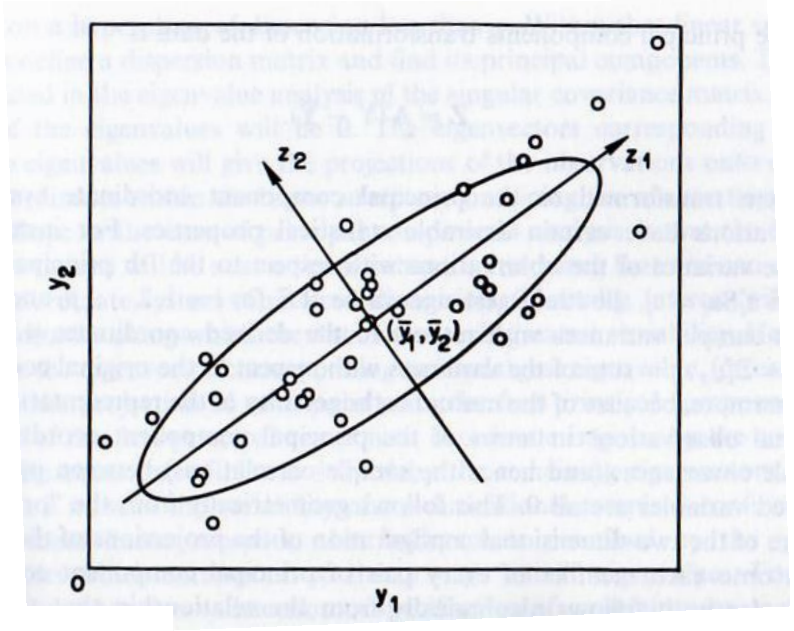
$$Y_i \in \Re^p \rightarrow Z_i = P'_{m\times p}\, Y_{i_{p\times 1}} \in \Re^m \quad \Rightarrow Z_{ki} = P_k'\, Y_{p\times 1}; \quad Var(Z_{ki}) = \lambda_k$$

Análise de Componentes Principais

Técnica de Redução Linear de Dimensionalidade de Variáveis

$(y-\bar{y})'\Sigma^{-1}(y-\bar{y})=c^2$ define

uma família de elipsóides



Transformação que preserva a variância total (Rotação ortogonal dos Eixos)

$$Y \quad \Rightarrow Z = AY$$

$$(y_1, y_2) \Rightarrow (z_1, z_2)$$

Z_1: primeiro componente principal

Z_2: segundo componente principal

$$Z_1 = a_1'Y \; ; \quad V(a_1'Y) = a_1'\Sigma\, a_1$$

$$Z_2 = a_2'Y \; ; \quad V(a_2'Y) = a_2'\Sigma\, a_2$$

$$V(a_1'Y) \geq V(a_2'Y)$$

$$Cov(Z_1, Z_2) = a_1'\Sigma\, a_2 = 0$$

Decomposição espectral de Σ (autovalores e autovetores) permite uma representação dos dados em eixos ortogonais e nas direções de máxima variação (total) dos dados.

Análise de Componentes Principais

Exemplo 1: $\Sigma = \sigma^2 I$; $\Sigma = P\Lambda P' \Rightarrow P = I$; $\Sigma P_j = \sigma^2 P_j$ σ^2 é autovalor com multiplicidade p.

$Z_{ji} = P_j' Y_i = Y_{ij}$ Não é possível reduzir nem ordenar as variáveis.

Exemplo 2: $\Sigma = diag(\sigma_{jj})$; $\Sigma = P\Lambda P' \Rightarrow P = I$; $\Sigma P_j = \sigma_{jj} P_j$

$Z_{ji} = P_j' Y_i = Y_{i(j)}$; $(\sigma_{jj}; P_j)$ Os CP são as variáveis originais ordenadas.

Exemplo 3: $\Sigma = (1-\rho)I + \rho 11'; \rho > 0$; $\Sigma = P\Lambda P' \Rightarrow \lambda_1 = 1+(p-1)\rho$; $P_1 = 1/\sqrt{p}1_p$

$\lambda_2 = ... = \lambda_p = 1-\rho$

$Z_{1i} = P_1' Y_i = \sum_{j=1}^{p} \frac{Y_{ij}}{\sqrt{p}}$ CP1 é um "índice" com pesos iguais, e de norma 1, para todas as variáveis

$$\%VarExpl = \frac{\lambda_1}{p} = \frac{1+(p-1)\rho}{p} = \rho + \frac{1-\rho}{p} \cong \rho \text{ se } \rho \to 1 \text{ ou p} \to \infty$$

Componentes Principais
Quantos Componentes Reter na Análise?

$$Y_i \in \Re^p \quad \rightarrow \quad Z_i = P'_{m\times p}\, Y_{i_{p\times 1}} \in \Re^m \qquad m?$$

- Preservar "grande" parte da variância total dos dados:

Para variáveis padronizadas: $\lambda_j \geq 1$

$$\frac{\lambda_1 + \lambda_2 \ldots + \lambda_m}{tr\Sigma} \geq ? \quad (0{,}70)$$

Devem ser retidos todos os CPj, com variância maior que a média:

$$\lambda_j \geq \frac{tr\Sigma}{p}$$

Critério de corte no *ScreePlot*: quando a variação entre os λs passa a ser pequena (*cotovelo*)

- Garantir Correlações "Altas" entre as variáveis Originais e as CP:

$$Y_{ij} = l_j{}' Y_i \quad ; l_j = (0,\ldots,0,1,0,\ldots,0)'$$

$$Z_{ki} = P_k{}' Y_i ; P_k = (a_{1k},\ldots,a_{pk})'$$

$$Cov(Y_{ij}, Z_{ki}) = Cov(l_j{}' Y_i, P_k{}' Y_i) = l_j{}' \Sigma P_k = \lambda_k a_{jk}$$

$$Corr(Y_{ij}, Z_{ki}) = \frac{Cov(Y_{ij}, Z_{ki})}{\sqrt{Var(Y_{ji})}\sqrt{Var(Z_{ki})}} = \frac{a_{jk}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}}$$

- Garantir "grande" parte da variabilidade de cada variável original:

$$Var(Y_{ij}) = Var(l'_j Y_i) = Var(l'_j P_{p\times m} Z_i) = \sum_{k=1}^{m} a_{jk}^2\, Var(Z_{ki}) = \sum_{k=1}^{m} a_{jk}^2\, \lambda_k \quad \Rightarrow \quad \sum_{k=1}^{m} \frac{a_{jk}^2\, \lambda_k}{\sigma_{jj}}$$

Análise de Componentes Principais

Obtenção dos Componentes Principais: Dados HATCO

Estatísticas Descritivas

	Variáveis						
	Y1	Y2	Y3	Y4	Y5	Y6	Y7
n	100	100	100	100	100	100	100
Média	3,515	2,364	7,894	5,248	2,916	2,665	6,971
Variância	1,74432	1,4296	1,92239	1,2801	0,564388	0,594217	2,51299

Matriz de Covariância Amostral

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Y1	1,74432	-0,55147	0,93262	0,07533	0,607131	0,07851	-1,01047
Y2	-0,55147	1,4296	-0,80769	0,36821	0,460784	0,171657	0,89036
Y3	0,93262	-0,80769	1,92239	-0,18213	0,06939	-0,036677	-0,98492
Y4	0,07533	0,36821	-0,18213	1,2801	0,253871	0,687455	0,35868
Y5	0,60713	0,46078	0,06939	0,25387	0,564388	0,139455	-0,06569
Y6	0,07851	0,17166	-0,03668	0,68745	0,139455	0,594217	0,21665
Y7	-1,01047	0,89036	-0,98492	0,35868	-0,065693	0,216652	2,51299

Análise de Componentes Principais

Decomposição espectral da matriz de covariância amostral (S):

Autovalores: 4,67716 2,12199 1,25753 1,07494 0,74288 0,16508 0,00842

Matriz dos Autovetores:

-0,444760	0,463895	0,082832	-0,370766	0,524883	0,003613	0,408834
0,383611	0,297166	0,404365	-0,323955	-0,560938	0,038013	0,424241
-0,495679	0,101991	-0,595498	-0,224854	-0,580861	-0,035533	0,007825
0,135095	0,612061	-0,153940	0,548036	-0,029365	-0,530750	0,023060
-0,030360	0,403650	0,251071	-0,345790	-0,037324	-0,005132	-0,807538
0,066693	0,377823	-0,156391	0,336346	-0,012315	0,845541	-0,011102
0,621058	0,076082	-0,603173	-0,417113	0,264651	-0,024851	-0,006317

P1 P2 P3 P4 P5 P6 P7

Análise de Componentes Principais

Coeficientes das Componentes Principais Z em função das variáveis Y e correspondentes variâncias

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	
Var(Y)	1,74432	1,4296	1,92239	1,2801	0,564388	0,594217	2,51299	Var(Z)
Z1	-0,444760	0,383611	-0,495679	0,135095	-0,030360	0,066693	0,621058	4,67716
Z2	0,463895	0,297166	0,101991	0,612061	0,403650	0,377823	0,076082	2,12199
Z3	0,082832	0,404365	-0,595498	-0,153940	0,251071	-0,156391	-0,603173	1,25753
Z4	-0,370766	-0,323955	-0,224854	0,548036	-0,345790	0,336346	-0,417113	1,07494
Z5	0,524883	-0,560938	-0,580861	-0,029365	-0,037324	-0,012315	0,264651	0,74288
Z6	0,003613	0,038013	-0,035533	-0,530750	-0,005132	0,845541	-0,024851	0,16508
Z7	0,408834	0,424241	0,007825	0,023060	-0,807538	-0,011102	-0,006317	0,00842

P'

$$\sum_{j=1}^{p} Var\left(Y_{ij}\right) = \sum_{j=1}^{p} Var\left(Z_{ji}\right) = 10,048$$

variância total

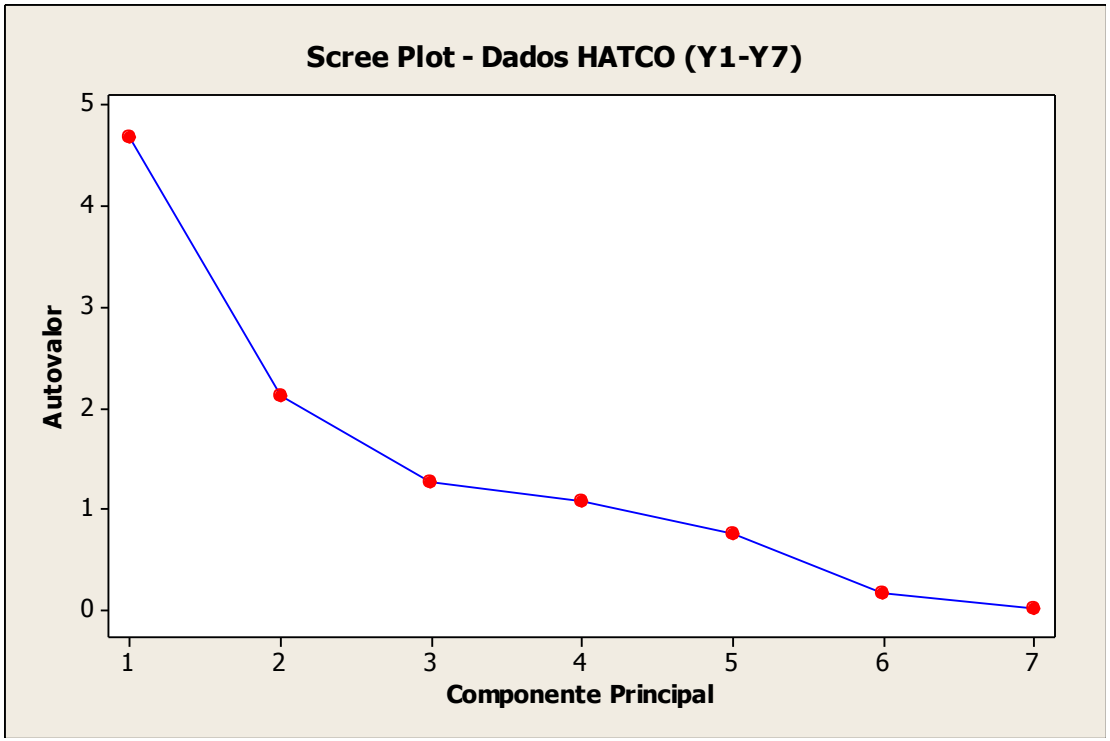
C.P.	%Var Total	% Acumul
Z1	0,465481	0,465481
Z2	0,211185	0,676666
Z3	0,125152	0,801818
Z4	0,10698	0,908798
Z5	0,073933	0,982731
Z6	0,016429	0,99916
Z7	0,000838	1

Quantos Componentes reter na análise?

Análise de Componentes Principais

Redução de dimensionalidade.

Quantos Componentes Principais Usar?



- acumular aproximadamente 70% da variabilidade total dos dados
- reter as C.P. que acumulem cerca de 50% da variabilidade de cada variável original
- Critério de corte: quando a variação entre os autovalores passa a ser pequena (*cotovelo* do gráfico)

Análise de Componentes Principais

Correlações entre as variáveis originais e as componentes principais

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Z1	-0,72829	0,69386	-0,77316	0,25823	-0,0874	0,18711	0,84728
Z2	0,51166	0,36205	0,10715	0,78803	0,78269	0,71398	0,06991
Z3	0,07033	0,37925	-0,48164	-0,15258	0,37477	-0,22751	-0,42668
Z4	-0,29106	-0,28091	-0,16814	0,5022	-0,47722	0,45238	-0,2728
Z5	0,34254	-0,40436	-0,36109	-0,02237	-0,04282	-0,01377	0,14389
Z6	0,00111	0,01292	-0,01041	-0,1906	-0,00278	0,44566	-0,00637
Z7	0,02841	0,03256	0,00052	0,00187	-0,09865	-0,00132	-0,00037

Como calcular tais coeficientes de correlação ?

A redução para os *dois primeiros CP* (Z1 e Z2) é apropriada?

Análise de Componentes Principais

Correlações (**R**) entre as variáveis originais e as componentes principais

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Z1	-0,72829	0,69386	-0,77316	0,25823	-0,0874	0,18711	0,84728
Z2	0,51166	0,36205	0,10715	0,78803	0,78269	0,71398	0,06991
Z3	0,07033	0,37925	-0,48164	-0,15258	0,37477	-0,22751	-0,42668
Z4	-0,29106	-0,28091	-0,16814	0,5022	-0,47722	0,45238	-0,2728
Z5	0,34254	-0,40436	-0,36109	-0,02237	-0,04282	-0,01377	0,14389
Z6	0,00111	0,01292	-0,01041	-0,1906	-0,00278	0,44566	-0,00637
Z7	0,02841	0,03256	0,00052	0,00187	-0,09865	-0,00132	-0,00037

Proporção da Variância de Y explicada por cada componente principal (**R^2**)

Z1 e Z2 explicam 79% da variância de Y1

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Var(Y)	1,74432	1,4296	1,92239	1,2801	0,564388	0,594217	2,51299
Z1	0,530406	0,481442	0,597776	0,066683	0,007639	0,03501	0,717883
Z2	0,261796	0,13108	0,011481	0,620991	0,612604	0,509767	0,004887
Z3	0,004946	0,143831	0,231977	0,023281	0,140453	0,051761	0,182056
Z4	0,084716	0,07891	0,028271	0,252205	0,227739	0,204648	0,07442
Z5	0,117334	0,163507	0,130386	0,0005	0,001834	0,00019	0,020704
Z6	0,000001	0,000167	0,000108	0,036328	0,000008	0,198613	0,000041
Z7	0,000807	0,00106	0	0,000003	0,009732	0,000002	0
	1						

Análise de Componentes Principais

Arquivo HATCO (Hair et al., 2005)

ID	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Z1	Z2	Z3	Z4	Z5	Z6	Z7
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2							
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4							
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2							
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8							
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5							
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7							
…														
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1							
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4							
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4							
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0							

Escores ou valores dos CP

$$Y_{100\times7};\quad Cov(Y)=\Sigma \quad \Rightarrow \quad Z_{100\times7};\quad Cov(Z)=\Lambda_{\lambda_j}$$

$$\sum_{j=1}^{p} Var(Y_j) \quad = \quad \sum_{j=1}^{p} Var(Z_j)$$

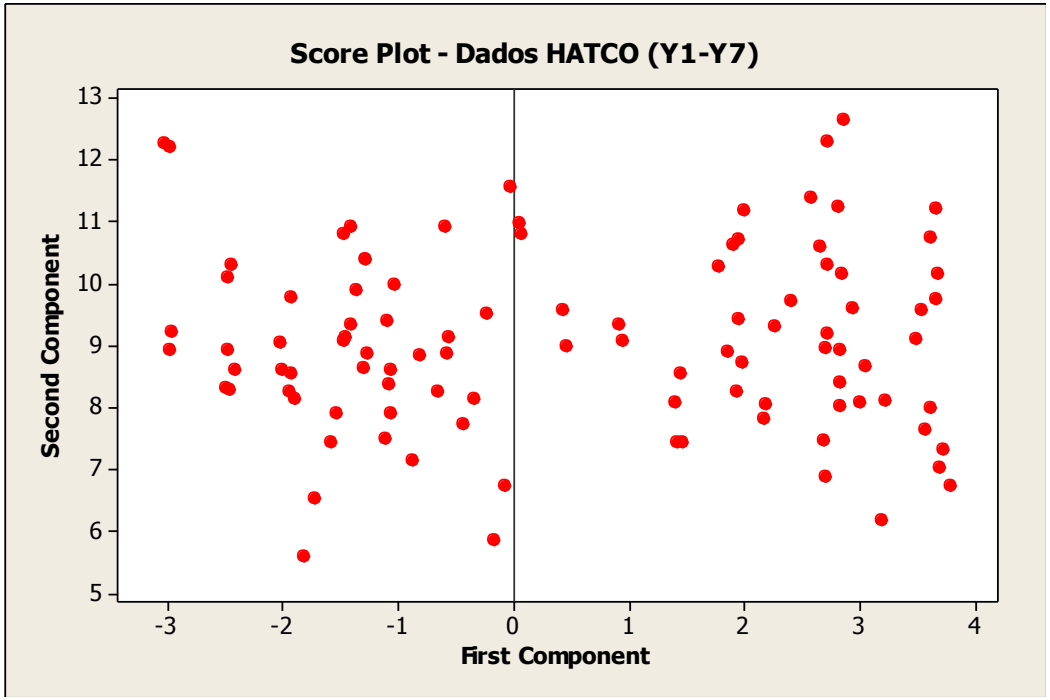
$$Corr(Y_j;Z_k)=a_{jk}\sqrt{\lambda_k}/\sigma_{jj}$$

Equivale a calcular a correlação de Pearson entre Yj e Zk

Análise de Componentes Principais

Y_{100x7}?

Diagrama de dispersão das observações de acordo com os dois primeiros componentes principais Z_{100x2}



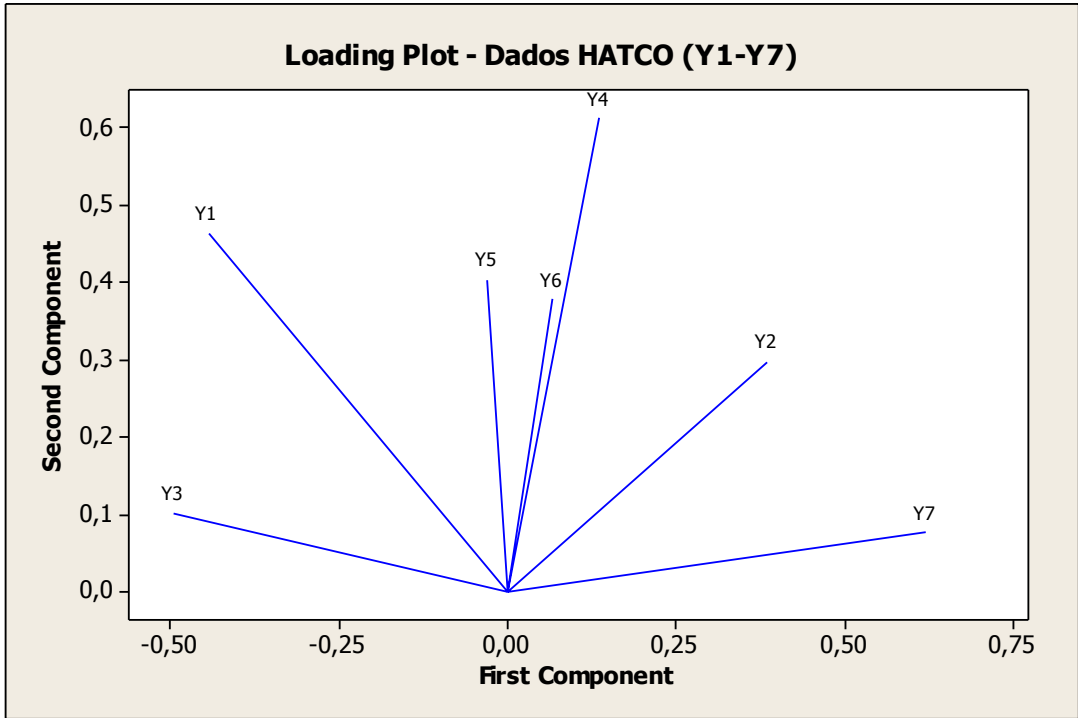
Como estes escores são calculados?

$(Z_{1i}; Z_{2i}), i = 1,2,\ldots,n$

- Visualização dos dados em $\Re 2$
- Como as observações estão dispersas (agrupadas)?
- Identificar observações atípicas multivariadas (além de usar d_M^2)

Análise de Componentes Principais

Visualização dos pesos das variáveis originais para os dois primeiros CP



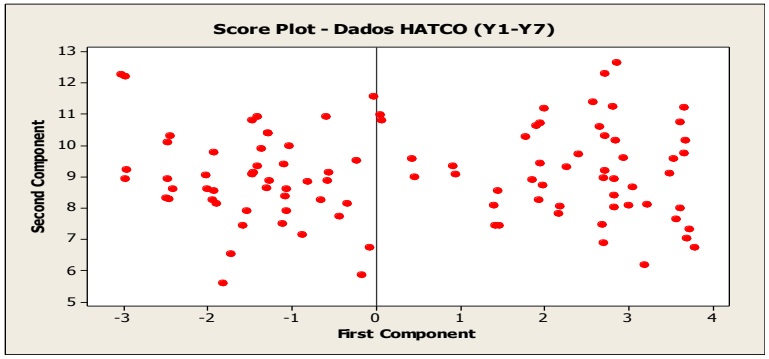
Veremos a construção do **Biplot**: representação gráfica simultânea dos escores e das cargas dos dois primeiros CP

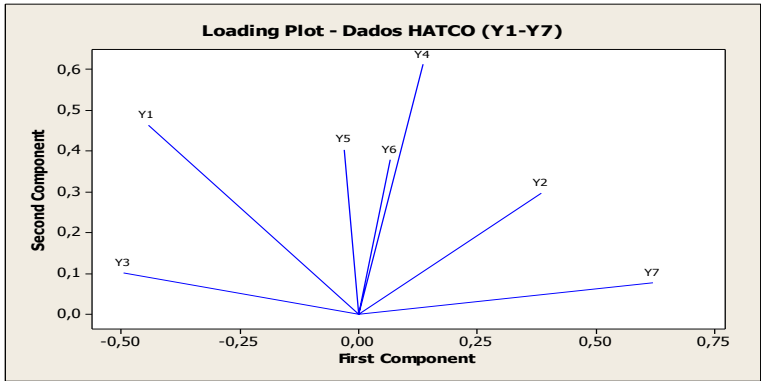
- Essa representação auxilia na interpretação dos CP
- Permite a avaliação da contribuição das variáveis originais aos CP

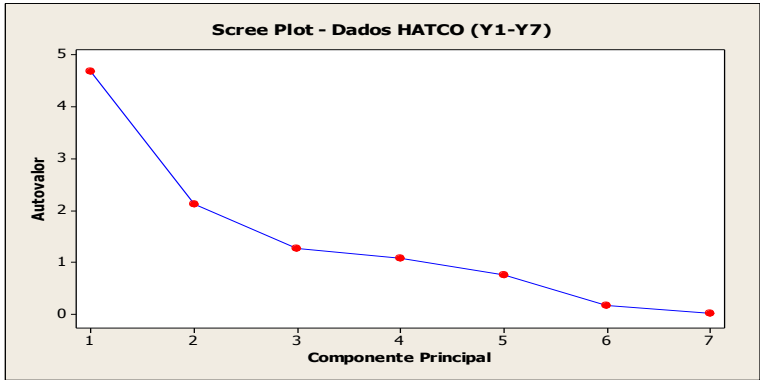
Análise de Componentes Principais

NÃO é invariante por padronização dos dados

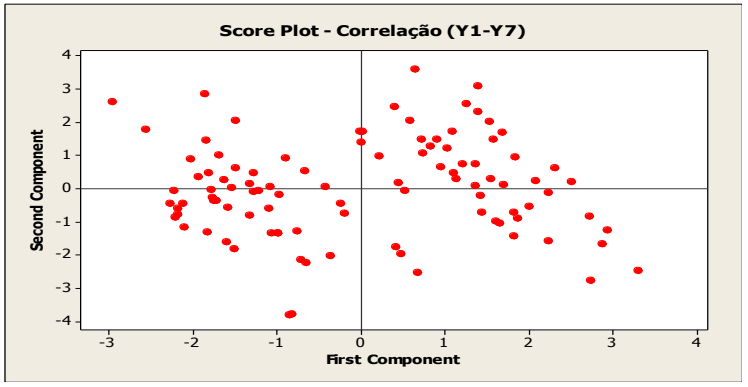
Dados Originais: decomposição de Σ

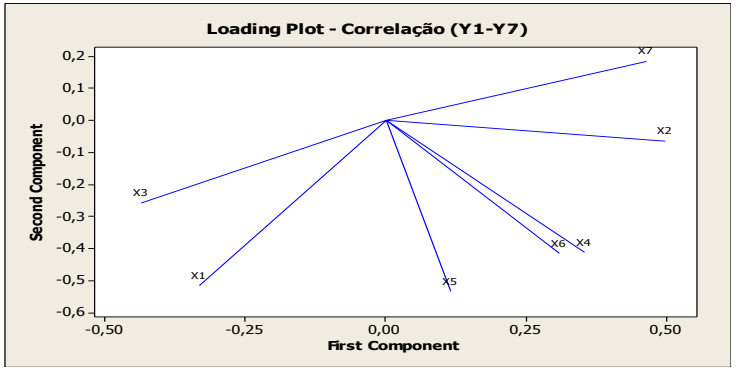


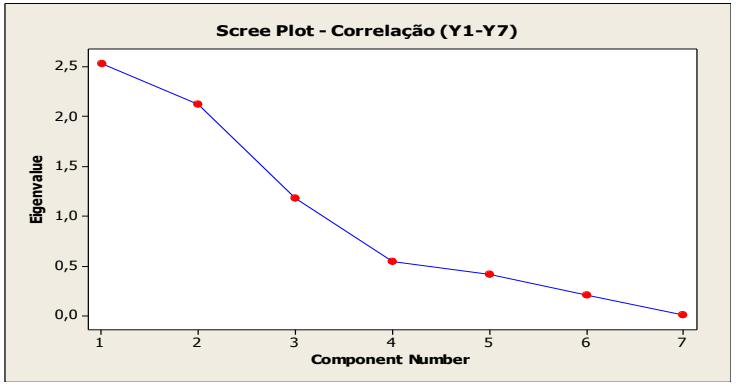




Dados Padronizados: decomposição de R







Análise de Componentes Principais

Na prática, Σ e R não são conhecidas e estimativas (MVS ou estimadores robustos) são utilizadas na decomposição espectral.

- Variáveis originais (Y) em escalas diferentes (com heterocedasticidade) podem ser padronizadas, o que equivale aos CP via R. Os resultados via Σ ou R NÃO são os mesmos e não há uma função relacionando-os.

- Quando o objetivo é o agrupamento de observações, em geral, não há necessidade de padronização das variáveis. Contudo, se o objetivo é a construção de índices (econômicos, de qualidade de vida, de desempenho do atleta, etc.), em geral, recomenda-se padronizar as variáveis.

- A interpretação das CP é fundamental (termos como "média ponderada" e "diferença entre médias ponderadas" das variáveis são comumente utilizados). Os coeficientes/cargas ("a_{jk}") e as correlações (r_{YjZk}) das variáveis originais com os CP são úteis na interpretação dos componentes principais.

- A estrutura de Σ é decisiva na análise de CP. Sob $\Sigma = (1-\rho)I_p + \rho 1_p 1_p^{'};\ \rho > 0$ as variáveis originais têm o mesmo "peso" na construção do CP1. Nesse caso,

$\lambda_1 = 1+(p-1)\rho$ e os restantes (p-1) autovalores são iguais a $(1-\rho)$, com o primeiro autovetor proporcional a 1_p. O teste dessa estrutura é conhecido como teste da esferecidade.

Escalonamento Multidimensional
Análise de Coordenadas Principais

Dados Multivariados

$$D = \begin{pmatrix} 0 & & & \\ d_{21} & 0 & & \\ \dots & \dots & \dots & \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}_{n \times n} \qquad C = \begin{pmatrix} 1 & & & \\ r_{21} & 1 & & \\ \dots & \dots & \dots & \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}_{n \times n}$$

Matriz de Distâncias entre indivíduos

Matriz de Similaridades entre indivíduos

Objetivos:

- A partir de matrizes de distância ou similaridade (D ou C) entre *n* objetos/observações obter uma representação das correspondentes observações Y_{nxp} que geraram D ou C;
- Obter Eixos Principais (Coordenadas Principais) $\Rightarrow$ Identificar dimensões não conhecidas de observações multivariadas

Escalonamento Multidimensional

Análise baseada no espaço linha da matriz de dados

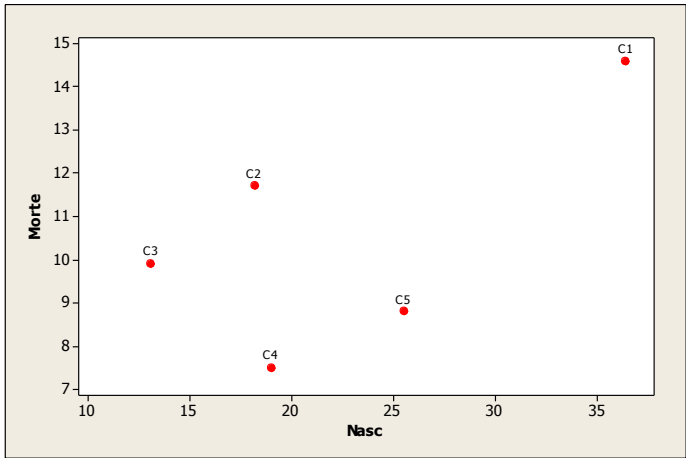
Escalonamento Multidimensional

Motivação: Matriz de Distâncias Euclidianas baseadas em taxas de Nascimento e Morte para 5 cidades

$Y_{5\times2}$: Matriz conhecida

Cidade	Nascimento	Morte
C1	36,4	14,6
C2	18,2	11,7
C3	13,1	9,9
C4	19	7,5
C5	25,5	8,8

Representação das observações (X conhecida)

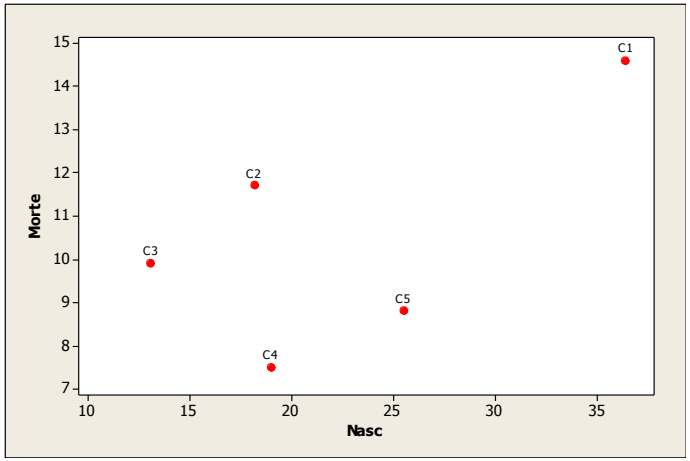


Considerando as taxas de nascimento e morte, a Cidade C1 está mais distante das demais. Isso pode ser confirmado calculando a distância (Euclidiana, por ex.) entre as cidades.

Escalonamento Multidimensional

Motivação: Matriz de Distâncias Euclidianas baseadas em taxas de Nascimento e Morte para 5 cidades

Cidade	Nascimento	Morte
C1	36,4	14,6
C2	18,2	11,7
C3	13,1	9,9
C4	19	7,5
C5	25,5	8,8



$$Y_{5\times 2} \Rightarrow D_{5\times 5}$$

$$d_{12}; \quad d_{12}^2 = (36{,}4 - 18{,}2)^2 + (14{,}6 - 11{,}7)^2$$

Matriz de distâncias Euclidianas

	C1	C2	C3	C4	C5
C1	0				
C2	18,43	0			
C3	23,76	5,41	0		
C4	18,79	4,28	6,37	0	
C5	12,34	7,85	12,45	6,63	0

Escalonamento Multidimensional

Matriz de Distâncias Euclidianas, D, baseadas nas taxas de nascimento e morte para 5 cidades.

	C1	C2	C3	C4	C5
C1	0				
C2	18,43	0			
C3	23,76	5,41	0		
C4	18,79	4,28	6,37	0	
C5	12,34	7,85	12,45	6,63	0

Supondo Y desconhecida, com base somente na matriz D, como as 5 cidades podem ser representadas em um gráfico (bidimensional)?

$$D_{5\times 5} \xrightarrow{?} Y_{5\times k}; \quad k = 2$$

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre as 7 primeiras observações do banco de dados HATCO (considerando as variáveis V1, V2, …, V5)

	1	2	3	4	5	6	7
1	0						
2	3,87685	0					
3	5,32823	3,37046	0				
4	1,98997	2,53574	5,13323	0			
5	4,64758	5,93212	6,67158	4,82908	0		
6	3,63043	2,43105	3,87685	2,90172	5,92453	0	
7	3,88201	4,40908	4,97192	3,83536	2,38118	3,83536	0

Como representar os 7 pontos em um gráfico?

$$D_{7\times 7} \xrightarrow{?} Y_{7\times k}; \quad k = 2$$

Escalonamento Multidimensional

Distâncias (em km) entre 12 cidades $\Rightarrow$ matriz de "distância" empírica

Não é matriz de distância Euclidiana é uma distância empírica!

	1	2	3	4	5	6	7	8	9	10	11	12
1	0											
2	244	0										
3	218	350	0									
4	284	77	369	0								
5	197	167	347	242	0							
6	312	444	94	463	441	0						
7	215	221	150	236	279	245	0					
8	469	583	251	598	598	169	380	0				
9	166	242	116	257	269	210	55	349	0			
10	212	53	298	72	170	392	168	531	190	0		
11	253	325	57	340	359	143	117	264	91	273	0	
12	270	168	284	164	277	378	143	514	174	111	256	0

Como representar os 12 pontos em um gráfico?

$$D_{12\times 12} \xrightarrow{?} Y_{12\times k}; \quad k = 2$$

Escalonamento Multidimensional

Matriz de Distância* (“postos”) entre 6 Docerias

	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

Não é matriz de distância Euclidiana é uma distância empírica!

*1: é o par mais similar 15: é o par menos similar (6(6-1)/2)

Como representar os 6 pontos em um gráfico?

$$D_{6\times 6} \xrightarrow{?} Y_{6\times k}; \quad k = 2$$

Escalonamento Multidimensional

Matriz de Similaridade: Porcentagem de vezes que sinais de Código Morse (para indicação de números) foram declarados como iguais num total de 598 jurados ($\Rightarrow$ matriz de "similaridade" empírica)

	1	2	3	4	5	6	7	8	9	0
1	84									
2	62	89								
3	16	59	86							
4	6	23	38	89						
5	12	8	27	56	90					
6	12	14	33	34	30	86				
7	20	25	17	24	18	65	85			
8	37	25	16	13	10	22	65	88		
9	57	28	9	7	5	8	31	58	91	
0	52	18	9	7	5	18	15	39	79	94

Não é matriz de distância Euclidiana

Como representar os 10 pontos em um gráfico?

$$C_{n\times n} = D\left(d_{ij}\right); \quad d_{ij} = \left(c_{ii} + c_{jj} - 2c_{ij}\right)^{1/2}$$

Escalonamento Multidimensional

Notação:

Dada uma matriz de distâncias D,

$$D = \left(d_{ij}\right)_{n \times n}$$

O objetivo do escalonamento multidimensional é encontrar pontos,

$Y_1, Y_2, \ldots, Y_n$ k-dimensionais, tal que, se $\hat{d}_{ij}$ é a distância

Euclidiana entre Y_i e Y_j, então $\hat{D} = \left(\hat{d}_{ij}\right)$ é uma "aproximação" para D em algum sentido.

Solução:

- Métodos métricos $\Rightarrow$ os pontos P são obtidos tal que $\hat{D} \cong D$
- Métodos não métricos $\Rightarrow$ baseados na ordenação das n(n-1)/2 distâncias e minimização de funções objetivo como o "stress"

Escalonamento Multidimensional

Solução Clássica em k dimensões

(Mardia, 1979)

Dado *D*: matriz de distância Euclidiana $\Leftrightarrow$ Existe $Y_{n\times p}$: matriz de dados

$$D=\left(d_{ij}\right)_{n\times n};\quad d_{ij}^2=\sum_{k=1}^{p}\left(y_{ik}-y_{jk}\right)^2$$

d_{ij} : conhecido

y_{ik} : desconhecido

Existe: $B_{n\times n}=Y_{n\times p}\,Y'_{p\times n}$ $\Rightarrow$ $b_{ij}=\sum_{k=1}^{p}y_{ik}y_{jk}$

Matriz de Produto Interno entre as linhas de Y

$$d_{ij}^2=b_{ii}+b_{jj}-2b_{ij}$$

$$b_{ij}=-\frac{1}{2}\left(d_{ij}^2-d_{i.}^2-d_{.j}^2+d_{..}^2\right)$$

B:calculada de D

Escalonamento Multidimensional

Solução Clássica em k dimensões

(Mardia, 1979)

$$D = \left(d_{ij}\right)_{n\times n} \Leftrightarrow \quad Y_{n\times p}\ ? \qquad d_{ij}^2 = \sum_{k=1}^{p}\left(y_{ik} - y_{jk}\right)^2$$

$$B = YY' \quad \Rightarrow \quad B = \left(b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)\right)$$

Logo, temos:

$B_{n\times n} = Y\,Y'$ Matriz p.s.d. (n>p) e sua Decomposição Espectral é:

$$= P\,\Lambda\,P' = P\Lambda^{1/2}\ \Lambda^{1/2}P' = \left(P\Lambda^{1/2}\right)\left(P\Lambda^{1/2}\right)'$$

$$Y = \left(P\Lambda^{1/2}\right)$$

Escalonamento Multidimensional

Solução Clássica em k dimensões

$$D = \left(d_{ij}\right)_{n\times n} \Leftrightarrow \qquad Y_{n\times p}\ ? \qquad d_{ij}^2 = \sum_{k=1}^{p}\left(y_{ik} - y_{jk}\right)^2$$

$$B = YY' \quad \Rightarrow \quad B = \left(b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)\right)$$

$$B_{n\times n} = P_{n\times n}\,\Lambda_{n\times n}\,P'_{n\times n} = P\,\Lambda^{1/2}\left(P\Lambda^{1/2}\right)' = Y\,Y'$$

Matriz de Dados Y é obtida de:

- Quando n>p, o posto de D é p. Logo, há (n-p) autovalores nulos.
- Podemos escolher uma representação para Y em uma dimensão k (k<p)

$$Y = P\Lambda^{1/2} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & & \dots & \\ a_{n1} & a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda_2} & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \sqrt{\lambda_n} \end{pmatrix}$$

$$n > p \Rightarrow \lambda_{(p+1)} = \dots = \lambda_n = 0$$

Escalonamento Multidimensional

Solução Clássica em k dimensões

$$D = \left(d_{ij}\right)_{n\times n}$$

Matriz de distâncias

$$\begin{cases} \Leftrightarrow & Y_{n\times p} \ ? \\ B_{n\times n} = \left(b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)\right) \end{cases}$$

Encontre os “k” primeiros componentes da decomposição espectral de B:

Autovalores: $\lambda_1 > \lambda_2 > \ldots > \lambda_k > \lambda_{k+1} > \ldots > \lambda_n > 0$

Autovetores normalizados: $P = \left(P_1, P_2, \ldots, P_k, P_{k+1}, \ldots, P_p, P_{p+1}, \ldots P_n\right)$

$\Rightarrow$ As coordenadas do vetor de resposta Y_i são obtidas a partir da *i*-ésima linha da matriz $P=(a_{ij})$

$$Y_i = \left(P_1 \ldots P_k\right)_i \ \Lambda^{1/2} = \left(a_{i1}\sqrt{\lambda_1}, a_{i2}\sqrt{\lambda_2}, \ldots, a_{ik}\sqrt{\lambda_k}\right)$$

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre as 7 primeiras observações do banco de dados HATCO

	1	2	3	4	5	6	7
1	0						
2	3,87685	0					
3	5,32823	3,37046	0				
4	1,98997	2,53574	5,13323	0			
5	4,64758	5,93212	6,67158	4,82908	0		
6	3,63043	2,43105	3,87685	2,90172	5,92453	0	
7	3,88201	4,40908	4,97192	3,83536	2,38118	3,83536	0

$$B = \begin{pmatrix} 5,97037 & -1,82095 & -4,4311 & 2,98791 & 0,037 & -0,96806 & -1,7724 \\ -1,82095 & 5,41773 & 3,8075 & 1,4766 & -7,0343 & 2,39062 & -4,23372 \\ -4,43115 & 3,80753 & 13,5573 & -4,4136 & -7,6245 & 1,90042 & -2,80392 \\ 2,98791 & 1,4766 & -4,4136 & 3,96546 & -1,8255 & 0,40948 & -2,59486 \\ 0,03699 & -7,03432 & -7,6245 & -1,82546 & 15,7036 & -7,06144 & 7,79422 \\ -0,96806 & 2,39062 & 1,9004 & 0,40948 & -7,0614 & 5,27351 & -1,94084 \\ -1,7724 & -4,23372 & -2,8039 & -2,59486 & 7,7942 & -1,94084 & 5,55482 \end{pmatrix}$$

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre as 7 primeiras observações do banco de dados HATCO

	1	2	3	4	5	6	7
1	0						
2	3,87685	0					
3	5,32823	3,37046	0				
4	1,98997	2,53574	5,13323	0			
5	4,64758	5,93212	6,67158	4,82908	0		
6	3,63043	2,43105	3,87685	2,90172	5,92453	0	
7	3,88201	4,40908	4,97192	3,83536	2,38118	3,83536	0

$B \Rightarrow \lambda = (\ 31{,}2312 \quad 15{,}2472 \quad 5{,}0860 \quad 3{,}8702 \quad 0{,}0085 \quad -0{,}0028 \quad 0{,}0026\ )$

$$P = \begin{pmatrix} 0{,}094195 & 0{,}475616 & 0{,}521284 & -0{,}471626 & 0{,}40608 & 0{,}267124 & -0{,}185722 \\ -0{,}34268 & 0{,}08044 & -0{,}082753 & 0{,}645763 & 0{,}598891 & 0{,}273346 & -0{,}137311 \\ -0{,}470631 & -0{,}608035 & 0{,}431356 & -0{,}120414 & -0{,}191785 & 0{,}263896 & -0{,}319076 \\ -0{,}013354 & 0{,}486511 & 0{,}01752 & 0{,}298922 & -0{,}565222 & 0{,}081535 & -0{,}589358 \\ 0{,}681413 & -0{,}246653 & 0{,}099089 & 0{,}241493 & -0{,}091019 & 0{,}626075 & 0{,}080287 \\ -0{,}279838 & 0{,}109615 & -0{,}630516 & -0{,}401399 & -0{,}087106 & 0{,}584723 & 0{,}038927 \\ 0{,}330745 & -0{,}29676 & -0{,}357312 & -0{,}193042 & 0{,}322954 & -0{,}209204 & -0{,}69967 \end{pmatrix}$$

$\Rightarrow$ Escolhendo k=2: explica 84% da variabilidade total

Escalonamento Multidimensional

Matriz de Distância Euclidiana entre as 7 primeiras observações do banco de dados HATCO

	1	2	3	4	5	6	7
1	0						
2	3,87685	0					
3	5,32823	3,37046	0				
4	1,98997	2,53574	5,13323	0			
5	4,64758	5,93212	6,67158	4,82908	0		
6	3,63043	2,43105	3,87685	2,90172	5,92453	0	
7	3,88201	4,40908	4,97192	3,83536	2,38118	3,83536	0

$B \Rightarrow \lambda = (31{,}2312 \quad 15{,}2472 \quad 5{,}0860 \quad 3{,}8702 \quad 0{,}0085 \quad -0{,}0028 \quad 0{,}0026)$

$0{,}094195\sqrt{31{,}2312}$

$0{,}475616\sqrt{15{,}2472}$

$Y =$

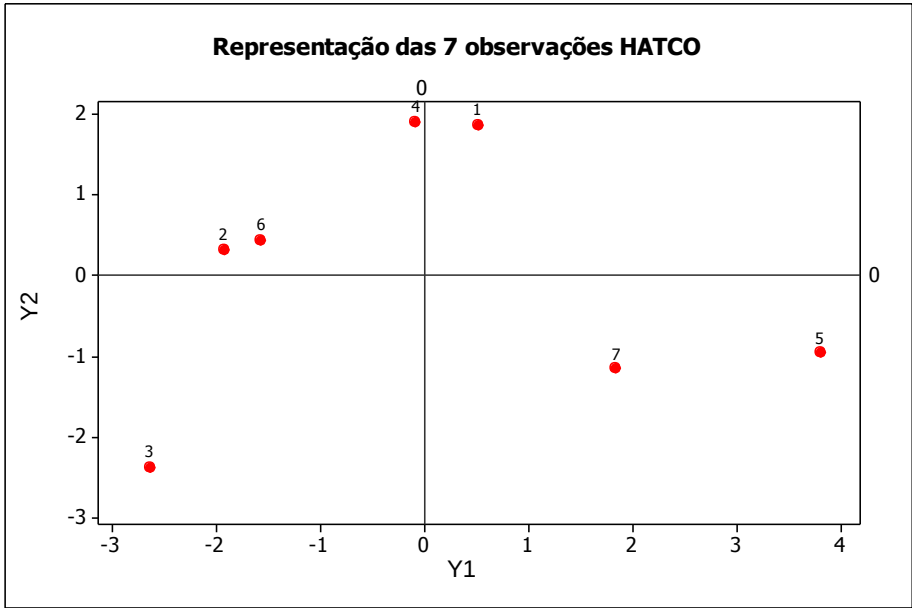
0,52641	1,85717
-1,91506	0,3141
-2,63012	-2,37423
-0,07463	1,89971
3,80807	-0,96312
-1,56387	0,42802
1,84837	-1,15878

As colunas de Y são denominadas as coordenadas principais de uma representação dos dados

Escalonamento Multidimensional

Representação das 7 primeiras observações do banco de dados HATCO obtida de D:

$$Y = \begin{pmatrix} 0{,}52641 & 1{,}85717 \\ -1{,}91506 & 0{,}3141 \\ -2{,}63012 & -2{,}37423 \\ -0{,}07463 & 1{,}89971 \\ 3{,}80807 & -0{,}96312 \\ -1{,}56387 & 0{,}42802 \\ 1{,}84837 & -1{,}15878 \end{pmatrix}$$



Escalonamento Multidimensional

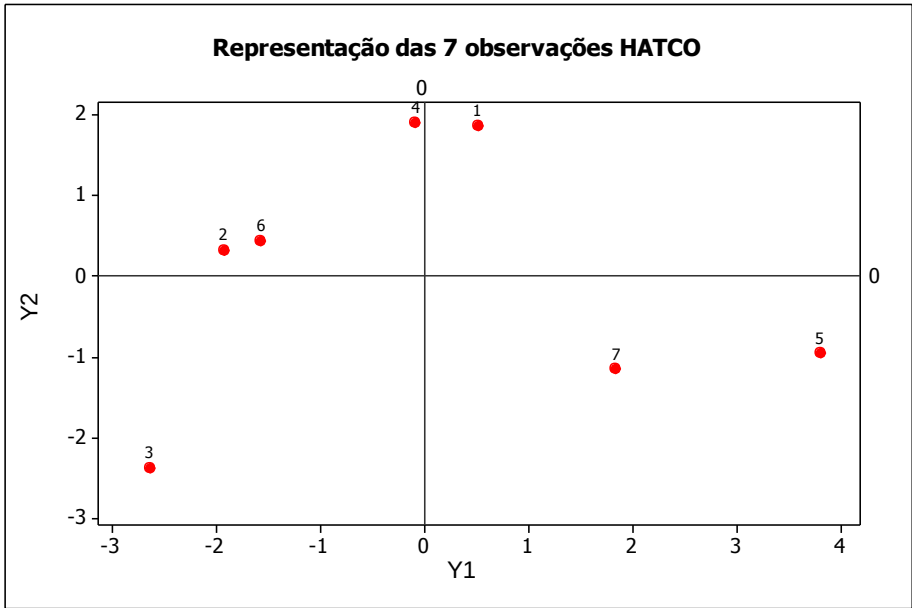
Matriz de Distância Euclidiana (observada e predita) entre as 7 primeiras observações do banco de dados HATCO

$\hat{D} \setminus D$

0	3,87685	5,32823	1,98997	4,64758	3,63043	3,88201
2,88822	0	3,37046	2,53574	5,93212	2,43105	4,40908
5,27905	2,7818	0	5,13323	6,67158	3,87685	4,97192
0,60254	2,42927	4,97967	0	4,82908	2,90172	3,83536
4,32705	5,86392	6,59102	4,82402	0	5,92453	2,38118
2,53214	0,3692	2,99825	2,09374	5,54915	0	3,83536
3,29295	4,04138	4,64049	3,61279	1,96945	3,76315	0

Y=

0,52641	1,85717
-1,91506	0,3141
-2,63012	-2,37423
-0,07463	1,89971
3,80807	-0,96312
-1,56387	0,42802
1,84837	-1,15878



Escalonamento Multidimensional

Dualidade entre Análise de Componentes Principais e

Análise de Coordenadas Principais (Escalonamento Multidimensional)

Análise de CP

$$Y_{n\times p};\quad \Sigma_{p\times p} = Y'Y = V_{p\times p}\Lambda_p V'_{p\times p} \quad\Rightarrow\quad Z_{n\times p} = YV$$

$$Y = U_{n\times n}\Lambda_n^{1/2}V'_{p\times p};\quad n \geq p \quad\Rightarrow\quad \lambda_{p+1} = \ldots = \lambda_n = 0$$

$$YV = U\,\Lambda^{1/2}V'V = U_{n\times n}\Lambda_p^{1/2}$$

↓ Componentes principais

↓ Coordenadas principais

Análise de EM

$$Y_{n\times p}; D_{n\times n} \rightarrow B = YY' = U_{n\times n}\Lambda_n U'_{n\times n} \quad\Rightarrow Y_{n\times p} = \left(U_1\sqrt{\lambda_1}\; U_2\sqrt{\lambda_2}\ldots U_p\sqrt{\lambda_p}\right) = U_{n\times p}\Lambda^{1/2}$$

- Os k primeiros Componentes Principais são "ótimos" $\Rightarrow$ a soma das variâncias é maior do que qualquer outro conjunto de k combinações lineares não correlacionadas
- as k primeiras Coordenadas Principais são "ótimas" $\Rightarrow$ a projeção de Y no sub-espaço de dimensão k de $\Re^p$ é mais próxima (em distância Euclidiana) da configuração original do que qualquer outra ($\hat{D} \cong D$)

Componentes Principais – Coordenadas Principais
Solução via Espaços Duais

$Y_{n\times p}$: Matriz de dados ("padronizados") multivariados de posto *r=min(n,p)*

Análise no espaço das variáveis: $\Re^{pxp}$

$$Y'Y = \Sigma_{p\times p} = V_{p\times p}\begin{pmatrix}\Lambda_r & 0\\ 0 & 0\end{pmatrix}V'_{p\times p} \Rightarrow \underbrace{Y_{n\times p}V_{p\times r}}_{\approx Y}$$

r Componentes Principais

Análise no espaço dos indivíduos: $\Re^{nxn}$

$$D_{n\times n} \Rightarrow B_{n\times n} = YY' = U_{n\times n}\begin{pmatrix}\Lambda_r & 0\\ 0 & 0\end{pmatrix}U'_{n\times n} \Rightarrow \underbrace{U\Lambda_r^{1/2}}_{\approx Y}$$

Escalonamento Multidimensional:
r Coordenadas Principais
obtidas da Matriz de Distâncias

Análise no espaço $\Re^{nxp}$

$$Y_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_r^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p} \Rightarrow \boxed{Y_{n\times p}V_{p\times r} = U_{n\times n}\Lambda_r^{1/2}}$$

Equivalência entre os Componentes Principais e as Cordenadas Principais

n<<p : Componentes Principais de Y podem ser obtidos da decomposição espectral da matriz de distâncias D (nxn), de dimensão muito menor que Σ (pxp)

Componentes Principais – Coordenadas Principais

Equivalência das Soluções em Espaços Duais

$Y_{n\times p}$: Matriz de dados ("originais") de posto r=min(n,p)

HY: linhas de Y centradas

$$(HY)_{n\times p} = U_{n\times n}\begin{pmatrix}\Lambda_r^{1/2} & 0\\ 0 & 0\end{pmatrix}V'_{p\times p}$$

B=HYY'H

Análise em $\Re^{nxn}$

$HYY'H = U\Lambda U'$

$(n-1)S_u = Y'HY$

Análise em $\Re^{pxp}$

$Y'HY = V\Lambda V'$

$$U_{n\times n}\Lambda_m^{1/2} = (HY)_{n\times p}V_{p\times m} \qquad m\le r$$

Coordenadas Principais

Componentes Principais (das linhas de HY)

Coordenadas Principais

Dados HatCo: 7 primeiras observações (com p=7 variáveis)

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
1	4.1	0.6	6.9	4.7	2.4	2.3	5.2
2	1.8	3.0	6.3	6.6	2.5	4.0	8.4
3	3.4	5.2	5.7	6.0	4.3	2.7	8.2
4	2.7	1.0	7.1	5.9	1.8	2.3	7.8
5	6.0	0.9	9.6	7.8	3.4	4.6	4.5
6	1.9	3.3	7.9	4.8	2.6	1.9	9.7
7	4.6	2.4	9.5	6.6	3.5	4.5	7.6

Matriz de Distância Euclidiana (D)

	1	2	3	4	5	6
2	5.31					
3	6.13	3.62				
4	3.27	3.11	5.16			
5	5.23	7.12	7.86	6.28		
6	5.80	3.47	4.23	3.49	8.33	
7	5.07	4.51	5.32	4.43	3.91	5.09

Escalonamento Multidimensional

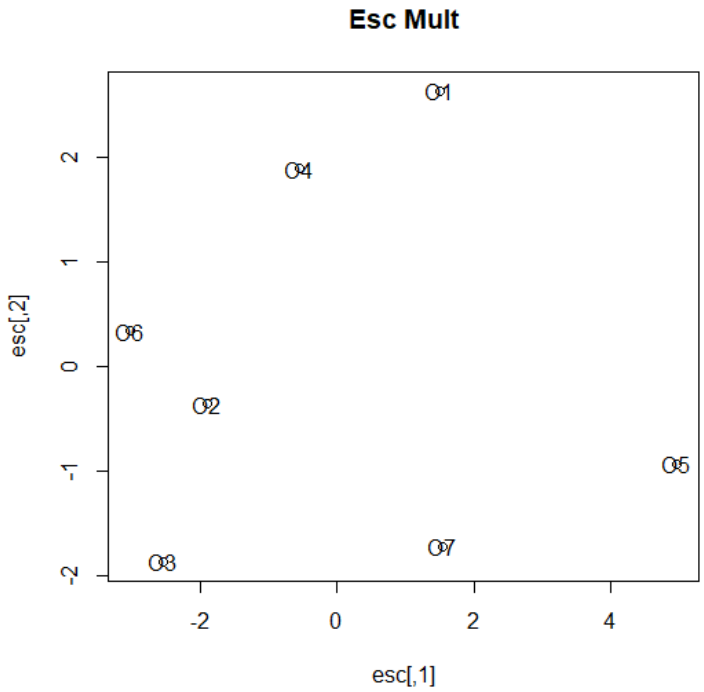
```
> cmdscale(diste)
        [,1]         [,2]
1   1.4981397   2.6360620
2  -1.8968160  -0.3552380
3  -2.5399671  -1.8649995
4  -0.5547189   1.8942258
5   4.9716633  -0.9304384
6  -3.0199393   0.3412407
7   1.5416383  -1.7208525
```

As colunas da matriz Y obtidas diretamente de D são denominadas Coordenadas Principais de uma escala construída aos dados.

Coordenadas Principais

Representação de Dados a partir de uma Matriz de Distâncias – Dados HatCo

	[,1]	[,2]
1	1.4981397	2.6360620
2	-1.8968160	-0.3552380
3	-2.5399671	-1.8649995
4	-0.5547189	1.8942258
5	4.9716633	-0.9304384
6	-3.0199393	0.3412407
7	1.5416383	-1.7208525



Matriz de Distância Euclidiana dos Dados Originais (D)

	1	2	3	4	5	6
2	5.31					
3	6.13	3.62				
4	3.27	3.11	5.16			
5	5.23	7.12	7.86	6.28		
6	5.80	3.47	4.23	3.49	8.33	
7	5.07	4.51	5.32	4.43	3.91	5.09

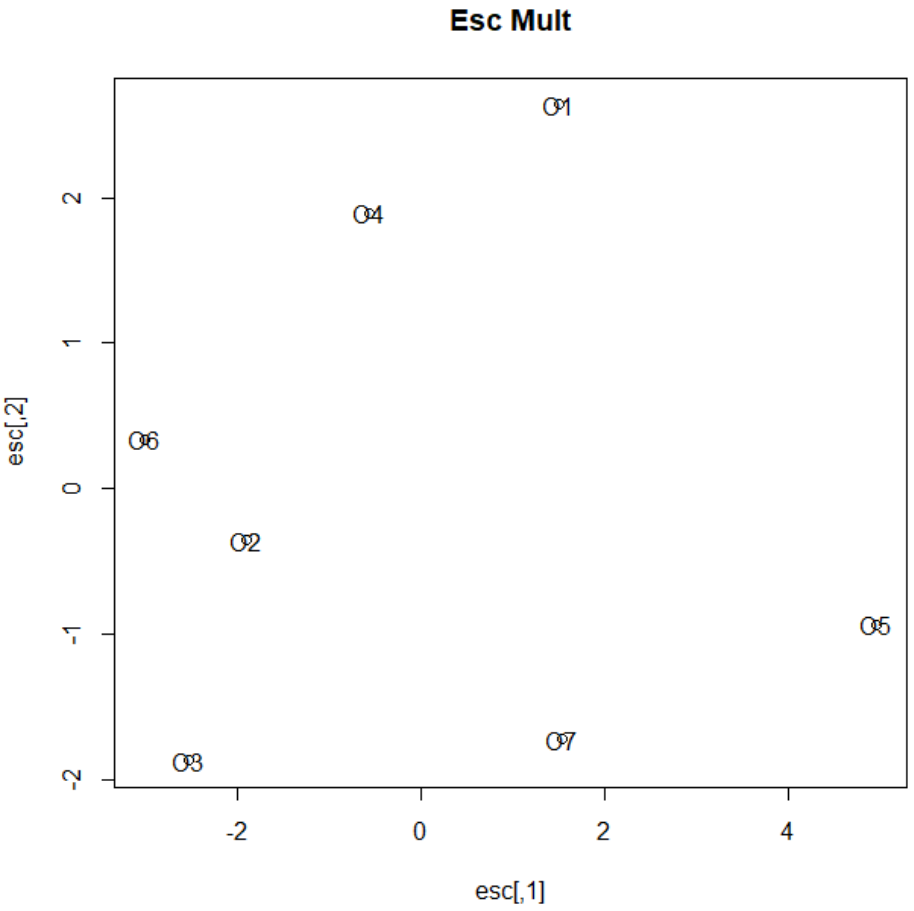
Avaliar a qualidade da representação em $\Re^2$

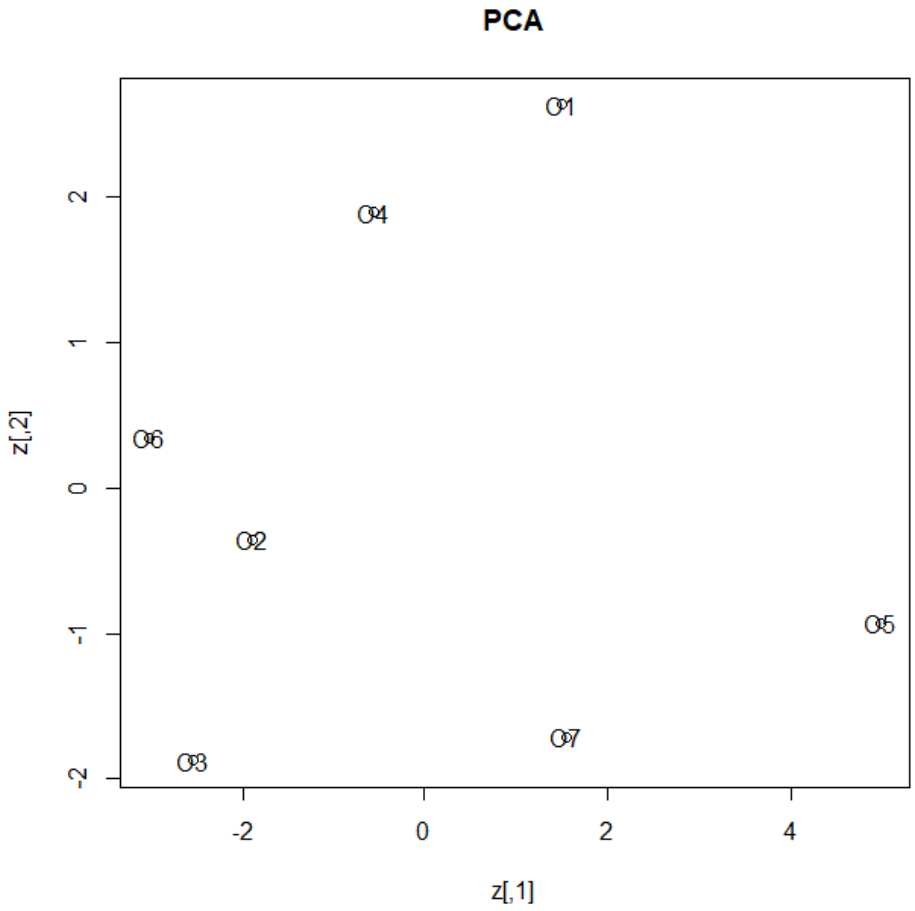
Matriz de Distância Euclidiana das Coordenadas Principais ($\widetilde{D}$)

	1	2	3	4	5	6
2	4.36					
3	5.23	3.73				
4	2.39	3.14	5.01			
5	4.93	5.30	6.34	5.49		
6	4.10	3.47	3.81	2.86	6.81	
7	4.62	3.54	4.56	4.48	2.77	4.81

Componentes Principais e Coordenadas Principais Equivalência

Coordenadas Principais extraídas de uma Matriz de Distância Euclidiana $\Rightarrow$ Representação (em $\Re^{2)}$ equivalente aos Componentes Principais extraídos de Σ





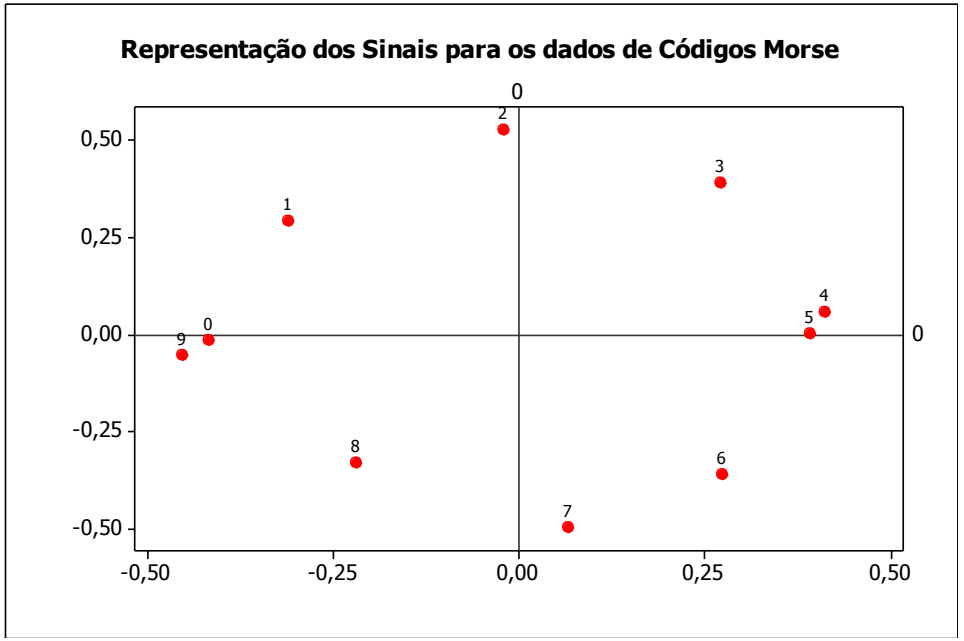
Escalonamento Multidimensional

Matriz de similaridade entre sinais de Código Morse

	1	2	3	4	5	6	7	8	9	0
1	84									
2	62	89								
3	16	59	86							
4	6	23	38	89						
5	12	8	27	56	90					
6	12	14	33	34	30	86				
7	20	25	17	24	18	65	85			
8	37	25	16	13	10	22	65	88		
9	57	28	9	7	5	8	31	58	91	
0	52	18	9	7	5	18	15	39	79	94

$$Y = \begin{pmatrix} -0{,}309585 & 0{,}290342 \\ -0{,}020226 & 0{,}523905 \\ 0{,}272386 & 0{,}3876 \\ 0{,}412688 & 0{,}055077 \\ 0{,}391557 & 0{,}000182 \\ 0{,}274401 & -0{,}35958 \\ 0{,}066838 & -0{,}49554 \\ -0{,}219191 & -0{,}33163 \\ -0{,}45251 & -0{,}05302 \\ -0{,}416357 & -0{,}01733 \end{pmatrix}$$

A orientação circular dos sinais somente é visualizada por meio da 2ª dimensão



Escalonamento Multidimensional

A análise de Coordenadas Principais (Escalonamento Multidimensional) é baseada em uma matriz de Distâncias (nxn) entre observações enquanto a análise de Componentes Principais é baseada em uma matriz de covariâncias (pxp) entre variáveis.
Equivalências entre essas análises:

1. A análise de Coordenadas Principais da matriz de distâncias Euclidianas é equivalente à análise de Componentes Principais da matriz de covariâncias.
2. A análise de Coordenadas Principais da matriz de distâncias de Penrose é equivalente à análise de Componentes Principais da matriz de correlação.

A análise de Coordenadas Principais pode ser aplicada de maneira mais geral, para diferentes escolhas de matriz de distâncias entre observações (Mahalanobis, Manhattan, entre outras). Neste caso, NÃO está garantida a equivalência entre as duas análises.

Componentes Principais, Coordenadas Principais e a Decomposição em Valores Singulares

Solução Métrica

$$Y = U_{n\times n}\Lambda_n^{1/2}V'_{p\times p};\quad n \geq p \quad \Rightarrow \quad \lambda_{p+1} = \ldots = \lambda_n = 0$$

$$YV = U\,\Lambda^{1/2}V'V = U_{n\times n}\Lambda_p^{1/2}$$

↓ Componentes principais

↓ Coordenadas principais

$$Y = \left(U_{n\times n}\Lambda_n^{1/2}\right)V'_{p\times p}$$

escores (pesos (coordenadas))

YV é uma rotação orthonormal de Y que permite representar os dados em um novo Sistema de Coordenadas $U\Lambda^{1/2}$ (podendo ser em baixa dimensão)

$$Y_{i\,p\times 1} \approx \left\{ \begin{bmatrix} U_1 & U_2 \end{bmatrix}_i \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}' \right\}$$

($\in \Re^p$; $'\in \Re^2$)

$$\approx \begin{bmatrix} \left(U_{i1}\sqrt{\lambda_1}\right)V_{11} + \left(U_{i2}\sqrt{\lambda_2}\right)V_{12} \\ \left(U_{i1}\sqrt{\lambda_1}\right)V_{21} + \left(U_{i2}\sqrt{\lambda_2}\right)V_{22} \end{bmatrix}$$

$$\approx \left(U_{i1}\sqrt{\lambda_1}\right)\begin{bmatrix} V_{11} \\ V_{21} \end{bmatrix} + \left(U_{i2}\sqrt{\lambda_2}\right)\begin{bmatrix} V_{12} \\ V_{22} \end{bmatrix};$$

($=Z_{1i}$; $=Z_{2i}$)

$$Z_{i\,2\times 1} = \begin{bmatrix} U_{i1}\sqrt{\lambda_1} \\ U_{i2}\sqrt{\lambda_2} \end{bmatrix}, \quad V_{2\times 2}$$

escores; sistema de coordenadas

$$\begin{pmatrix} 2 \\ 3 \end{pmatrix} = 2\begin{pmatrix} 1 \\ 0 \end{pmatrix} + 3\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 3 \end{pmatrix} = 2.5\begin{pmatrix} 1 \\ 1 \end{pmatrix} + (-1)\begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$$

Componentes Principais, Coordenadas Principais e a Decomposição em Valores Singulares

$$d_{ik}^2 = \underset{\in \Re^p}{(Y_i - Y_k)'}(Y_i - Y_k) \approx \left\{ \underset{\in \Re^2}{(Z_i - Z_k)'} [V_1 \quad V_2]' \right\} \left\{ (Z_i - Z_k)' [V_1 \quad V_2]' \right\}'$$

$$= (Z_i - Z_k)'(Z_i - Z_k)$$

$$= (Z_{1i} - Z_{1k})^2 + (Z_{2i} - Z_{2k})^2$$

A distância entre as observações originais em $\Re^p$ pode ser aproximada pela distância entre os componentes principais em $\Re^2 \Rightarrow$ permite a vizualização gráfica das observações.

Escalonamento Multidimensional

Métodos Não-Métricos

- **D** é considerada uma matriz de "dissimilaridade" geral (não precisa ser de distância Euclidiana)
- Os elementos de **D** podem ser ordenados

$$d_{ij}^{(1)} \le d_{ij}^{(2)} \le \ldots \le d_{ij}^{(m)}; \quad m = n(n-1)/2$$

- Seja $\hat{D}$, tal que os elementos $\hat{d}_{ij}$ estão monotonicamente relacionados aos elementos d_{ij}

$$d_{ij} < d_{rs} \quad \Rightarrow \quad \hat{d}_{ij} \le \hat{d}_{rs} \quad ; \mathrm{i} < \mathrm{j}, \mathrm{r} < \mathrm{s}$$

- Seja Y uma configuração em $\Re^k$ com distâncias $\hat{d}_{ij}$. Y é ótima no sentido de minimizar a seguinte medida:

$$S^2(Y) = \frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i<j} \left(d_{ij} - \bar{d} \right)^2}$$

Medida de stress de Y: mede quanto da variância de d_{ij} NÃO é explicada pelas k coordenadas principais

Escalonamento Multidimensional

Medidas de Distâncias (em Postos) entre 6 Docerias

	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

⇒ Localize os objetos A, B, C e D em uma única dimensão

⇒ Pense em um possível significado desta representação.

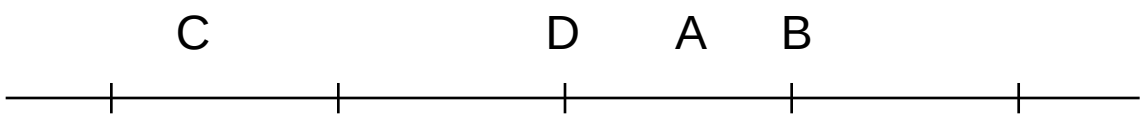
Escalonamento Multidimensional

Medidas de Distâncias (em Postos) entre 6 Docerias

	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

Localize os objetos A, B, C e D em uma única dimensão:

$$d_{AB} < d_{AD} < d_{BD} < d_{CD} < d_{BC} < d_{AC}$$



$$\hat{d}_{AB} < \hat{d}_{AD} < \hat{d}_{BD} < \hat{d}_{CD} < \hat{d}_{AC} < \hat{d}_{BC} \quad ?$$

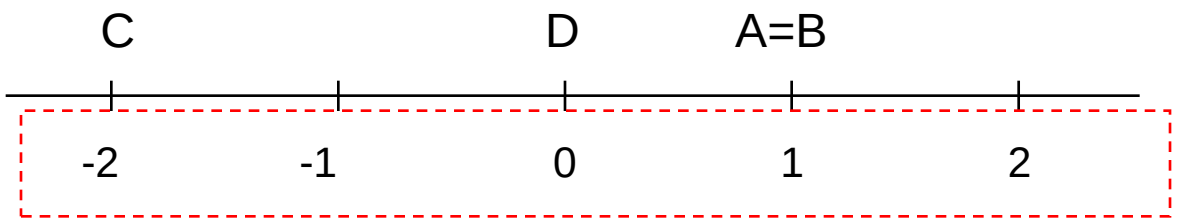
Escalonamento Multidimensional

Medidas de Distâncias (em Postos) entre 6 Docerias

	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

Localize os objetos A, B, C e D em uma única dimensão:

$$d_{AB} < d_{AD} < d_{BD} < d_{CD} < d_{BC} < d_{AC} \qquad \hat{d}_{AB} < \hat{d}_{AD} \leq \hat{d}_{BD} < \hat{d}_{CD} < \hat{d}_{AC} \leq \hat{d}_{BC}$$



C=-1⇒$S^2(Y)$=3,32

C=-2⇒$S^2(Y)$=2,75

$$S^2(Y) = \frac{\sum_{i<j}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j}\left(d_{ij} - 7,67\right)^2}$$

Escalonamento Multidimensional

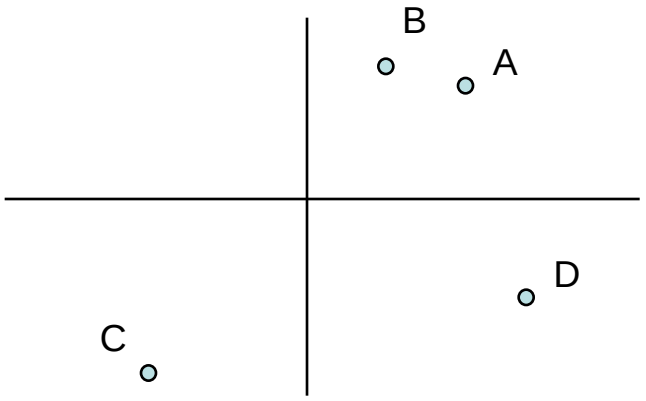
Medidas de Distâncias (em Postos) entre 6 Docerias

	A	B	C	D	E	F
A	-					
B	2	-				
C	13	12	-			
D	4	6	9	-		
E	3	5	10	1	-	
F	8	7	11	14	15	-

Agora, localize os objetos A, B, C e D em k=2 dimensões:

$$d_{AB} < d_{AD} < d_{BD} < d_{CD} < d_{BC} < d_{AC}$$

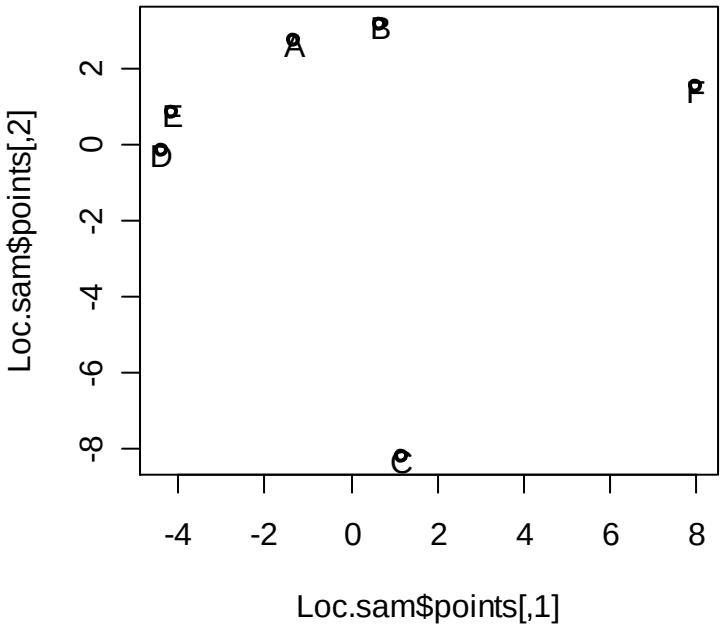
$$S^2(Y) = \frac{\sum_{i<j} \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum_{i<j} \left(d_{ij} - 7,67 \right)^2} \Rightarrow \hat{d}_{ij} \quad ?$$



$\Rightarrow$ Algoritmos de Busca da configuração ótima: funções “isoMDS” e “sammon” da biblioteca MASS do R utilizam o método de ordenação (não métrico)

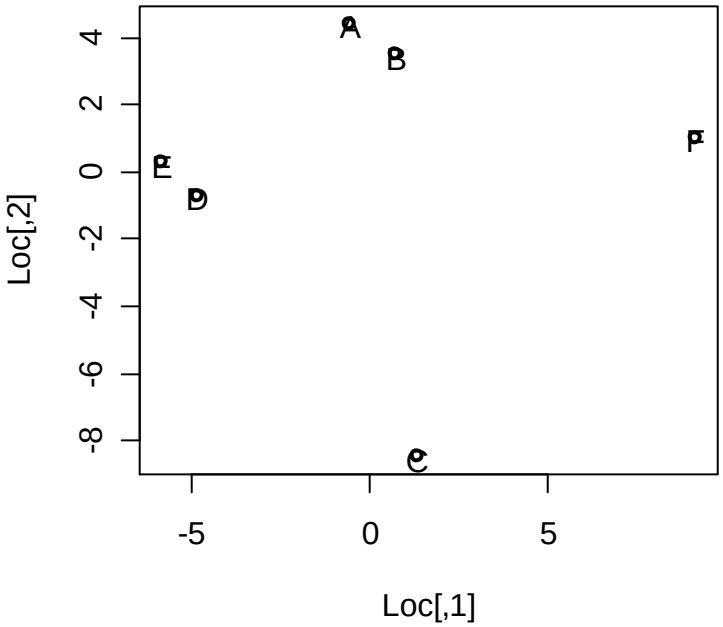
Solução Não Métrica: Sammon

```
          [,1]       [,2]
[1,] -1.3189094  2.7526337
[2,]  0.6507351  3.1832453
[3,]  1.2047833 -8.1931474
[4,] -4.3857141 -0.1449374
[5,] -4.1556192  0.8768785
[6,]  8.0047243  1.5253274
```



Solução Métrica

```
          [,1]       [,2]
[1,] -0.5455504  4.4014942
[2,]  0.7278073  3.4943629
[3,]  1.3306293 -8.4694243
[4,] -4.8265497 -0.7007641
[5,] -5.8277804  0.2607899
[6,]  9.1414439  1.0135414
```



Escalonamento Multidimensional

Recursos do R: Método Clássico (solução métrica)

```
 ?cmdscale
 ddat <-  matrix(c(0,2,13,4,3,8, 2,0,12,6,5,7,
13,12,0,9,10,11, 4,6,9,0,1,14, 3,5,10,1,0,15,
8,7,11,14,15,0), nrow=6, ncol=6)
 ddat
[,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    2   13    4    3    8
[2,]    2    0   12    6    5    7
[3,]   13   12    0    9   10   11
[4,]    4    6    9    0    1   14
[5,]    3    5   10    1    0   15
[6,]    8    7   11   14   15    0
 Loc <- cmdscale(ddat)
 X <- Loc[,1]
 Y <- Loc[,2]
 plot(X,Y)
 plot(Loc)
 text(Loc, labels=c("A", "B","C", "D","E", "F"), lwd=3)
```

⇒ O Método Clássico é aplicado muitas vezes com o objetivo de obtenção de uma solução inicial (semeste para a solução não métrica)

Análise de Correspondência

Análise de Correspondência

u.a. / Variável Linha	Variável Coluna 1	2	…	j	…	J
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1J}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2J}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{iJ}
…	…	…	…	…		…
I	Y_{I1}	Y_{I2}		Y_{Ij}		Y_{IJ}

Identificar a estrutura dos dados multivariados com “Tabelas de Contingência”

Objetivos:

- Descrever graficamente os dados dispostos em tabelas de contingência, de tal forma a representar o padrão de associação entre variáveis $\Rightarrow$ os vetores linha e os vetores coluna da tabela são visualizados como pontos em um espaço vetorial
- Decompor a estatística χ^2 do teste de independência em tab. de contingência

TÉCNICA GRÁFICA MULTIDIMENSIONAL (similar ao Escalonamento!!)

(essencialmente descritiva, não adota qualquer modelo estrutural, auxilia a análise inferencial)

Análise de Correspondência

			Ano			
Jornal	**1976**	**1977**	**1978**	**1979**	**1980**	**Total**
A	64	58	67	59	60	308
B	18	18	23	20	17	96
C	12	10	9	12	9	52
D	36	25	34	31	27	153
E	29	21	25	20	20	115
F	133	115	116	107	89	560
G	34	28	30	26	29	147
H	178	143	180	150	148	799
I	8	8	5	6	6	33
J	101	113	143	112	107	576
K	66	56	60	58	53	293
L	87	69	79	68	69	372
M	23	19	17	19	17	95
N	34	24	29	26	23	136
O	70	56	60	55	50	291
P	29	20	25	19	18	111
Q	46	40	38	38	33	195
R	123	122	149	122	112	628
S	79	68	70	61	57	335
T	130	109	148	110	100	597
U	22	17	19	15	16	89
Total	1322	1139	1326	1134	1060	5981

Ao longo de 5 anos, em cada ano, cerca de 1000 pessoas de uma cidade foram amostradas e questionadas sobre quais jornais, dentre 21, eles liam regularmente.

Como representar o hábito de leitura de jornais dos cidadãos e sua variação ao longo do tempo?

Análise de Correspondência

Distribuição de 5.387 estudantes escoceses de acordo com a cor dos olhos e dos cabelos (Fisher, 1940)

	Cor do cabelo					
Cor olhos	**Claro**	**Ruivo**	**Médio**	**Escuro**	**Preto**	**Total**
Claros	688	116	584	188	4	1580
Azul	326	38	241	110	3	718
Médio	343	84	909	412	26	1774
Escuro	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Como descrever graficamente o padrão de associação entre as variáveis cor dos olhos e dos cabelos dos estudantes escoceses ?

Análise de Correspondência

Distribuição dos funcionários de uma empresa de acordo com o tabagismo.

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Para aderir a uma campanha nacional anti-tabagismo, o gerente de Recursos Humanos de uma empresa deseja conhecer o hábito de fumar dos funcionários. Os dados acima foram coletados para esta finalidade.

A representação gráfica dos dados é, em geral, de fácil entendimento. Como representar o padrão de associação entre o nível do funcionário e o hábito de fumar em um gráfico ?

Análise de Correspondência

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	Faixa Etária				
	< 16	16-17	17-18	18-19	19-20
Nenhum namorado	21	21	14	13	8
Namoro sem sexo	8	9	6	8	2
Namoro com sexo	2	3	4	10	10
Total	31	33	24	31	20

Como descrever graficamente o padrão de associação entre as variáveis faixa etária da adolescente e o tipo de namoro ?

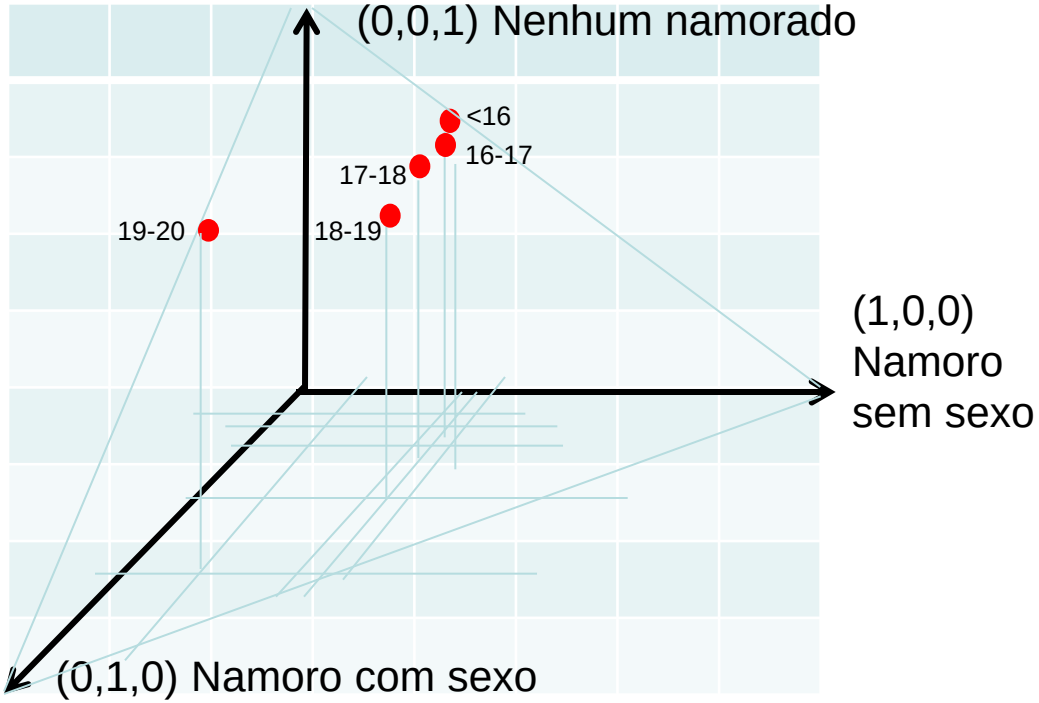
Análise de Correspondência
Representação Simplex

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	Faixa Etária				
	< 16	16-17	17-18	18-19	19-20
Nenhum namorado	21 (68)	21 (64)	14 (58)	13 (42)	8 (40)
Namoro sem sexo	8 (26)	9 (27)	6 (25)	8 (26)	2 (10)
Namoro com sexo	2 (6)	3 (9)	4 (17)	10 (32)	10 (50)
Total	31 (100%)	33 (100%)	24 (100%)	31 (100%)	20 (100%)

Como representar, em um gráfico bidimensional, as 5 faixas etárias? Note que os dados de cada coluna podem ser representados como variáveis Trinomiais.

NÃO há perda de informação nessa representação (as trinomiais estão originalmente na dimensão 2)



Análise de Correspondência
Representação Simplex

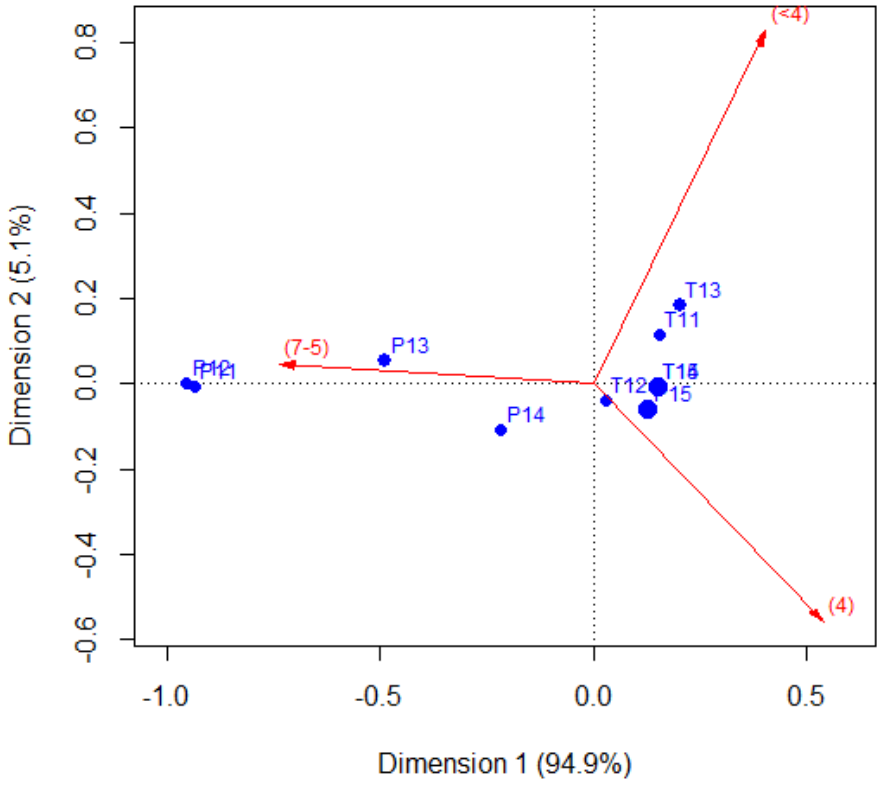
Distribuição do número de bulbilhos de alho de acordo com o tamanho (7-5, 4 e <4), tratamento e ano de plantio.

Ano	Tratamento*	Tamanho dos bulbilhos			Total
		7-5	4	<4	
2011	Padrão	417	36	0	453
	Teste	164	176	90	400
2012	Padrão	357	27	0	384
	Teste	169	161	54	384
2013	Padrão	800	240	103	1143
	Teste	412	458	274	1144
2014	Padrão	273	176	39	488
	Teste	185	220	83	488
2015	Padrão	1521	1794	585	3900
	Teste	1420	1681	635	3736

Represente as 10 trinomiais (variáveis nas linhas da tabela de contingência) no simplex. Este gráfico permite visualizar o padrão de heterogeneidade entre as populações trinomiais de acordo com o tamanho dos bulbilhos de alho. Interprete. Quais anos e qual tratamento produz os maiores bulbilhos?

Análise de Correspondência - Representação Simplex

Distribuição do número de bulbilhos de alho de acordo com tamanho, tratamento e ano de plantio.



A variável Tamanho do bulbilho de alho está em coordenadas padrão e Tratamento_Ano está em coordenadas principais. Neste caso, nenhuma informação é perdida nesta representação dos dados.

Bulbilhos de tamanho 7-5 estão mais associados ao tratamento Padrão em 2011 e 2012, seguidos de 2013. O tratamento Padrão em 2014 mostra associação (mais fraca) com bulbilhos tanto de tamanho 7-5 e 4. Já os tratamentos Teste de 2011 a 2015, bem como o tratamento Padrão em 2015, estão mais associados com bulbilhos de tamanho menor (4 e <4).

Análise de Correspondência

(Everitt, 2004)

- A AC em tabelas de contingência é um método de decomposição da estatística Qui-Quadrado em componentes que correspondem a “eixos principais” que mais explicam a heterogeneidade entre as variáveis coluna (ou linhas) da tabela.
- Método que simultaneamente atribui uma escala às linhas e, separadamente, uma escala às colunas da tabela de tal forma a maximizar a correlação entre as duas escalas.
- Método de obtenção de coordenadas para representar as categorias das variáveis linha e das variáveis coluna da tabela. O padrão de associação entre as variáveis fica representado graficamente $\Rightarrow$ é uma análise de Escalonamento Multidimensional para uma medida de distância específica para dados categorizados, conhecida como distância Qui-Quadrado.

Análise de Correspondência e Escalonamento Multidimensional

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Perfis Linha

	Variável Coluna			
Variável Linha	1	…	J	Total
1	$p_{11}=n_{11}/n_{1.}$		$p_{1J}=n_{1J}/n_{1.}$	1
…	…	…	…	…
I	$p_{I1}=n_{I1}/n_{I.}$		$p_{IJ}=n_{IJ}/n_{I.}$	1

$$\Rightarrow p_{ij}^{L} = \frac{n_{ij}}{n_{i.}}; \quad \bar{p}_{.j} = \frac{n_{.j}}{n}$$

Perfis Coluna

	Variável Coluna		
Variável Linha	1	…	J
1	$p_{11}=n_{11}/n_{.1}$		$p_{1J}=n_{1J}/n_{.J}$
…	…	…	…
I	$p_{I1}=n_{I1}/n_{.1}$		$p_{IJ}=n_{IJ}/n_{.J}$
Total	1	…	1

$$\Rightarrow p_{ij}^{C} = \frac{n_{ij}}{n_{.j}}; \quad \bar{p}_{i.} = \frac{n_{i.}}{n}$$

Análise de Correspondência e Escalonamento Multidimensional

1: Análise das Matrizes Quadradas D_{IxI} e D_{JxJ}

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Distância Qui-Quadrado – Perfis Linha

$$p_{ij}^{L} = \frac{n_{ij}}{n_{i.}} \qquad i = 1,2,...,I$$

Distância Euclidiana ponderada entre os perfis linha i e j (de proporções)

$$d_{ij}^{2\,Linhas} = \sum_{k=1}^{J} \frac{\left(p_{ik}^{L} - p_{jk}^{L}\right)^2}{\bar{p}_{.k}}$$

Distância Qui-Quadrado – Perfis Coluna

$$p_{ij}^{c} = \frac{n_{ij}}{n_{.j}} \qquad j = 1,2,...,J$$

Distância Euclidiana ponderada entre os perfis coluna i e j (de proporções)

$$d_{ij}^{2\,Colunas} = \sum_{k=1}^{I} \frac{\left(p_{ki}^{C} - p_{kj}^{C}\right)^2}{\bar{p}_{k.}}$$

$\Rightarrow$ Extrair as Coordenadas Principais das Matrizes de distâncias

D^{Linhas} e $D^{Colunas}$ $\Rightarrow$ resultados equivalentes à solução via *dvs* de Y^L e Y^C

que veremos a seguir .

Análise de Correspondência e Escalonamento Multidimensional

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	Faixa Etária				
	< 16	16-17	17-18	18-19	19-20
Nenhum namorado	21 (68)	21 (64)	14 (58)	13 (42)	8 (40)
Namoro sem sexo	8 (26)	9 (27)	6 (25)	8 (26)	2 (10)
Namoro com sexo	2 (6)	3 (9)	4 (17)	10 (32)	10 (50)
Total	31 (100%)	33 (100%)	24 (100%)	31 (100%)	20 (100%)

Matriz D_{IxI}

d_{ij} Colunas	**< 16**	**16-17**	**17-18**	**18-19**	**19-20**
<16	0,00	0,09	0,26	0,66	1,07
16-17		0,00	0,19	0,59	1,01
17-18			0,00	0,41	0,83
18-19				0,00	0,51
19-20					0,00

Matriz D_{JxJ}

d_{ij} Linhas	**Sem Nam**	**Nam**	**NamSexo**
Sem Nam	0,00	0,21	0,93
Nam		0,00	0,93
NamSexo			0,00

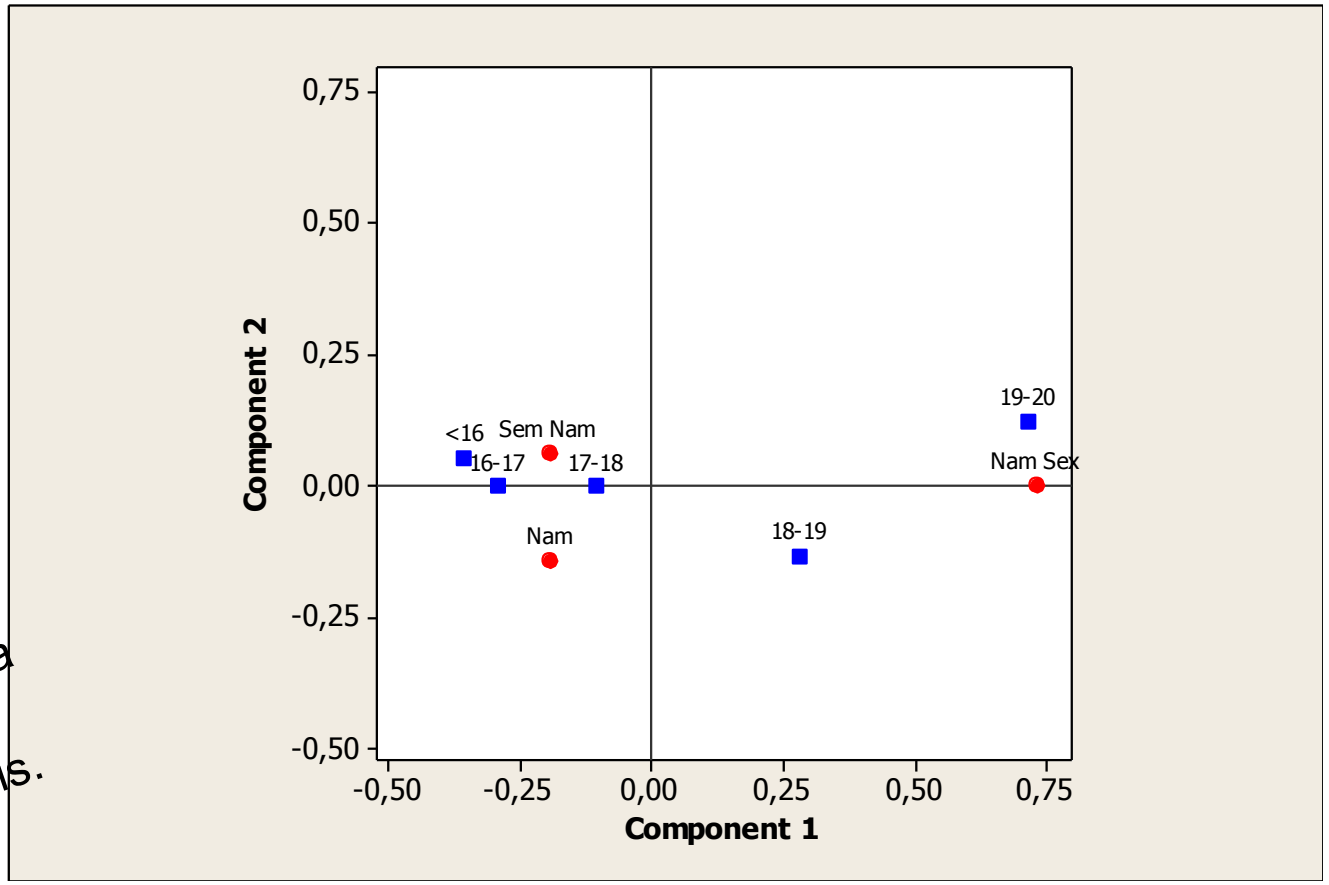
Obter as coordenadas principais (via Escalonamento Multidimensional) a partir das matrizes de distâncias Qui-Quadrado.

Análise de Correspondência e Escalonamento Multidimensional

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	CP1	CP2
Sem Nam	-0,1933	0,0610
Nam	-0,1924	-0,1425
Nam Sex	0,7322	0,0002
<16	-0,3547	0,0550
16-17	-0,2897	-0,0003
17-18	-0,1033	-0,0001
18-19	0,2806	-0,1342
19-20	0,7169	0,1234

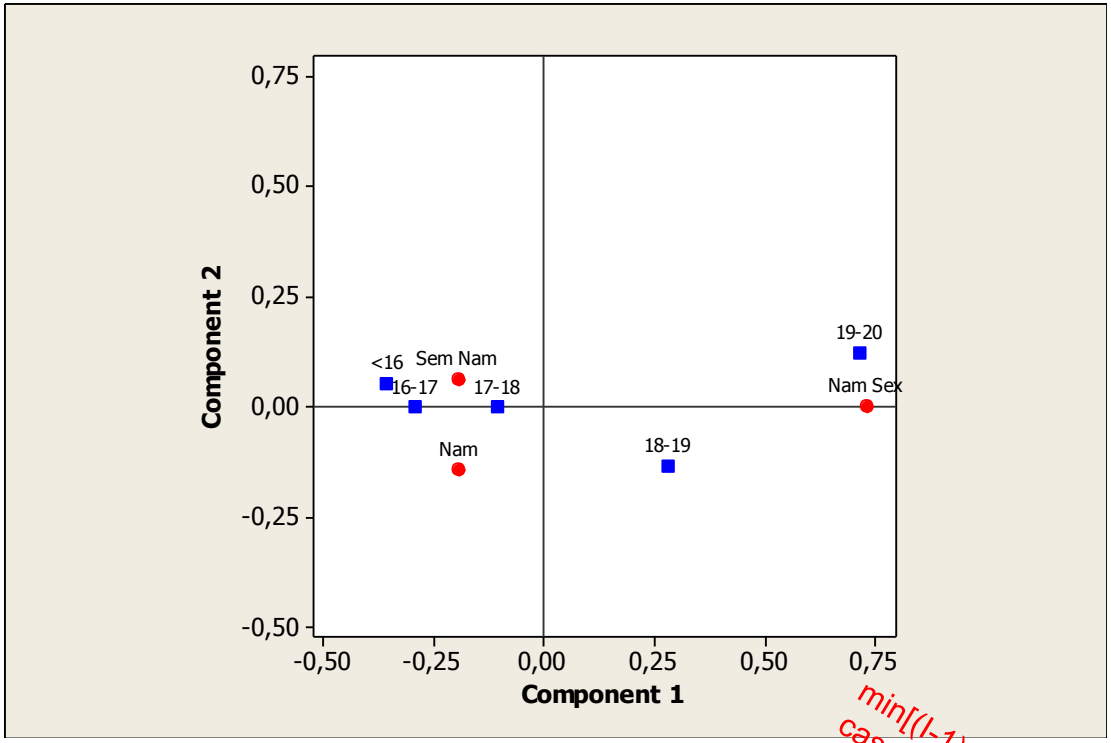
Variáveis Linha e Coluna representadas em Coordenadas Principais.



Análise de Correspondência e Escalonamento Multidimensional

Influência da idade da adolescente no tipo de namoro (Everitt, 2004)

	CP1	CP2
Sem Nam	-0,1933	0,0610
Nam	-0,1924	-0,1425
Nam Sex	0,7322	0,0002
<16	-0,3547	0,0550
16-17	-0,2897	-0,0003
17-18	-0,1033	-0,0001
18-19	0,2806	-0,1342
19-20	0,7169	0,1234



$$d_{Euclid}(<16,16-17) = \sqrt{(-0,3547+0,2897)^2 + (0,055+0,0003)^2} = 0,09$$

$$d_{Qui-Quad}(<16,16-17) = \sqrt{\frac{(0,68-0,64)^2}{0,55} + \frac{(0,26-0,27)^2}{0,24} + \frac{(0,06-0,09)^2}{0,21}} = 0,09$$

min[(I-1),(J-1)]=2. Nesse caso, com 2 eixos, não há perda de informação

Análise de Correspondência

2: A estatística Qui-Quadrado é uma medida de Distância Euclidiana Ponderada entre os Perfis Linha (ou Coluna) e o centróide

Representação dos Perfis Linha da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

Estatística Qui-Quadrado:

$$\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = \frac{\left(O_{11}-E_{11}\right)^2}{E_{11}} + \frac{\left(O_{12}-E_{12}\right)^2}{E_{12}} + \frac{\left(O_{IJ}-E_{IJ}\right)^2}{E_{IJ}}; \qquad O_{ij} = n_{ij} \quad E_{ij} = \frac{n_{i.}n_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(\frac{n_{i.}}{n_{i.}}O_{ij} - \frac{n_{i.}}{n_{i.}}E_{ij}\right)^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{i.}\frac{\left(p_{ij}^{L}-\bar{p}_{.j}\right)^2}{\bar{p}_{.j}} = \sum_{i=1}^{I} n_{i.}\sum_{j=1}^{J}\frac{\left(p_{ij}^{L}-\bar{p}_{.j}\right)^2}{\bar{p}_{.j}} = \sum_{i=1}^{I} n_{i.}d_i^2$$

Distância entre os perfis linha e centróide

massas

pesos

$$d_i^2 = \left(\mathbf{p}_i^L - \bar{\mathbf{p}}^L\right)' \mathbf{D}_{\bar{\mathbf{p}}}^{-1}\left(\mathbf{p}_i^L - \bar{\mathbf{p}}^L\right)$$

$\mathbf{p}_i^L$: perfil de freqüências relativas da linha i

Análise de Correspondência

Representação dos Perfis Linha da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$\mathbf{p}_i^L = \left(p_{i1}^L, p_{i2}^L, ..., p_{iJ}^L \right)'; \quad p_{ij}^L = \frac{n_{ij}}{n_{i.}}; \quad \sum_{j=1}^{J} p_{ij}^L = 1$$

$$\bar{\mathbf{p}}^L = \left(\frac{n_{.1}}{n}, \frac{n_{.2}}{n}, ..., \frac{n_{.J}}{n} \right)'; \quad \bar{p}_{.j}^L = \frac{n_{.j}}{n}; \quad \sum_{j=1}^{J} \bar{p}_{.j}^L = 1$$ Centróide (linha)

$$d_i^2 = \left(\mathbf{p}_i^L - \bar{\mathbf{p}}^L \right)' \mathbf{D}_{\bar{\mathbf{p}}}^{-1} \left(\mathbf{p}_i^L - \bar{\mathbf{p}}^L \right) = \sum_{j=1}^{J} \frac{\left(p_{ij}^L - \bar{p}_{.j}^L \right)^2}{\bar{p}_{.j}^L}$$ Distância Euclidiana ponderada ao quadrado do perfil de freqüências relativas da linha i ao centróide

Como representar tais perfis linha em um espaço multidimensional?

Análise de Correspondência

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$\mathbf{p_i} = \left(p_{i1}, p_{i2}, ..., p_{iJ}\right)' \qquad \bar{\mathbf{p}} = \left(\frac{n_{.1}}{n}, \frac{n_{.2}}{n}, ..., \frac{n_{.J}}{n}\right) \qquad d_i^2 = \left(\mathbf{p_i} - \bar{\mathbf{p}}\right)' \mathbf{D}_{\bar{\mathbf{p}}}^{-1} \left(\mathbf{p_i} - \bar{\mathbf{p}}\right)$$

$$\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{\left(O_{11} - E_{11}\right)^2}{E_{11}} + \frac{\left(O_{12} - E_{12}\right)^2}{E_{12}} + ... + \frac{\left(O_{IJ} - E_{IJ}\right)^2}{E_{IJ}};$$

$$in(I) = \chi^2 / n$$

Ponderação da estatística por n

in(I): medida de Inércia total do conjunto dos I perfis. Mede a variação dos perfis individuais $\mathbf{p_i}$ em torno do centróide $\bar{\mathbf{p}}$.

O objetivo da análise de Correspondência é encontrar um subespaço de "baixa" dimensão que melhor contenha os perfis $\mathbf{p_i}$

3: Análise da Matriz rectangular Y_{IxJ} e sua dvs

Análise de Correspondência

$$\mathbf{Y}^{L}_{I\times J} = \begin{pmatrix} \mathbf{p}^{L}_{1_{1\times I}} \\ \dots \\ \mathbf{p}^{L}_{I_{1\times I}} \end{pmatrix}$$

Matriz de dados (frequências relativas) com a soma de cada linha igual a uma constante c (c=1). O vetor centróide, é dado por:

$$\bar{\mathbf{p}}^{L} = \left(\bar{p}^{L}_{1}, \dots, \bar{p}^{L}_{J} \right)$$

Considere as seguintes matrizes:

Matriz de pesos: $D_{pesos\,J\times J} = diag\left(1/\ \bar{p}^{L}_{j} = n/n_{.j}\right) = D^{-1}_{\bar{p}}$

Matriz de massas: $D_{massas I\times I} = diag\left(n_{i.}/n\right)$ associada à marginal fixada

Então, os eixos principais dos perfis linha $\mathbf{p_i}$ podem ser obtidos da decomposição em valores singulares de ***Y*** . Para k dimensões e com I<J, tem-se:

$$Y^{L}_{I\times J} = N^{L}_{I\times I}\ D^{L}_{\lambda I\times I}\ M'^{L}_{I\times J} \quad ; \quad N^{L}{}' D^{L}_{massas} N^{L} = M^{L}{}' D^{L}_{pesos} M^{L} = I \Rightarrow X^{L}_{I\times k} = N^{L}_{I\times I} D^{L}_{\lambda_{I\times k}}$$

Eixos principais das linhas

$$\Rightarrow in(I) = \chi^2/n = \sum_{i=1}^{I} \lambda_i^2$$: inércia total $$\Rightarrow \frac{\lambda^2{}_1 + \lambda^2{}_2 + \dots + \lambda^2{}_k}{\sum_i \lambda_i^2}$$: proporção da inércia descrita pelos eixos

Análise de Correspondência

Nível do funcionário vs tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Expected Frequencies

	F0	F1	F2	F3
N1	3,48	2,56	3,53	1,42
N2	5,69	4,20	5,78	2,33
N3	16,12	11,89	16,38	6,61
N4	27,81	20,52	28,27	11,40
N5	7,90	5,83	8,03	3,24

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Chi-Square Distances

	F0	F1	F2	F3	Total
N1	0,079	0,124	0,081	0,232	0,516
N2	0,502	0,341	0,256	1,194	2,293
N3	4,893	0,301	1,173	1,028	7,395
N4	3,463	0,591	0,792	0,225	5,070
N5	0,557	0,005	0,132	0,474	1,168
Total	9,493	1,362	2,434	3,153	16,442

0,232/16,442

Relative Inertias

	F0	F1	F2	F3	Total
N1	0,005	0,008	0,005	0,014	0,031
N2	0,031	0,021	0,016	0,073	0,139
N3	0,298	0,018	0,071	0,063	0,450
N4	0,211	0,036	0,048	0,014	0,308
N5	0,034	0,000	0,008	0,029	0,071
Total	0,577	0,083	0,148	0,192	1,000

Estatística $\chi^2 = 16,442$ (p=0,172)

Inércia total=16,442/193=0,08518

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Perfis Linha da Tabela

Y_{5x4}

	F0	F1	F2	F3	Mass
N1	0,364	0,182	0,273	0,182	0,057
N2	0,222	0,167	0,389	0,222	0,093
N3	0,490	0,196	0,235	0,078	0,264
N4	0,205	0,273	0,375	0,148	0,456
N5	0,400	0,240	0,280	0,080	0,130
Mass	0,316	0,233	0,321	0,130	

Vetor de pesos (inverso)

Vetor de massas

$$\Rightarrow \quad Y_{5\times4} = N\ D_{\lambda}\ M' \qquad X^{L}_{(k=2)} = N_{5\times5} D_{\lambda_{5\times2}}$$

$\lambda = (\ 0{,}2734 \quad 0{,}1001 \quad 0{,}0203\ )$

$$X^{L}_{(k=2)} = \begin{pmatrix} 0{,}066 & -0{,}194 \\ -0{,}259 & -0{,}243 \\ 0{,}381 & -0{,}011 \\ -0{,}233 & 0{,}058 \\ 0{,}201 & 0{,}078 \end{pmatrix}$$

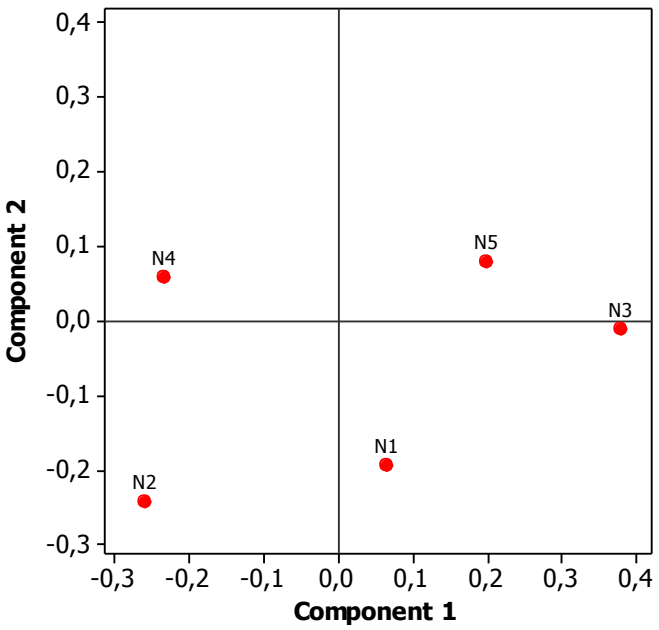
Nível1
Nível2
Nível3
Nível4
Nível5

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Representação dos perfis linha



$$\Rightarrow in(eixo1) = \lambda_1^2 = (0{,}2734)^2 = 0{,}0748$$

$$\Rightarrow in(eixo2) = \lambda_2^2 = (0{,}1001)^2 = 0{,}01$$

0,0848/0,08518=0,995

$\Rightarrow$ 99,5% da inércia total dos dados está representada no plano

$\Rightarrow$ os funcionários níveis N5 e N3 são mais semelhantes em seu hábito de fumar. N2 e N4 estão mais distantes deste grupo, sendo mais semelhantes entre si. N1 ocupa uma posição intermediária entre estes grupos.

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Problema Dual:

Estudar o padrão de variação da variável hábito de fumar em função do nível funcional na empresa

$\Rightarrow$ Como representar os perfis das frequências relativas das colunas?

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$p_{ij}^{c} = \frac{n_{ij}}{n_{.j}}; \qquad \sum_{i=1}^{I} p_{ij}^{c} = 1$$

Estatística Qui-Quadrado:

$$\chi^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{\left(O_{11} - E_{11}\right)^2}{E_{11}} + \frac{\left(O_{12} - E_{12}\right)^2}{E_{12}} + \ldots + \frac{\left(O_{IJ} - E_{IJ}\right)^2}{E_{IJ}}; \qquad O_{ij} = n_{ij} \quad E_{ij} = \frac{n_{i.} n_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{.j} \frac{\left(p_{ij}^{C} - \bar{p}_{i.}^{C}\right)^2}{\bar{p}_{i.}^{C}} = \sum_{j=1}^{J} n_{.j} \sum_{i=1}^{I} \frac{\left(p_{ij}^{C} - \bar{p}_{i.}^{C}\right)^2}{\bar{p}_{i.}^{C}} = \sum_{j=1}^{J} n_{.j} d_{.j}^2$$

Distância entre os perfis coluna e o centróide

pesos

massas

$$d_{.j}^2 = \left(\mathbf{p}_j^{C} - \bar{\mathbf{p}}^{C}\right)' \mathbf{D}_{\bar{\mathbf{p}}^{c}}^{-1} \left(\mathbf{p}_j^{C} - \bar{\mathbf{p}}^{C}\right)$$

$\mathbf{p}_j^{C}$: perfil de freqüências relativas da coluna j

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

$$\mathbf{p}_j^C = \left(p_{1j}^C, p_{2j}^c, ..., p_{Ij}^C \right)'; \qquad p_{ij}^C = \frac{n_{ij}}{n_{.j}}; \quad \sum_{i=1}^{I} p_{ij}^C = 1$$

$$\bar{\mathbf{p}} = \bar{\mathbf{p}}^C = \left(\frac{n_{1.}}{n}, \frac{n_{2.}}{n}, ..., \frac{n_{I.}}{n} \right)'; \quad \sum_{i=1}^{I} \bar{p}_{i.}^C = 1$$ Centróide (coluna)

$$d_j^2 = \left(\mathbf{p}_j^C - \bar{\mathbf{p}}^C \right)' \mathbf{D}_{\bar{\mathbf{p}}^c}^{-1} \left(\mathbf{p}_j^C - \bar{\mathbf{p}}^C \right) = \sum_{i=1}^{I} \frac{\left(p_{ij}^C - \bar{p}_i^C \right)^2}{\bar{p}_i^C}$$ Distância Euclidiana ponderada ao quadrado do perfil de freqüências relativas da coluna j ao centróide

Como representar tais perfis coluna em um espaço multidimensional?

Análise de Correspondência

Representação dos Perfis Linha e Coluna da Tabela

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

linhas

$$\mathbf{Y}^{L}{}_{I\times J} = \begin{pmatrix} \mathbf{p}^{L}_{1_{1\times J}} \\ \dots \\ \mathbf{p}^{L}_{I_{1\times J}} \end{pmatrix}$$

$$\bar{\mathbf{p}}^{L}_{1\times J} = \left(\bar{p}^{L}_{.1},\dots,\bar{p}^{L}_{.J}\right)'$$

$$D^{L}_{pesos\,J\times J} = diag\left(1/\,\bar{p}^{L}_{.j}\right)$$

$$D^{L}_{massa\,I\times I} = diag\left(n_{i.}/n\right)$$

$\Rightarrow$

colunas

$$\mathbf{Y}^{C}{}_{J\times I} = \begin{pmatrix} \mathbf{p}^{C}_{1_{1\times I}} \\ \dots \\ \mathbf{p}^{C}_{J_{1\times I}} \end{pmatrix} \qquad \bar{\mathbf{p}}^{C}{}_{1\times I} = \left(\bar{p}^{C}_{1.},\dots,\bar{p}^{C}_{I.}\right)'$$

$$D^{C}_{peso\,I\times I} = diag\left(1/\,\bar{p}^{C}_{i.}\right)$$: matriz de pesos

$$D^{C}_{massa\,J\times J} = diag\left(n_{.j}/n\right)$$: matriz de massas

Análise de Correspondência

Representação dos Perfis Coluna da Tabela

Perfis Linha

$$\mathbf{Y}^{L}{}_{I\times J} = \begin{pmatrix} \mathbf{p}^{L}_{1_{1\times I}} \\ \dots \\ \mathbf{p}^{L}_{I_{1\times I}} \end{pmatrix}$$

Perfis Coluna

$$\mathbf{Y}^{C}{}_{J\times I} = \begin{pmatrix} \mathbf{p}^{C}_{1_{1\times I}} \\ \dots \\ \mathbf{p}^{C}_{J_{1\times I}} \end{pmatrix}$$

$$\bar{\mathbf{p}}^{C} = \left(\bar{p}^{C}_{1.},\dots,\bar{p}^{C}_{I.}\right)'$$

$$D^{C}_{pesos I\times I} = diag\left(1/\,\bar{p}^{C}_{i.}\right) \qquad D^{C}_{massas J\times J} = diag\left(n_{.j}/n\right)$$

Obter os eixos principais dos perfis colunas $\mathbf{p}^{C}_{j} \Rightarrow$ obter a decomposição espectral da matriz $\boldsymbol{Y^{C}}$, tal que, para dimensões de ordem k, tem-se:

Eixos principais das colunas

$$X^{C}_{(k)} = N^{C}_{J\times J} D^{C}_{\lambda(k)} \Rightarrow Y^{C}{}_{J\times I} = N^{C}\, D^{C}_{\lambda}\, M^{C\,\prime} \quad ; \quad N^{C\,\prime} D^{C}_{massa} N^{C} = M^{C\,\prime} D^{C}_{peso} M^{C} = I$$

$$\Rightarrow D^{C}_{\lambda} = D^{L}_{\lambda}$$

$\Rightarrow$ Os valores singulares da representação dos perfis linha e coluna são os mesmos (exceto pelos autovalores nulos)
$\Rightarrow$ o subespaço ótimo para a representação dos perfis linha e coluna é o mesmo !!

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Perfis Coluna da Tabela

$Y^{C'}_{4x5}$

	F0	F1	F2	F3	Mass
N1	0,066	0,044	0,048	0,080	0,057
N2	0,066	0,067	0,113	0,160	0,093
N3	0,410	0,222	0,194	0,160	0,264
N4	0,295	0,533	0,532	0,520	0,456
N5	0,164	0,133	0,113	0,080	0,130
Mass	0,316	0,233	0,321	0,130	

$$\Rightarrow \quad Y^{C'}{}_{4\times5} = N^{C}\ D^{C}_{\lambda}\ M^{C'} \qquad X^{C}_{(k)} = N^{C}_{(k)} D_{\lambda(k)}$$

$\lambda = (\ 0{,}2734 \quad 0{,}1001 \quad 0{,}0203\ )$

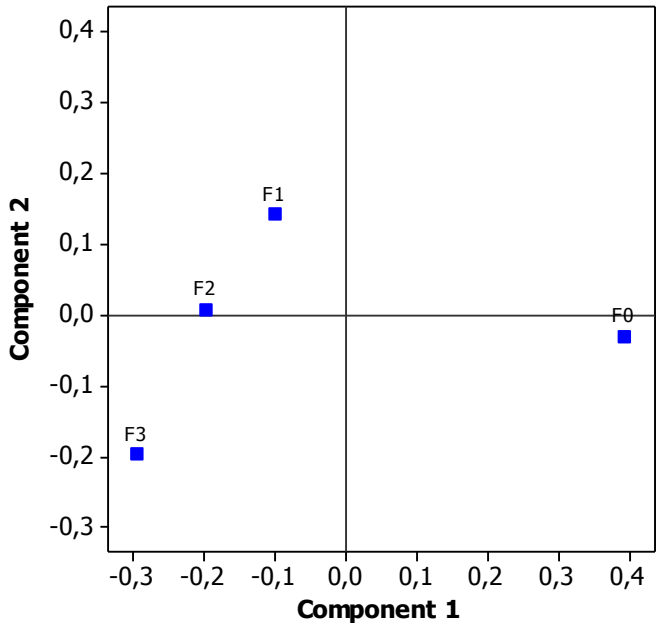
$$X^{C}_{(k=2)} = \begin{pmatrix} 0{,}393 & -0{,}031 \\ -0{,}1 & 0{,}141 \\ -0{,}196 & 0{,}007 \\ -0{,}294 & -0{,}198 \end{pmatrix} \begin{matrix} \leftarrow F0 \\ \leftarrow F1 \\ \leftarrow F2 \\ \leftarrow F3 \end{matrix}$$

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Representação dos perfis coluna



$$\Rightarrow in(eixo1) = \lambda_1^2 = (0{,}2734)^2 = 0{,}0748$$

$$\Rightarrow in(eixo2) = \lambda_2^2 = (0{,}1001)^2 = 0{,}01$$

0,0848/0,08518=0,995

$\Rightarrow$ 99,5% da inércia total dos dados está representada no plano

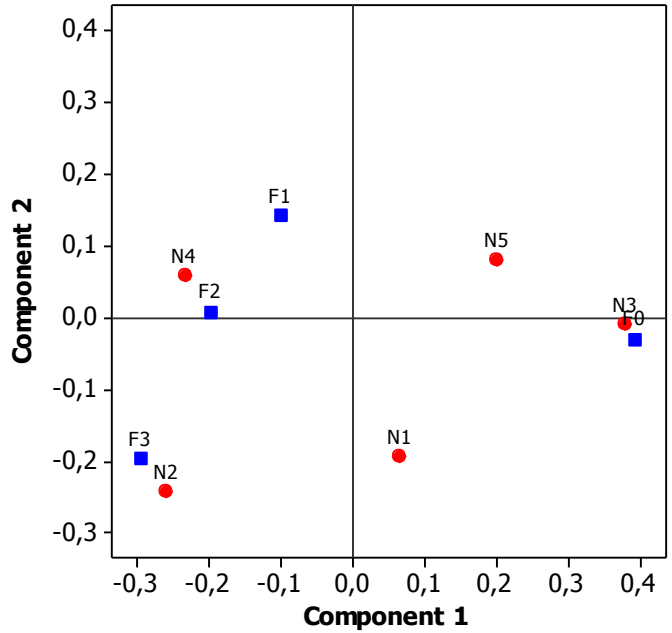
$\Rightarrow$ disposição linear (C1) dos níveis de hábito de fumar. O grupo de não fumantes está bem distante dos demais

Análise de Correspondência

Distribuição de funcionários de acordo com o tabagismo

	Hábito de Fumar				
Funcionário	**Não**	**Pouco**	**Médio**	**Muito**	**Total**
Nível 1	4	2	3	2	11
Nível 2	4	3	7	4	18
Nível 3	25	10	12	4	51
Nível 4	18	24	33	13	88
Nível 5	10	6	7	2	25
Total	61	45	62	25	193

Representação dos perfis linha e coluna



Biplot: Representação conjunta dos perfis de frequência relativa das linhas e colunas da tabela

Análise de Correspondência

Exemplo: (Greenacre, 2007)

Tabela A

	C1	C2	C3
L1	11	10	9
L2	10	11	9
L3	10	9	10
L4	9	9	12
L5	10	11	10

	C1	C2	C3
L1	0.37	0.33	0.30
L2	0.33	0.37	0.30
L3	0.34	0.31	0.34
L4	0.30	0.30	0.40
L5	0.32	0.35	0.32

X2=1.133, p=0.9973

Tabela C

	C1	C2	C3
L1	17	5	3
L2	3	20	4
L3	19	5	2
L4	6	8	35
L5	5	12	6

	C1	C2	C3
L1	0.68	0.20	0.12
L2	0.11	0.74	0.15
L3	0.73	0.19	0.08
L4	0.12	0.16	0.71
L5	0.22	0.52	0.26

X2=88.843, p=7.982e-16

Tabela B

	C1	C2	C3
L1	13	8	9
L2	6	14	10
L3	14	7	8
L4	7	9	18
L5	10	12	5

	C1	C2	C3
L1	0.43	0.27	0.30
L2	0.20	0.47	0.33
L3	0.48	0.24	0.28
L4	0.21	0.26	0.53
L5	0.37	0.44	0.19

X2=16.513, p=0.0356

Tabela D

	C1	C2	C3
L1	20	1	0
L2	0	24	1
L3	24	2	0
L4	2	0	47
L5	4	23	2

	C1	C2	C3
L1	0.95	0.05	0.00
L2	0.00	0.96	0.04
L3	0.92	0.08	0.00
L4	0.04	0.00	0.96
L5	0.14	0.79	0.07

X2=235.731, p< 2.2e-16

Em cada caso calcule: vetores de proporções das ***trinomiais*** (linha), centróide, vetor de massas, vetor de pesos, distância Qui-Quadrado entre L1 e L2 e entre L1 (L2) e o centróide, inércia total. Obtenha a representação das 5 trinomiais no simplex correspondente. Interprete.

Análise de Correspondência

5 populações trinomiais

Tabela B

	C1	C2	C3	Total
L1	13	8	9	30
	0,43	0,27	0,30	0,20
L2	6	14	10	30
	0,20	0,47	0,33	0,20
L3	14	7	8	29
	0,48	0,24	0,28	0,19
L4	7	9	18	34
	0,21	0,26	0,53	0,23
L5	10	12	5	27
	0,37	0,44	0,19	0,18
Total	50	50	50	150
	0,33	0,33	0,33	1

Vetor de massas

Vetor centróide (baricentro)

0,33=0,20(0,43)+0,20(0,20)+0,19(0,48)+0,23(0,21)+0,18(0,37)

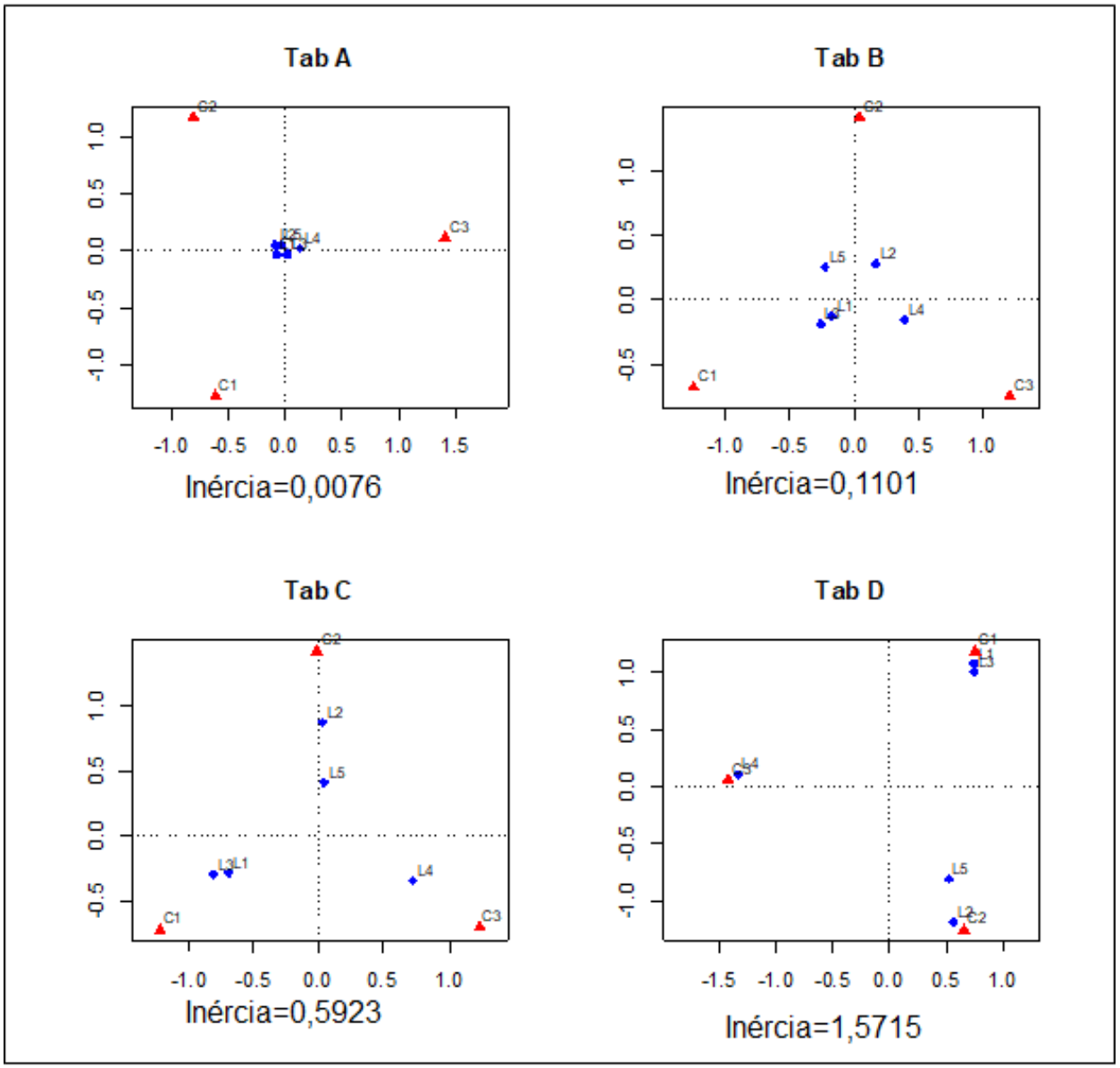
Análise das Linhas

$$d_{ij}^{2\,Linhas} = \sum_{k=1}^{J} \frac{\left(p_{ik}^{L} - p_{jk}^{L}\right)^2}{\bar{p}_{.k}}$$

$$d_{i}^{L2} = \sum_{j=1}^{J} \frac{\left(p_{ij}^{L} - \bar{p}_{.j}^{L}\right)^2}{\bar{p}_{.j}^{L}}$$

Inércia=16,513/150=0,1101

BiPlot e Inércias



Representação BiPlot
Mapa Assimétrico

- Linhas (trinomiais) em Coordenadas Principais

-Colunas em Coordenadas Padrão (vértices do simplex)

Existem diferentes construções do BiPlot, visando diferentes padronizações dos eixos.
Ex.: Mapa Simétrico: ambos, variáveis das Linhas e Colunas, em coordenadas principais.

Análise de Correspondência e Escalonamento Multidimensional

	Variável Coluna					
Variável Linha	1	…	j	…	J	Total
1	n_{11}		n_{1j}		n_{1J}	$n_{1.}$
…	…	…	…		…	
i	n_{i1}		n_{ij}	…	n_{iJ}	$n_{i.}$
…	…	…	…		…	
I	n_{I1}		n_{Ij}		n_{IJ}	$n_{I.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.J}$	n

- Em um gráfico de coordenadas principais representando somente as categorias linha (ou coluna) da tabela, as distâncias entre os pontos são distâncias Euclidianas.

- MAS, em um gráfico onde ambos os espaços (linha e coluna) estão representados simultaneamente, é preciso ter cuidado com a comparação entre categorias das linhas e colunas pois, neste caso, a medida de distância Euclidiana pode não ser válida $\Rightarrow$ uma melhor aproximação pode ser obtida com a padronização das coordenadas principais (dividir os valores das coordenadas pela raiz quadrada da inércia do componente) $\Rightarrow$ Coordenadas Assimétricas

Análise de Fatores (Análise Fatorial)

(Análise de Fatores Comuns e Específicos)

Análise de CP - Análise Fatorial Exploratória

Como obter as variáveis originais a partir das componentes principais?

$$Y_{n\times p}\ ;\quad Y_i \overset{iid}{\sim} (\mu;\Sigma);\quad \Sigma = P\Lambda P'\ ;\quad P=(a_{ij});\quad P'=(a_{ji})$$

$$Z_{ji} = P_j'\, Y_i = a_{1j}Y_{1i} + a_{2j}Y_{2i} + \ldots + a_{pj}Y_{pi}$$

j-ésima coluna de P

autovetores

$$P_{p\times p} = \begin{pmatrix} P_1 & P_2 & & P_p \\ a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ & & \ldots & \\ a_{p1} & a_{p2} & \ldots & a_{pp} \end{pmatrix}$$

$$Z_i = P'\,Y_i \Leftrightarrow P\,Z_i = P\,P'\,Y_i = Y_i \Rightarrow Y_i = P\,Z_i$$

$$Y_{ij} = a_{j1}Z_{1i} + a_{j2}Z_{2i} + \ldots + a_{jp}Z_{pi}$$

j-ésima linha de P

Pense no sistema de equações das p variáveis Y_{ij} ($j=1,2,\ldots,p$) definidas em função de um conjunto de "fatores comuns" Z_k , com $k=1,2,\ldots,m$, $m<p$. Este é um dos objetivos da *Análise Fatorial Exploratória*.

Análise Fatorial

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

Objetivos:

- Decompor a covariação entre as *p*-variáveis por meio de *m*-fatores comuns (à todas as *p* variáveis) mais fatores específicos

Análise Fatorial Exploratória $\Rightarrow$ obter constructos, variáveis latentes (não observáveis) que expliquem a correlação entre as variáveis.

Análise Fatorial Confirmatória $\Rightarrow$ verificar se uma específica estrutura (decomposição ou grafo) se ajusta aos dados (à matriz de correlações)

Avaliar se Σ pode ser estruturada ou decomposta em fatores comuns + específicos

Análise Fatorial Exploratória

Como explicar o comportamento de variáveis observadas em função de um conjunto de variáveis latentes (não observáveis, constructos)?

$$\mathrm{Y}_{i_{p\times 1}} \overset{iid}{\sim} (\mu;\Sigma) \quad \Rightarrow \quad Y_{ij} = \mu_j + \phi_{j1}F_{1i} + \phi_{j2}F_{2i} + ... + \phi_{jm}F_{mi} + e_{ij}$$

Modelo estrutural:

$$\begin{cases} Y_{i1} - \mu_1 = \phi_{11}F_{1i} + \phi_{12}F_{2i} + ... + \phi_{1m}F_{mi} + e_{1i} \\ Y_{i2} - \mu_2 = \phi_{21}F_{1i} + \phi_{22}F_{2i} + ... + \phi_{2m}F_{mi} + e_{2i} \\ ... \\ Y_{ip} - \mu_p = \phi_{p1}F_{1i} + \phi_{p2}F_{2i} + ... + \phi_{pm}F_{mi} + e_{pi} \end{cases}$$

$\mathbf{f} = (F_1,...,F_m)'$: fatores comuns

$e = (e_1,...,e_p)'$: fatores específicos

$\Phi = (\phi_{ij})$: cargas fatoriais

Notação Matricial

$$Y_i - \mu = \Phi_{p\times m}\,\mathbf{f}_{i\ m\times 1} + e_{i\ p\times 1}$$

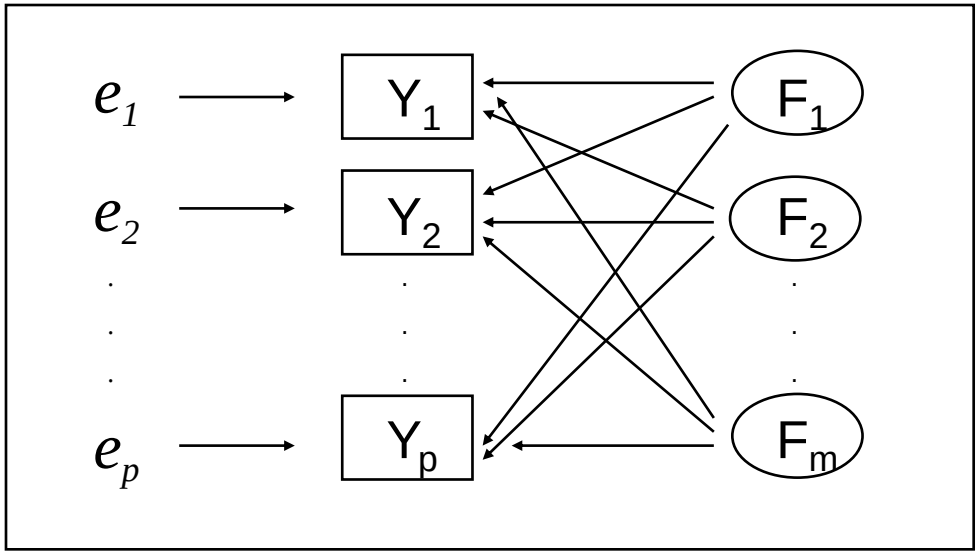
Como obter $\begin{cases} \Phi_{p\times m} \\ \mathbf{f}_{i\ m\times 1} \end{cases}$?

Análise Fatorial Confirmatória

$$Y_i \Rightarrow \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i$$

Equações de mensuração

Diagrama de Caminhos (Grafo) de um modelo de Análise Fatorial ortogonal



Var. Observadas: retângulos

Var. Latentes (constructo): círculos

Erros: sem representação gráfica

As setas (direcionadas) partem de uma variável independente e atingem uma variável dependente

Se existirem correlações (entre os fatores específicos ou entre os comuns), estas devem ser representadas por arcos

Modelo de Equações Estruturais (MEE)

- Equações estruturais das variáveis latentes
- Equações estruturais das variáveis observadas (equações de mensuração)

(Bollen, 1989)

Pacotes computacionais: LISREL, Lavan-R, (fa-R)

Análise Fatorial

- Modelo estrutural: $Y_i - \mu = \Phi\, \mathbf{f}_i + e_i$
- Suposições do modelo de fatores ortogonais (**f** e e são variáveis aleatórias):

$$\mathbf{f}_{i_{m\times 1}} = \begin{bmatrix} F_{1i} \\ \dots \\ F_{mi} \end{bmatrix} \overset{iid}{\sim} (0; I_m); \quad e_{i_{p\times 1}} \overset{iid}{\sim} \left(0; \Psi = diag(\psi_1, \dots, \psi_p)\right); \quad Cov(\mathbf{f}, e) = 0$$

Matriz de Covariância (marginal) de Y:

Decomposição muito particular da matriz Σ!

$$Cov(Y_i) = \Sigma_{p\times p} = Cov(\Phi \mathbf{f}_i + e_i) \Rightarrow \boxed{\Sigma = \Phi\, \Phi' + \Psi}$$

componente de covariâncias devido ao fator comum

componente de variâncias devido ao fator específico

$$Var(Y_{ij}) = \phi_{j1}^2 + \phi_{j2}^2 + \dots + \phi_{jm}^2 + \psi_j = h_j^2 + \psi_j$$

comunalidade da variável Y_{ij}

especificidade de Y_{ij}

Análise Fatorial

$$Y_i \in \Re^p; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \Leftrightarrow \quad \Sigma = \Phi\Phi' + \Psi$$

$$Var(Y_{ij}) = \phi_{j1}^2 + \phi_{j2}^2 + ... + \phi_{jm}^2 + \psi_j = h_j^2 + \psi_j$$

$$\overline{h}_j^2 = \frac{h_j^2}{Var\left(Y_{ij}\right)}$$

% da Var(Y_{ij}) explicada pelo conjunto dos fatores comuns

$$H^2 = \frac{\sum_{j=1}^{p} h_j^2}{\sum_{j=1}^{p} Var\left(Y_{ij}\right)}$$

% da Variância Total de Y explicada pelo conjunto dos fatores comuns

$$H_{F_k}^2 = \frac{\sum_{j=1}^{p} \varphi_{jk}^2}{\sum_{j=1}^{p} Var\left(Y_{ij}\right)}$$

% da Variância Total explicada pelo fator comum Fk

$$Cov(Y_{ij}, Y_{ij'}) = \phi_{j1}\phi_{j'1} + \phi_{j2}\phi_{j'2} + ... + \phi_{jm}\phi_{j'm} \quad \Rightarrow \text{depende somente de fatores comuns}$$

$$Cov(Y_{ij}, F_{ki}) = Cov\left(\varphi_{j1}F_{1i} + ... + \varphi_{jk}F_{ki} + ... + \varphi_{jm}F_{mi}; F_{ki}\right) = Cov\left(\varphi_{jk}F_{ki}; F_{ki}\right) = \varphi_{jk}$$

$$Corr(Y_{ij}, F_{ki}) = \varphi_{jk} / \sqrt{Var\left(Y_{ij}\right)} = \varphi_{jk} / \sqrt{h_j^2 + \psi_j}$$

Análise Fatorial

$$Y_i \in \Re^p; \quad Y_i - \mu = \Phi\, \mathbf{f}_i + e_i$$

$$\Leftrightarrow \quad \Sigma = \Phi\Phi' + \Psi$$

$$\begin{cases} Y_{1i} - \mu_1 = \phi_{11}F_{1i} + \phi_{12}F_{2i} + \ldots + \phi_{1m}F_{mi} + e_{1i} \\ Y_{2i} - \mu_2 = \phi_{21}F_{1i} + \phi_{22}F_{2i} + \ldots + \phi_{2m}F_{mi} + e_{2i} \\ \ldots \\ Y_{pi} - \mu_p = \phi_{p1}F_{1i} + \phi_{p2}F_{2i} + \ldots + \phi_{pm}F_{mi} + e_{pi} \end{cases}$$

Não unicidade das cargas e possibilidade de rotacionar soluções

Para qualquer matriz orthogonal Γ, tem-se:

$$\Phi^* = \Phi\Gamma; \quad \Gamma\Gamma' = I \quad \Rightarrow \Phi^*\Phi^{*\prime} + \Psi = \Phi\Gamma\Gamma'\Phi + \Psi = \Sigma$$

Como obter:

- Matriz de Coeficientes (Φ)
- Componentes Específicos (Ψ)
- Escores Fatoriais (F_{ki})

- Via Componentes Principais
- Via Máxima Verossimilhança

Análise Fatorial via Componentes Principais

$$Y_i \in \Re^p; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \Leftrightarrow \quad \Sigma = \Phi\Phi' + \Psi$$

$$\Sigma = P\Lambda P' \quad \Rightarrow \quad \Sigma = \lambda_1 P_1 P_1' + \ldots + \lambda_m P_m P_m' + \ldots + \lambda_p P_p P_p'$$

$\Re^p \rightarrow R^m$

$$\Sigma \approx \lambda_1 P_1 P_1' + \ldots + \lambda_m P_m P' = \Phi\Phi'$$

Aproximação usando *m* componentes
$\Rightarrow$ define os termos comuns!

Note que:
$\sum_{j=1}^{p} \phi_{jk}^{\ 2} = \lambda_k$

- $\Phi = (\phi_1, \ldots, \phi_m) = \left(\sqrt{\lambda_1} P_1, \ldots, \sqrt{\lambda_m} P_m\right) \Rightarrow \phi_{jk} = \sqrt{\lambda_k}\ a_{jk}$

- $\Psi \Rightarrow \Psi = diag\left(\sigma_1^2 - h_1^2, \ldots, \sigma_p^2 - h_p^2\right) \Rightarrow \psi_j = \sigma_{jj} - \sum_{k=1}^{m} \phi_{jk}^2$

Componente específico

- Qual o valor do escore factorial? $\mathbf{f}_i$, $i = 1,2,\ldots,n$

$$Y_{i\,(p\times 1)} \Rightarrow \qquad Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \Rightarrow \quad \mathbf{f}_i = Z_i D_{\lambda_j}^{-1/2}$$

Os escores fatoriais são os componentes principais padronizados

Análise Fatorial via Componentes Principais

Obtenção do modelo de fatores comuns e específicos

$$Y_i \in \Re^p; \quad Y_i - \mu = \Phi \mathbf{f}_i + e_i \quad \Leftrightarrow \quad \Sigma_{p\times p} = \Phi_{p\times m}\Phi'_{m\times p} + \Psi_{p\times p}$$

$$\Sigma \approx \lambda_1 P_1 P_1' + \ldots + \lambda_m P_m P_m' = \Phi\Phi'$$

$$\Phi = \left(\phi_{jk}\right) = \left(\sqrt{\lambda_j}\, a_{jk}\right) \qquad \Psi = diag\left(\sigma_{jj} - h_j^2\right)$$

$$R_{res} = \Sigma - \left(\Phi\Phi' + \Psi\right)$$ Matriz residual

Os elementos da diagonal estão bem aproximados. Para os elementos fora da diagonal de Σ a aproximação pode não ser boa!!

Um critério de bondade de ajuste é:

S.Q. das entradas de $R_{res} \leq \lambda_{m+1}^2 + \lambda_{m+2}^2 + \ldots + \lambda_p^2$

Análise Fatorial via Componentes Principais

Exemplo: Considere a matriz de covariância de Y=(Y_1 Y_2 Y_3) dada por:

$$\Sigma = \begin{pmatrix} 1 & 0,9 & 0,7 \\ 0,9 & 1 & 0,4 \\ 0,7 & 0,4 & 1 \end{pmatrix}$$

Obter a solução do modelo fatorial para Y via Componentes Principais:

$$Y - \mu = \Phi \mathbf{f} + e$$

$$\Sigma \approx \Phi\Phi' + \Psi \qquad \Rightarrow \quad \Phi ? \quad \Psi ?$$

Análise Fatorial via Componentes Principais

$$\Sigma = \begin{pmatrix} 1 & 0,9 & 0,7 \\ 0,9 & 1 & 0,4 \\ 0,7 & 0,4 & 1 \end{pmatrix}$$

Autovalores: 2,35364 0,61602 0,03035

Matriz dos autovetores:

$$\begin{pmatrix} 0,643624 & -0,111080 & 0,757238 \\ 0,576635 & -0,580180 & -0,575225 \\ 0,503230 & 0,806878 & -0,309365 \end{pmatrix}$$

$$m = 1 \Rightarrow \Sigma \cong \lambda_1 a_1 a_1' \Rightarrow \quad \Phi = \sqrt{2,35} \begin{pmatrix} 0,6436 \\ 0,5766 \\ 0,5032 \end{pmatrix} = \begin{pmatrix} 0,987 \\ 0,885 \\ 0,772 \end{pmatrix}$$

$$Y - \mu = \Phi \mathbf{f} + e \quad \Rightarrow \quad Y_1 - \mu_1 = 0,987 F_1 + e_1$$

$$Y_2 - \mu_2 = 0,885 F_1 + e_2$$

$$Y_3 - \mu_3 = 0,772 F_1 + e_3$$

Análise Fatorial via Componentes Principais

$$Y - \mu = \Phi \mathbf{f} + e \quad \Rightarrow \quad \begin{aligned} Y_1 - \mu_1 &= 0{,}987 F_1 + e_1 \\ Y_2 - \mu_2 &= 0{,}885 F_1 + e_2 \\ Y_3 - \mu_3 &= 0{,}772 F_1 + e_3 \end{aligned}$$

$$\Sigma = \begin{pmatrix} 1 & 0{,}9 & 0{,}7 \\ 0{,}9 & 1 & 0{,}4 \\ 0{,}7 & 0{,}4 & 1 \end{pmatrix}$$

Matriz de covariância de Y

$tr\ \Sigma = 3$: variância total

$$\Sigma \approx \Phi\, \Phi' + \Psi$$

ψ_1=1-0,9742

$$\Sigma = \begin{pmatrix} 1 & 0{,}9 & 0{,}7 \\ 0{,}9 & 1 & 0{,}4 \\ 0{,}7 & 0{,}4 & 1 \end{pmatrix} \approx \begin{pmatrix} 0{,}9742 & 0{,}8735 & 0{,}7620 \\ 0{,}8735 & 0{,}7832 & 0{,}6832 \\ 0{,}7620 & 0{,}6832 & 0{,}5960 \end{pmatrix} + \begin{pmatrix} 0{,}0258 & 0 & 0 \\ 0 & 0{,}2168 & 0 \\ 0 & 0 & 0{,}4040 \end{pmatrix}$$

A variância total está preservada mas os termos fora da diagonal podem não estar bem aproximados!

Análise Fatorial via Máxima Verossimilhança

Suponha que os fatores comuns F e os específicos e seguem distribuição Normal, tal que, a distribuição marginal de Y é :

$$\mathbf{Y}_{ip\times 1} \overset{iid}{\sim} N_p\left(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Sigma}_{p\times p} = \Phi\Phi' + \Psi\right)$$

Para uma amostra de n vetores independentes de Y a função de verossimilhança é:

$$L(\mu,\Phi,\Psi \mid \mathbf{Y}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)'\Sigma^{-1}(Y_i-\mu)} = (2\pi)^{-np/2}|\Sigma|^{-n/2} e^{-\frac{1}{2}tr\left[\Sigma^{-1}\left(\sum_{i=1}^{n}(Y_i-\bar{Y})(Y_i-\bar{Y})'+n(\bar{Y}-\mu)(\bar{Y}-\mu)'\right)\right]}$$

com$\Sigma=\Phi\Phi$' + Ψ. Para $\hat{\mu} = \bar{Y}$ temos (exceto por constantes):

$$\ln L\left(\Phi,\Psi \mid S,\hat{\mu}\right) = -\frac{n}{2}\left(\ln|\Sigma| + tr\left(\Sigma^{-1}S\right)\right)$$

Everitt, 2004;
Johson and Whichern, 1992)

Maximizar *lnL* em Φ e Ψ é equivalente a minimizar a função (S e p são constantes na maximização):

$$D\left(\Phi,\Psi \mid S;\hat{\mu}\right) = \ln|\Sigma| - \ln|S| + tr\left(\Sigma^{-1}S\right) - p = tr\left(\Sigma^{-1}S\right) - \ln\left|\Sigma^{-1}S\right| - p$$

Considerando a falta de unicidade para rotações na matriz Φ, é adotada a restrição $\Phi'\Psi^{-1}\Phi = Diag_m$. É necessário usar métodos numéricos para obter os estimadores de Φ e Ψ que minimizem a função D.

Análise Fatorial via Máxima Verossimilhança

- A solução da análise fatorial via MVS é comumente obtida considerando a matriz de correlação de Y (equivalente a maximizar a função de verossimilhança dos dados padronizados Y*). Esse procedimento envolve menos parâmetros a serem estimados comparado a Σ . Assim, é necessário minimizar

$$D(\Phi_R, \Psi_R \,|R) = tr[(\Phi_R \Phi_R' + \Psi_R)^{-1} R] - \ln|(\Phi_R \Phi_R' + \Psi_R)^{-1} R\,| - p$$

com R a matriz de correlação amostral com divisor n e com a restrição de que $(\Phi_R' \Psi_R^{-1} \Phi_R)$ é matriz diagonal, obtendo-se

$$\hat{\Phi}_{R(p\times m)} \quad \text{e} \quad \hat{\Psi}_{R(p\times p)}; \quad R = \hat{\Phi}_R \hat{\Phi}_R' + \hat{\Psi}_R$$

Considerando que os estimadores de MVS são invariantes por escala, isto é, se $D(\Phi_R, \Psi_R \,|R)$ é minimizado em $\hat{\Phi}_R$ e $\hat{\Psi}_R$, $D(\Phi, \Psi \,|S)$ é minimizado em

$$\hat{\Phi}_{p\times m} = D_{s_{jj}}^{1/2} \hat{\Phi}_R; \quad \hat{\Psi}_{p\times p} = D_{s_{jj}}^{1/2} \hat{\Psi}_R D_{s_{jj}}^{1/2}; \quad \hat{\Sigma} = \hat{\Phi}\hat{\Phi}' + \hat{\Psi}$$

com s_{jj} a variância amostral de Y_j calculada com divisor “n”.

Análise Fatorial via Máxima Verossimilhança

É interessante avaliar situações em que o modelo fatorial oferece uma interpretação mais simplificada (em uma dimensão mais baixa) para as observações Y.

Como o vetor de parâmetros de locação μ não é de interesse na análise, usamos uma estimativa e podemos avaliar a aproximação da matriz de covariância amostral por meio do seguinte sistema de equações:

$$S_{p\times p} \cong \hat{\Phi}_{p\times m}\hat{\Phi}'_{m\times p} + \hat{\Psi}_{p\times p} \quad \Rightarrow \quad \mathrm{Re}\,s = S - \left(\hat{\Phi}\hat{\Phi}' + \hat{\Psi}\right)$$

com a restrição: $(\Phi'\Psi^{-1}\Phi)$ é matriz diagonal.

Além disso, a diferença no número de parâmetros envolvidos é:

$$\delta = \frac{p(p+1)}{2} - \left[pm + p - \left(\frac{m(m-1)}{2} \right) \right]$$

parâmetros em S; em Φ; em Ψ; sob a restrição de unicidade

δ<0: sistema com mais parâmetros do que equações e o modelo fatorial não está bem definido aos dados

δ=0: soluções exatas são possíveis mas o modelo fatorial não oferece simplificação

δ>0: a simplificação/redução é possível por meio do modelo fatorial

Análise Fatorial

$$R = \begin{pmatrix} 1{,}00 & 0{,}61 & 0{,}63 \\ 0{,}61 & 1{,}00 & 0{,}45 \\ 0{,}63 & 0{,}45 & 1{,}00 \end{pmatrix}$$

Solução do Modelo Fatorial para a extração de m=2 componentes comuns via Componentes Principais e Máxima Verossimilhança:

	CP				Max. Veross.			
Variável	F1	F2	Comunalid	Especif	F1	F2	Comunalid	Especif
Y1	-0,895	0,026	0,801	0,199	0,733	0,127	0,554	0,446
Y2	-0,81	-0,542	0,949	0,051	0,897	-0,38	0,949	0,051
Y3	-0,821	0,506	0,93	0,067	0,755	0,6	0,93	0,067
Variância	2,1303	0,5505	2,6808	0,032	1,9136	0,5199	2,4335	0,564
% Explic	0,71	0,183	0,894	0,0106	0,638	0,173	0,811	0,188

(Uso do R)

- As comunalidades obtidas via CP são usadas como valores iniciais no algoritmo iterativo de maximização da função de verossimilhança

Análise Fatorial

$$R = \begin{pmatrix} 1{,}00 & 0{,}61 & 0{,}63 \\ 0{,}61 & 1{,}00 & 0{,}45 \\ 0{,}63 & 0{,}45 & 1{,}00 \end{pmatrix}$$

Matriz Residual via Componentes Principais:

$$R - \left(\Phi_R \, \Phi_R' + \Psi_R\right) = \begin{pmatrix} 0{,}000000 & 0{,}100858 & 0{,}117951 \\ 0{,}100858 & 0{,}000000 & -0{,}059242 \\ 0{,}117951 & -0{,}059242 & 0{,}000000 \end{pmatrix}$$

Matriz Residual via Máxima Verossimilhança:

$$R - \left(\Phi_R \, \Phi_R' + \Psi_R\right) = \begin{pmatrix} 0{,}000000 & -0{,}0007590 & -0{,}000385 \\ -0{,}000759 & 0{,}0000000 & -0{,}000765 \\ -0{,}000385 & -0{,}0007650 & 0{,}000000 \end{pmatrix}$$

$\Rightarrow$ A solução por máxima verossimilhança apresenta melhor resultado na aproximação de Σ apesar da % da variância total explicada ter sido menor

Análise Fatorial – Rotação dos Fatores

$$Y_{p\times 1} \Rightarrow \qquad Y - \mu = \Phi\,\mathbf{f} + e$$

$$\begin{cases} Y_1 - \mu_1 = \phi_{11}F_1 + \phi_{12}F_2 + e_1 \\ Y_2 - \mu_2 = \phi_{21}F_1 + \phi_{22}F_2 + e_2 \\ \quad \ldots \\ Y_p - \mu_p = \phi_{p1}F_1 + \phi_{p2}F_2 + e_p \end{cases}$$

Φ é solução

$$\Rightarrow Cov(Y) = Cov(\Phi\mathbf{f} + e) = \Phi\Phi' + \Psi = \Sigma$$

Seja $\Phi^* = \Phi\Gamma; \qquad \Gamma\Gamma' = I$

$$\Rightarrow \quad \Phi^*\Phi^{*\prime} + \Psi = \Phi\Gamma\Gamma'\Phi' + \Psi = \Phi\Phi' + \Psi = \Sigma$$

$\Rightarrow \quad \Phi^*$ é solução

$\Phi^*=\Phi\Gamma \Rightarrow$ Geometricamente é uma rotação de eixos (novos fatores)

$\Rightarrow$ Podemos buscar rotações que conduzam a soluções fáceis de interpretar (médias, contrastes, formas canônicas)

Análise Fatorial – Rotação dos Fatores

	Geografia	Inglês	História	Aritmética	Algebra	Geometria
R =	1	0,439	0,41	0,288	0,329	0,248
	0,439	1	0,351	0,354	0,32	0,329
	0,41	0,351	1	0,164	0,19	0,181
	0,288	0,354	0,164	1	0,595	0,47
	0,329	0,32	0,19	0,595	1	0,464
	0,248	0,329	0,181	0,47	0,464	1

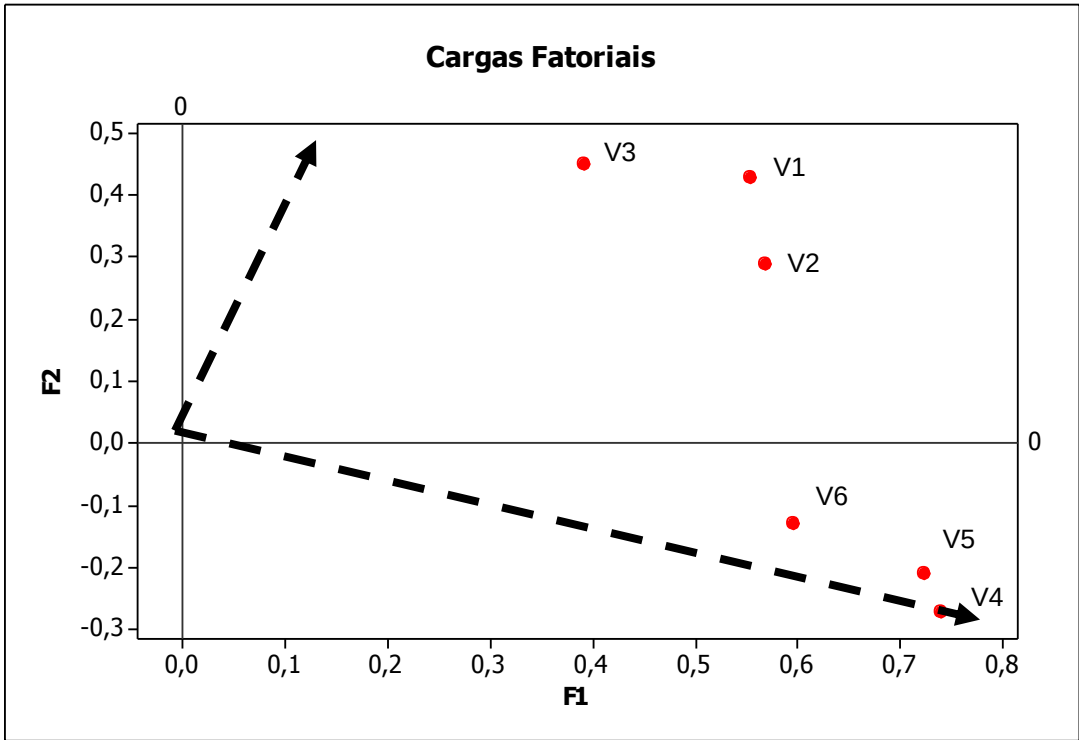
Variável	**Fator1**	**Fator2**	**Comunalidade**
Geogr	0,553	0,429	0,49
Inglês	0,568	0,288	0,406
Hist	0,392	0,45	0,356
Aritm	0,74	-0,273	0,623
Algebra	0,724	-0,211	0,569
Geom	0,595	-0,132	0,372
Variance	2,2094	0,6057	2,8151
% Var	0,368	0,101	0,469

- Solução da análise fatorial via máxima verossimilhança
- F1: cargas positivas ⇒resposta média ⇒ fator de inteligência geral
- F2: metade das cargas é positiva e metade é negativa ⇒fator bipolar

Rotacionar os fatores

Análise Fatorial – Rotação dos Fatores

Rotação de 20^0 eixos originais



⇒Nos novos eixos o fator de inteligência geral (F1) está particionado nos fatores F1* e F2*

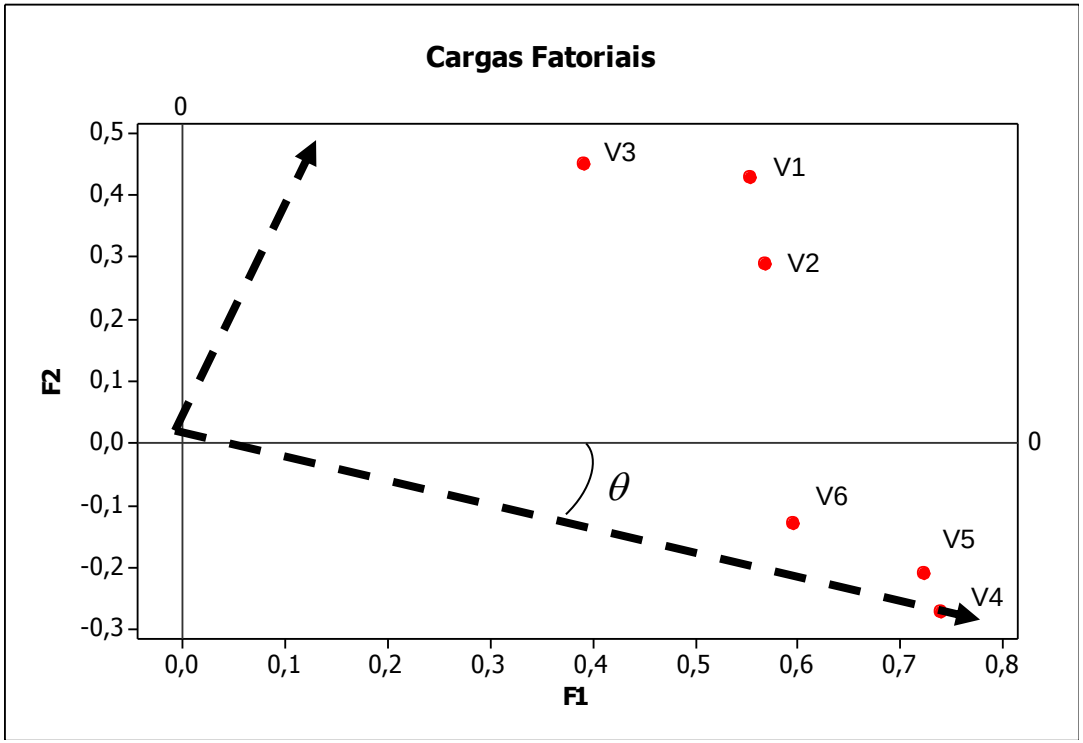
- Os eixos originais podem ser rotacionados de tal forma que todos os pontos caem no 1° quadrante
- As variáveis V4, V5 e V6 (da área exata) recebem carga alta em F1* e baixa em F2*
- As variáveis V1, V2 e V3 (da área de humanas) recebem carga alta em F2* e carga moderada/baixa em F1*

⇒F1*: habilidade matemática

⇒ F2*: habilidade verbal

Análise Fatorial – Rotação dos Fatores

Rotação de 20^0 eixos originais



Obtenha as novas coordenadas sob uma rotação de 20 graus nos eixos originais (sentido horário)
(Johnson and Wichern, 1978).

$$\Gamma = \begin{pmatrix} \cos\theta & sen\theta \\ -sen\theta & \cos\theta \end{pmatrix}$$

Cosseno de 20^0 = 0,9397
Seno de 20^0 = 0,3420

$$\Phi^* = \Phi\Gamma; \qquad \Gamma\Gamma' = I$$

Obtenha também as coordenadas sob uma rotação de 40 graus (Everitt, 2007).

Análise Fatorial – Rotação dos Fatores

	Fatores Originais			Fatores Rotacionados ($\theta=40^0$)		
Variável	**Fator1**	**Fator2**	**Comunalid**	**Fator1***	**Fator2***	**Comunalid**
Geogr	0,553	0,429	0,49	0,232	0,66	0,49
Inglês	0,568	0,288	0,406	0,321	0,551	0,406
Hist	0,392	0,45	0,356	0,085	0,591	0,356
Aritm	0,74	-0,273	0,623	0,77	0,173	0,623
Algebra	0,724	-0,211	0,569	0,723	0,215	0,569
Geom	0,595	-0,132	0,372	0,572	0,213	0,372
Variance	2,2094	0,6057	2,8151	1,6057	1,2094	2,8151
% Var	0,368	0,101	0,469	0,268	0,202	0,469

habilidade matemática

habilidade verbal

$\Rightarrow$ As comunalidades são invariantes por rotação ortogonal dos fatores

$$\Rightarrow \quad \Phi\Phi' = \Phi\Gamma\Gamma'\Phi = \Phi^*\Phi^{*\prime}$$

Análise Fatorial – Rotação dos Fatores

Não existe uma solução única para representar os fatores

Como escolher e obter uma solução/rotação ?

$\Rightarrow \quad \Phi^* = \left(\phi^*_{jk}\right)$: novas cargas dos fatores rotacionados

Conseguir valores 0's e 1's

Na prática o objetivo dos métodos de rotação é SIMPLIFICAR as linhas e colunas da matrix de cargas para facilitar a interpretação

$$Y - \mu = \Phi \mathbf{f} + e$$

$$\begin{cases} Y_1 - \mu_1 = \phi_{11}F_1 + \phi_{12}F_2 + \ldots + \phi_{1m}F_m + e_1 \\ Y_2 - \mu_2 = \phi_{21}F_1 + \phi_{22}F_2 + \ldots + \phi_{2m}F_m + e_2 \\ \quad \ldots \\ Y_p - \mu_p = \phi_{p1}F_1 + \phi_{p2}F_2 + \ldots + \phi_{pm}F_m + e_p \end{cases}$$

Matriz de Cargas

$$\begin{pmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1m} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2m} \\ & & \cdots & \\ \phi_{p1} & \phi_{p1} & \cdots & \phi_{pm} \end{pmatrix}$$

linhas

$\Rightarrow$ Variabilid. das variáveis Yj

colunas $\Rightarrow$ variabilidade dos fatores Fk

Análise Fatorial – Rotação dos Fatores

Não existe uma solução única para representar os fatores

Como escolher e obter uma solução/rotação ?

Métodos de Rotação Ortogonal:

- Rotação Varimax: simplifica as colunas da matriz de cargas
- Rotação Quartimax: simplifica as linhas da matriz de cargas
- Rotação Equimax: é um compromisso entre as duas outras técnicas

Existem ainda as rotações obliquas. Neste caso, as comunalidades são não invariantes.

Análise Fatorial – Rotação dos Fatores

	Fatores Originais			Fatores Rotacionados		
Variável	**Fator1**	**Fator2**	**Comunalid**	**Fator1***	**Fator2***	**Comunalid**
Geogr	0,553	0,429	0,49	0,232	0,66	0,49
Inglês	0,568	0,288	0,406	0,321	0,551	0,406
Hist	0,392	0,45	0,356	0,085	0,591	0,356
Aritm	0,74	-0,273	0,623	0,77	0,173	0,623
Algebra	0,724	-0,211	0,569	0,723	0,215	0,569
Geom	0,595	-0,132	0,372	0,572	0,213	0,372
Variance	2,2094	0,6057	2,8151	1,6057	1,2094	2,8151
% Var	0,368	0,101	0,469	0,268	0,202	0,469

	Fatores Rotacionados - Quartimax			Fatores Rotacionados - Equimax		
Variável	**Fator1**	**Fator2**	**Comunalid**	**Fator1***	**Fator2***	**Comunalid**
Geogr	0,26	0,65	0,49	0,232	0,66	0,49
Inglês	0,344	0,536	0,406	0,321	0,551	0,406
Hist	0,111	0,587	0,356	0,085	0,591	0,356
Aritm	0,777	0,139	0,623	0,77	0,173	0,623
Algebra	0,731	0,184	0,569	0,723	0,215	0,569
Geom	0,58	0,188	0,372	0,572	0,213	0,372
Variance	1,6733	1,1418	2,8151	1,6057	1,2094	2,8151
% Var	0,279	0,19	0,469	0,268	0,202	0,469

Análise Fatorial – Escores Fatoriais

Escore Fatorial: valor que cada indivíduo na amostra tem para cada um dos fatores comuns

$$Y_{p\times 1} \Rightarrow \qquad Y - \mu = \Phi\,\mathbf{f} + e$$

$$\begin{cases} Y_1 - \mu_1 = \phi_{11}F_1 + \phi_{12}F_2 + e_1 \\ Y_2 - \mu_2 = \phi_{21}F_1 + \phi_{22}F_2 + e_2 \\ \quad \ldots \\ Y_p - \mu_p = \phi_{p1}F_1 + \phi_{p2}F_2 + e_p \end{cases}$$

Para o indivíduo i :

$$Y_{i\,(p\times 1)} \Rightarrow \qquad Y_i - \mu = \Phi\,\mathbf{f}_i + e_i$$

$$\begin{cases} Y_{1i} - \mu_1 = \phi_{11}F_{1i} + \phi_{12}F_{2i} + e_{1i} \\ Y_{2i} - \mu_2 = \phi_{21}F_{1i} + \phi_{22}F_{2i} + e_{2i} \\ \quad \ldots \\ Y_{pi} - \mu_p = \phi_{p1}F_{1i} + \phi_{p2}F_{2i} + e_{pi} \end{cases}$$

Qual o valor de $\mathbf{f}_i$, ? $\qquad i = 1,2,\ldots,n$

Análise Fatorial – Escores Fatoriais

Qual o valor de $\mathbf{f}_i$, ? $\quad i = 1,2,...,n$

- Método de Componentes Principais:

m primeiros componentes principais padronizados

$$Y_{i\,(p\times 1)} \Rightarrow \quad Y_i - \mu = \Phi\,\mathbf{f}_i + e_i \quad \Rightarrow \quad \mathbf{f}_i = Z_i D_{\lambda_j}^{-1/2}$$

- Método de Mínimos Quadrados Ponderados (Bartlett):

Supondo μ, Φ e Ψ conhecidos $\Rightarrow$ modelo factorial pode ser formulado como um modelo de regressão linear heterocedástico nas variáveis preditoras Φ.

O estimador (preditor) de $\mathbf{f_i}$ é dado por:

Restrição na EMVS

$$\hat{\mathbf{f}}_i = \underbrace{\left(\Phi'\Psi^{-1}\Phi\right)^{-1}\Phi'\Psi^{-1}}_{(mxp)}\left(Y_i - \mu\right) \qquad (px1)$$

Coeficiente do fator

Análise Fatorial – Escores Fatoriais

$Y_{i\,(p\times 1)} \Rightarrow \qquad Y_i - \mu = \Phi\,\mathbf{f}_i + e_i$ Qual o valor de $\mathbf{f}_i$, ? $\qquad i = 1,2,...,n$

- Método da Regressão: (μ , Φ e ψ são assumidos conhecidos)

$$\mathbf{f}_i \sim N_m(0, I_m) \quad e_i \sim N_p(0, \Psi) \quad \Rightarrow \quad \begin{pmatrix} e_i \\ \mathbf{f}_i \end{pmatrix} \sim N_{p+m}\left(0, \begin{pmatrix} \Psi & 0 \\ 0 & I_m \end{pmatrix}\right)$$

$$Y_i - \mu = \Phi\,\mathbf{f}_i + e_i \sim N_p(0, \Sigma = \Phi\Phi' + \Psi) \quad \Rightarrow \quad \begin{pmatrix} Y_i - \mu \\ \mathbf{f}_i \end{pmatrix} \sim N_{p+m}\left(0, \begin{pmatrix} \Sigma & \Phi \\ \Phi' & I_m \end{pmatrix}\right)$$

$$\mathbf{f}_i / Y_i \sim N_m\left(\Phi'\Sigma^{-1}(Y_i - \mu); I_m - \Phi'\Sigma^{-1}\Phi\right)$$

Esperança condicional dos fatores

O preditor de $\mathbf{f_i}$ é dado por: $\hat{\mathbf{f}}_i = \Phi'\Sigma^{-1}(Y_i - \mu) = \underbrace{\Phi'(\Phi\Phi' + \Psi)^{-1}}_{\text{(mxp)}}(Y_i - \mu)$

Coeficiente do fator

Análise Fatorial

Arquivo HATCO (Hair et al., 2005)

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45	4,4	0	1	1	2
…														
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1	0	60	5,2	1	0	1	3
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4	1	38	3,7	0	1	0	1
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4	1	42	4,3	0	1	0	1
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0	0	33	4,4	1	0	0	1

Análise Fatorial

Arquivo HATCO (Hair et al., 2005)

	Fatores Originais - CP			Fatores Rotacionados - Varimax		
Variável	Fator1	Fator2	Comunalidade	Fator1	Fator2	Comunalidade
X1	0,627	-0,514	0,658	-0,787	0,194	0,658
X2	-0,759	0,068	0,58	0,714	0,266	0,58
X3	0,73	-0,337	0,646	-0,804	-0,011	0,646
X4	-0,494	-0,798	0,882	0,102	0,933	0,882
X6	-0,425	-0,832	0,872	0,025	0,934	0,872
X7	-0,767	0,168	0,616	0,764	0,179	0,616
Variance	2,5135	1,7395	4,253	2,3698	1,8832	4,253
% Var	0,419	0,29	0,709	0,395	0,314	0,709

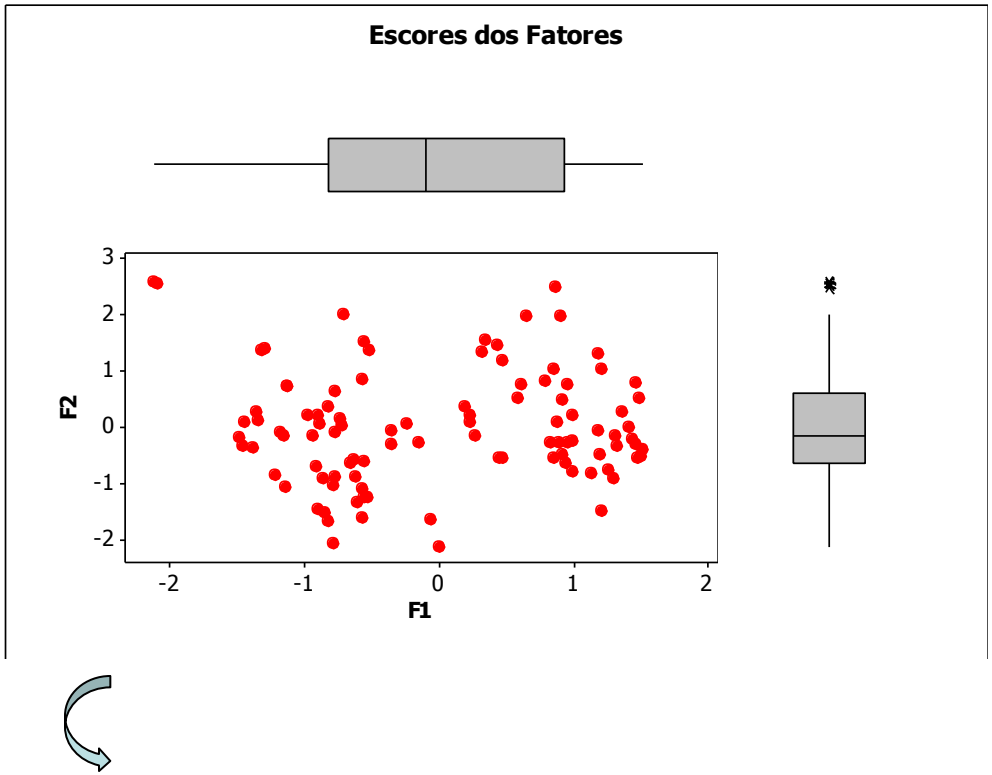
Valor básico Imagem

- X1, X2 e X3: variáveis associadas a preço do produto (X7: qualidade)
- X4 e X6: variáveis associadas à imagem da CO

$\Rightarrow$ Como validar os resultados de uma Análise Fatorial?

Análise Fatorial

Arquivo HATCO (Hair et al., 2005)



Coeficientes dos Fatores (Φ)

Variável	Fator1	Fator2
X1	-0,352	0,159
X2	0,289	0,095
X3	-0,345	0,05
X4	-0,02	0,499
X6	-0,053	0,504
X7	0,317	0,044

$$\hat{\mathbf{f}}_i = \underbrace{\Phi'\left(\Phi\Phi' + \Psi\right)^{-1}}\left(Y_i - \bar{Y}\right)$$

Matriz (mxp) dos coeficientes dos fatores

Gráfico de dispersão das observações (empresas). Os escores fatoriais podem ser usados para Análise de diagnóstico de observações discrepantes (o que também pode ser feito com os escores dos Componentes Principais, bem como com base na distância de Mahalanobis).

Componentes Principais x Análise Fatorial

- Ambas buscam uma Redução de Dimensionalidade, por explicar um conjunto de dados multivariados usando um conjunto menor de variáveis

- CONTUDO, os critérios de otimalidade usados em cada caso são diferentes:

$\Rightarrow$ Análise Fatorial é ótima no sentido de explicar as covariâncias (e correlações) entre as variáveis por meio de fatores comuns.

$\Rightarrow$ Análise de CP explica a variância total das variáveis observadas.

$\Rightarrow$ Na análise de CP se o número de componentes retidos aumenta, isto NÃO altera os anteriores, mas isto pode não acontecer na Análise Fatorial sob a solução de MVS.

$\Rightarrow$ Cálculo dos escores das observações nos CPs tem solução única. No caso de Análise fatorial (via MVS) existem diferentes procedimentos inferenciais propostos.

Componentes Principais x Análise Fatorial

⇒ As análises de CPs via matriz de covariância (Σ=Cov(Y)) ou de correlação (R=Cov(Y*)) são diferentes não sendo possível relacioná-las. Na Análise Fatorial via MVS a solução para a matriz de covariância (Φ) é obtida da solução da matriz de correlação (Φ*) :

$$\Phi = D_{s_{jj}}^{1/2}\ \Phi^*, \qquad \Psi = D_{s_{jj}}^{1/2}\ \Psi^*\ D_{s_{jj}}^{1/2}$$

⇒ Teste (assintótico) da adequação do modelo fatorial:

$$H_0 : \Sigma = \Phi\Phi' + \Psi \qquad H_1 : \Sigma \text{ com estrutura geral}$$

A estatística da razão de verossimilhanças (sob normalidade) é:

$$-2\ln\frac{L_0}{L_1} = n\ln\left(\frac{\left|\hat{\Phi}\hat{\Phi}' + \hat{\Psi}\right|}{|S_n|}\right)$$

Usando a correção de Bartlett, rejeita-se H0 a um nível de significância α se:

$$\left(n - 1 - (2p + 4m + 5)/6\right)\ln\left(\frac{\left|\hat{\Phi}\hat{\Phi}' + \hat{\Psi}\right|}{|S_n|}\right) > \chi^2_{\left[(p-m)^2 - p - m\right]/2}(\alpha)$$

Análise Discriminante

Redução de Dimensionalidade
Análise Discriminante

Análise de CP (CoP, ACo)
Análise Fatorial

$$Y_{n\times p};\quad n>p \Rightarrow\ Y_{i_{p\times 1}} \overset{iid}{\sim}(\mu;\Sigma);\quad \Re^p \rightarrow \Re^m, m<\min(n,p)$$

Semelhança com MANOVA

Análise Discriminante: $Y_{n\times p};\quad n=\sum_{g=1}^{G} n_g;\quad Y_{gi_{p\times 1}} \overset{iid}{\sim}\left(\mu_g;\Sigma_g\right), g=1,2,...,G$

$$\Re^p \rightarrow \Re^m$$

- Populações Estratificadas: Caso em que G=2
  - Solução de Fisher
  - Solução Probabilística (Regra Discriminante de Bayes)

- Populações Estratificadas: Caso em que G>2
  - Solução de Fisher
  - Solução Probabilística

População Estratificada

τ_1 τ_2 ... τ_G

$Y_i \in \Re^p$

$E(Y_i \mid \tau_1) = \mu_{1(p\times 1)}$ $E(Y_i \mid \tau_2) = \mu_{2(p\times 1)}$ ... $E(Y_i \mid \tau_G) = \mu_{G(p\times 1)}$

$Cov(Y_i \mid \tau_1) = \Sigma_{1(p\times p)}$ $Cov(Y_i \mid \tau_2) = \Sigma_{2(p\times p)}$ ... $Cov(Y_i \mid \tau_G) = \Sigma_{G(p\times p)}$

n_1 n_2 n_G

Amostra

Grupos	Unidades amostrais	Variáveis					
		1	2	…	j	…	p
1	1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
	2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
	…	…	…	…	…		…
…	i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
	…	…	…	…	…	…	…
G	n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

$$Y_{n\times p} = \begin{pmatrix} Y_1 \\ \dots \\ Y_G \end{pmatrix}$$

$$Y_{g(n_g \times p)}$$

$$n = \sum_{g=1}^{G} n_g$$

Objetivos: ANÁLISE DISCRIMINANTE

Problema de P-Integração ?

- Obter Funções Discriminantes das “p” variáveis
  - Redução de dimensionalidade: $\Re^p \rightarrow \Re^m$; $m \leq \min(n,p,(G-1))$
  - Classificar “novas” observações

Análise Discriminante - Motivação

Banco	Condição	Y1	Y2	Y3	Y4
B1	1	0,8888	0,7391	1,0255	0,3938
B2	1	1,6655	0,7268	0,878	0,0004
B3	1	2,2111	0,9166	0,9492	0,342
B4	1	1,4351	0,9133	0,9577	0,2325
B5	1	2,1414	0,002	1,0245	0,3966
B6	1	1,192	0,4972	1,034	0,3095
B7	1	1,5895	0,2593	1,0453	0,557
B8	1	1,3272	0,4126	1,0448	0,3482
B9	1	1,8847	0,388	0,9864	0,0337
B10	1	0,5229	0,9473	1,1244	0,118
n		10	10	10	10
Média		1,4852	0,5802	1,007	0,2732
D.P.		0,533	0,319	0,0674	0,1762
B11	2	0,4922	0,3166	1,1127	0,1628
B12	2	1,4427	0,0589	0,9019	0,1355
B13	2	0,5438	0,5358	1,03	0,1481
B14	2	0,1904	0,7087	0,9917	0,2625
B15	2	0,1102	0,7378	1,528	0,0783
B16	2	2,006	0,014	1,0321	0,0816
B17	2	0,2321	0,9234	0,9753	0,0045
B18	2	0,9019	0,1634	1,1414	0,5485
B19	2	1,9757	0,3395	0,9997	0,0751
B20	2	0,7276	0,3139	1,1077	0,2957
n		10	10	10	10
Média		0,862	0,4112	1,0821	0,1793
D.P.		0,712	0,3055	0,1726	0,1567

Dados: Barroso e Artes, 2002

Condição:

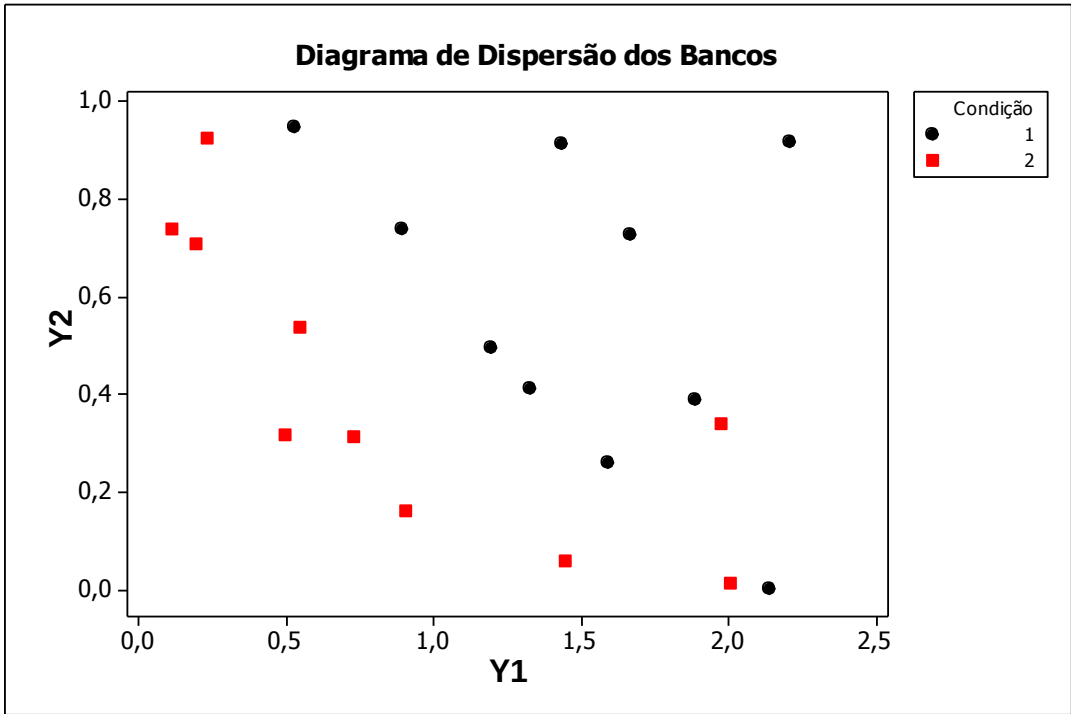
1: Com problemas

2: Sem problemas

Objetivo:

Com base nas variáveis econômicas Y1 e Y2, obter uma função discriminante que permita a classificação de um banco em Com ou Sem problemas.

Análise Discriminante



Considerando as variáveis Y1 e Y2, qual seria uma direção “otima” (função linear de Y1 e Y2) para a discriminação das instituições bancárias?

Análise Discriminante

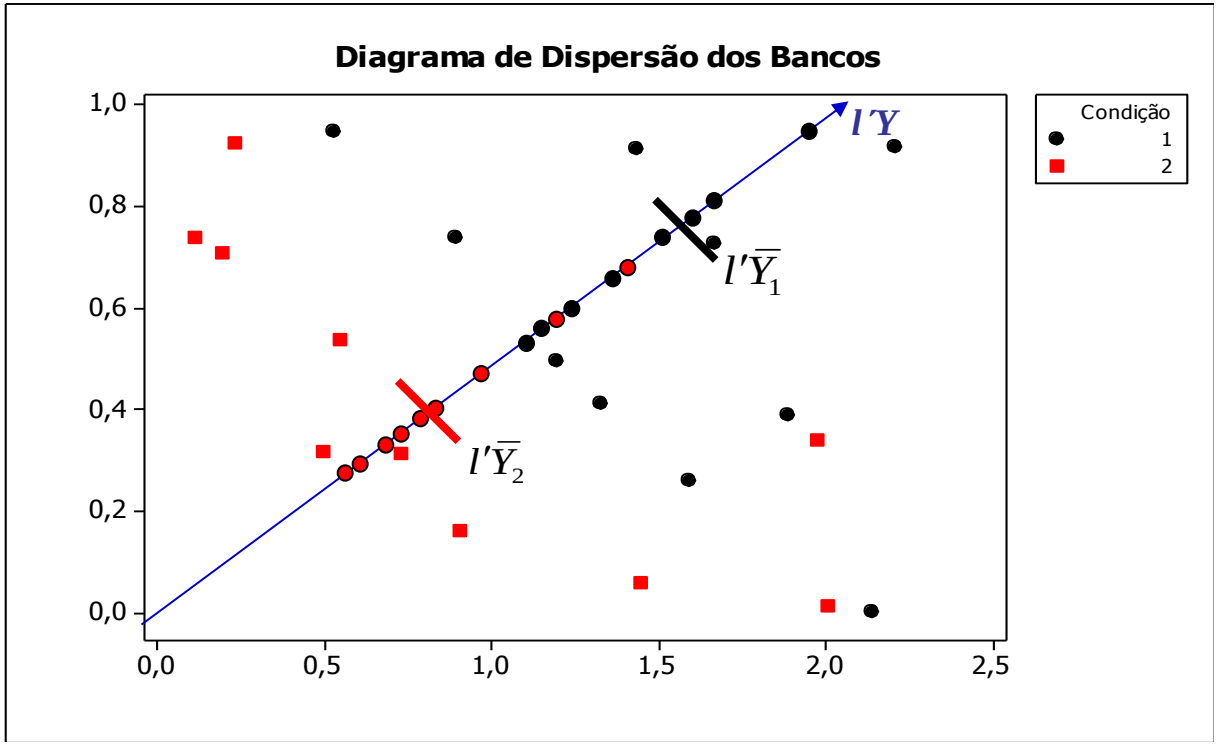


Gráfico de dispersão das observações. Indicação de um terceiro eixo $l'Y$ (em azul) que define uma função discriminante linear.

Função Discriminante Linear de Fisher

Formulação de Fisher – Caso de 2 Populações

O grupo ao qual a observação pertence é conhecido.

Considere uma População constituída por observações multivariadas (quantitativas) e estratificada em dois subgrupos, tal que:

$$Y_{n\times p} = \begin{bmatrix} Y_{1(n_1\times p)} \\ Y_{2(n_2\times p)} \end{bmatrix} \Rightarrow \begin{cases} Y_i \in \Re^p; & E(Y_i \mid \tau_1) = \mu_{1(p\times 1)} \quad Cov(Y_i \mid \tau_1) = \Sigma_{1(p\times p)} \\ & E(Y_i \mid \tau_2) = \mu_{2(p\times 1)} \quad Cov(Y_i \mid \tau_2) = \Sigma_{2(p\times p)} \end{cases}$$

Suposição $\Rightarrow \quad \Sigma_1 = \Sigma_2 = \Sigma$

Matrizes de covariâncias homogêneas

Para G=2 - Proposta de Fisher: obter combinações lineares de Yi$\in \Re^p$ tais que:

$$Y_i \in \Re^p \rightarrow X = l' Y_i; \qquad l = \arg\max_{l;=l'Y} \frac{(\mu_{X1} - \mu_{X2})^2}{\sigma_X^2} = \arg\max_l \frac{(l'\mu_1 - l'\mu_2)^2}{l'\Sigma l}$$

Função Discriminante Linear de Fisher

$$Y_{n\times p}=\begin{bmatrix} Y_{1(n_1\times p)} \\ Y_{2(n_2\times p)} \end{bmatrix} \Rightarrow \begin{cases} Y_i \in \Re^p; \quad E(Y_i \mid \tau_1)=\mu_{1(p\times 1)} \\ E(Y_i \mid \tau_2)=\mu_{2(p\times 1)} \end{cases} \qquad \Sigma_1=\Sigma_2=\Sigma_{p\times p}$$

Matrizes de covariâncias homogêneas

Seja X uma combinação linear das variáveis multidimensionais $Yi_{(px1)}$. Então,

$$\Rightarrow \quad X_i = l'\,Y_i \quad \begin{cases} \mu_{X_1}=E(X_i \mid \tau_1)=E(l'Y_i \mid \tau_1)=l'\mu_1 \\ \mu_{X_2}=E(X_i \mid \tau_2)=E(l'Y_i \mid \tau_2)=l'\mu_2 \end{cases} \qquad \sigma_X^2=Var(l'Y_i)=l'\Sigma l$$

Desigualdade de Cauchy-Schwarz

$$\max_{l;X=l'Y} \frac{(\mu_{X_1}-\mu_{X_2})^2}{\sigma_X^2}; \quad \frac{l'(\mu_1-\mu_2)(\mu_1-\mu_2)'l}{l'\Sigma l}=\frac{l'\delta\,\delta'l}{l'\Sigma l}=\frac{(l'\delta)^2}{l'\Sigma l} \le (\delta'\Sigma^{-1}\delta)$$

$$(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)$$

Distância de Mahalanobis

$$l=\Sigma^{-1}(\mu_1-\mu_2); \quad X_i=l'\,Y_i=(\mu_1-\mu_2)'\Sigma^{-1}\,Y_i$$

Discriminação sob Estimação

$$Y_{n\times p}=\begin{bmatrix} Y_{1(n_1\times p)} \\ Y_{2(n_2\times p)} \end{bmatrix} \Rightarrow \qquad Y_i \in \Re^p \rightarrow X_i = l'\,Y_i = \left(\mu_1-\mu_2\right)'\Sigma^{-1}Y_i$$

Para dados amostrais: Adotar estimadores "apropriados" de $\hat{\mu}_1,\ \hat{\mu}_2,\ \hat{\Sigma}$

$$X_i = l'Y_i = \left(\bar{Y}_1-\bar{Y}_2\right)'S_c^{-1}Y_i$$

$$S_{c_{p\times p}} = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}; \quad S_g = \left(n_g-1\right)^{-1}\sum_{g=1}^{2}\sum_{i=1}^{n_g}\left(Y_{gi}-\bar{Y}_g\right)\left(Y_{gi}-\bar{Y}_g\right)'$$

Regra de Classificação Amostral: Alocação de uma nova observação

$$Y_0\,? \qquad \begin{cases} X_0 \geq c \Rightarrow Y_0 \in \tau_1 \\ X_0 < c \Rightarrow Y_0 \in \tau_2 \end{cases}$$

$$c = \bar{X} = \frac{1}{2}\left(\bar{X}_1+\bar{X}_2\right) = \frac{1}{2}\left(l'\bar{Y}_1+l'\bar{Y}_2\right) = \frac{1}{2}l'\left(\bar{Y}_1+\bar{Y}_2\right) = \frac{1}{2}\left(\bar{Y}_1-\bar{Y}_2\right)'S_c^{-1}\left(\bar{Y}_1+\bar{Y}_2\right)$$

Análise Discriminante - Exemplo

Banco	Condição	Y1	Y2	Y3	Y4
B1	1	0,8888	0,7391	1,0255	0,3938
B2	1	1,6655	0,7268	0,878	0,0004
B3	1	2,2111	0,9166	0,9492	0,342
B4	1	1,4351	0,9133	0,9577	0,2325
B5	1	2,1414	0,002	1,0245	0,3966
B6	1	1,192	0,4972	1,034	0,3095
B7	1	1,5895	0,2593	1,0453	0,557
B8	1	1,3272	0,4126	1,0448	0,3482
B9	1	1,8847	0,388	0,9864	0,0337
B10	1	0,5229	0,9473	1,1244	0,118
n		10	10	10	10
Média		1,4852	0,5802	1,007	0,2732
D.P.		0,533	0,319	0,0674	0,1762
B11	2	0,4922	0,3166	1,1127	0,1628
B12	2	1,4427	0,0589	0,9019	0,1355
B13	2	0,5438	0,5358	1,03	0,1481
B14	2	0,1904	0,7087	0,9917	0,2625
B15	2	0,1102	0,7378	1,528	0,0783
B16	2	2,006	0,014	1,0321	0,0816
B17	2	0,2321	0,9234	0,9753	0,0045
B18	2	0,9019	0,1634	1,1414	0,5485
B19	2	1,9757	0,3395	0,9997	0,0751
B20	2	0,7276	0,3139	1,1077	0,2957
n		10	10	10	10
Média		0,862	0,4112	1,0821	0,1793
D.P.		0,712	0,3055	0,1726	0,1567

$$\bar{Y}_1 = \begin{pmatrix} 1{,}485 \\ 0{,}58 \end{pmatrix} \qquad \bar{Y}_2 = \begin{pmatrix} 0{,}862 \\ 0{,}414 \end{pmatrix}$$

$$\bar{Y}_1 - \bar{Y}_2 = \begin{pmatrix} 0{,}624 \\ 0{,}166 \end{pmatrix}$$

$$S_1 = \begin{pmatrix} 0{,}284 & -0{,}07 \\ -0{,}07 & 0{,}101 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 0{,}506 & -0{,}164 \\ -0{,}164 & 0{,}091 \end{pmatrix}$$

$$S_c = \begin{pmatrix} 0{,}395 & -0{,}117 \\ -0{,}164 & 0{,}096 \end{pmatrix}$$

$$l' = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}$$

$$X = l'Y = 3{,}274Y_1 + 5{,}719Y_2$$

Análise Discriminante - Exemplo

Banco	Condição	Função Discr	Classificação
B1	1	7,13	1
B2	1	9,61	1
B3	1	12,48	1
B4	1	9,92	1
B5	1	7,02	1
B6	1	6,74	1
B7	1	6,68	2
B8	1	6,7	1
B9	1	8,39	1
B10	1	7,13	1
B11	2	3,42	2
B12	2	5,06	2
B13	2	4,84	2
B14	2	4,67	2
B15	2	4,58	2
B16	2	6,8	1
B17	2	6,04	2
B18	2	3,89	2
B19	2	8,41	1
B20	2	4,17	2

Qual é a proporção de acerto de X?

$$\overline{X} = 6{,}681$$

$$X_0 \geq \overline{X} \Rightarrow \text{Condição 1}$$

$$X_0 < \overline{X} \Rightarrow \text{Condição 2}$$

Matriz de classificação

	Predito		
Verdade	τ_1	τ_2	
τ_1	9	1	10
τ_2	2	8	10

17/20=0,85

Taxa de acerto superestimada: a amostra foi usada no treinamento e na validação do procedimento

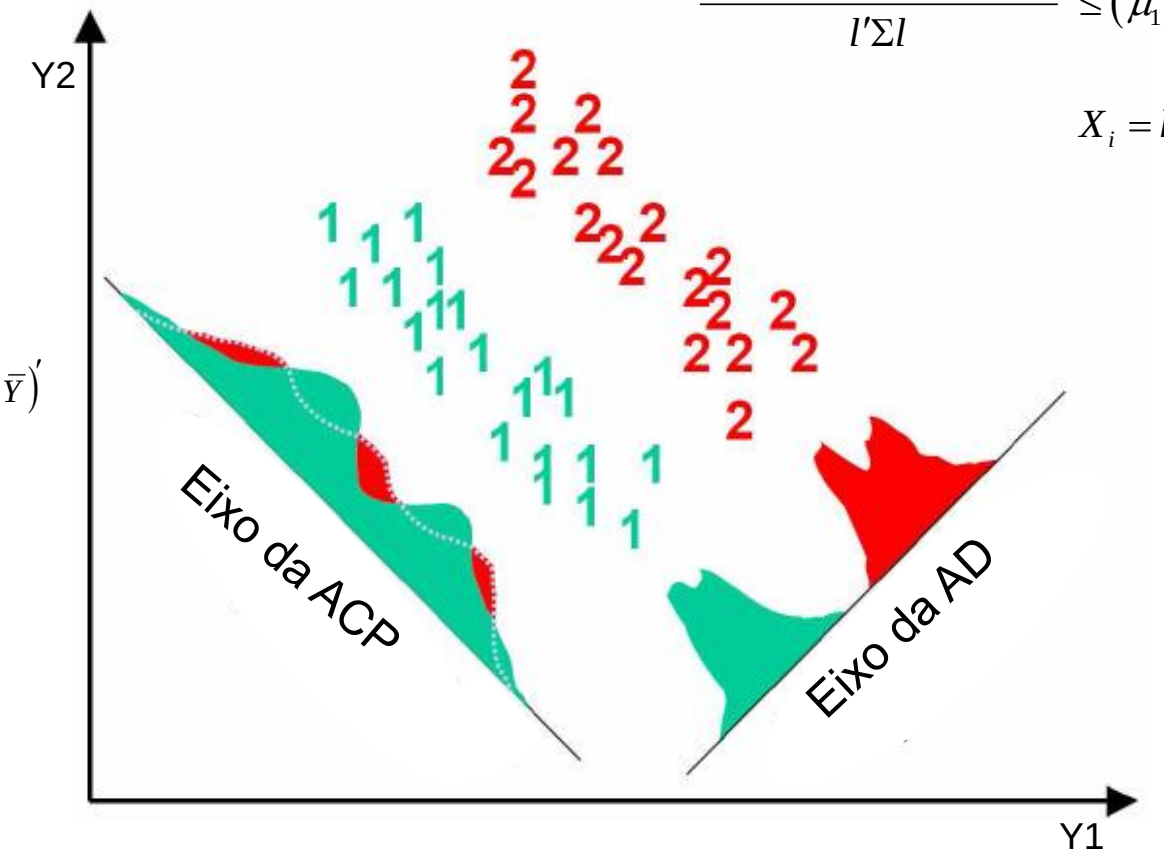
Alternativas: Dividir a amostra (Treinamento e Validação), Método de validação cruzada

Análise de Componentes Principais
G=2

$$\max_{\|a\|=1} \frac{a'\Sigma a}{a'a}$$

$$\Sigma = P\Lambda P';$$
$$a_j = P_j;$$
$$\Sigma\ P_j = \lambda_j P_j$$

$$\hat{\Sigma} = \frac{1}{n-1} S_T$$
$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(Y_i - \bar{Y}\right)'$$



Análise Discriminante
G=2

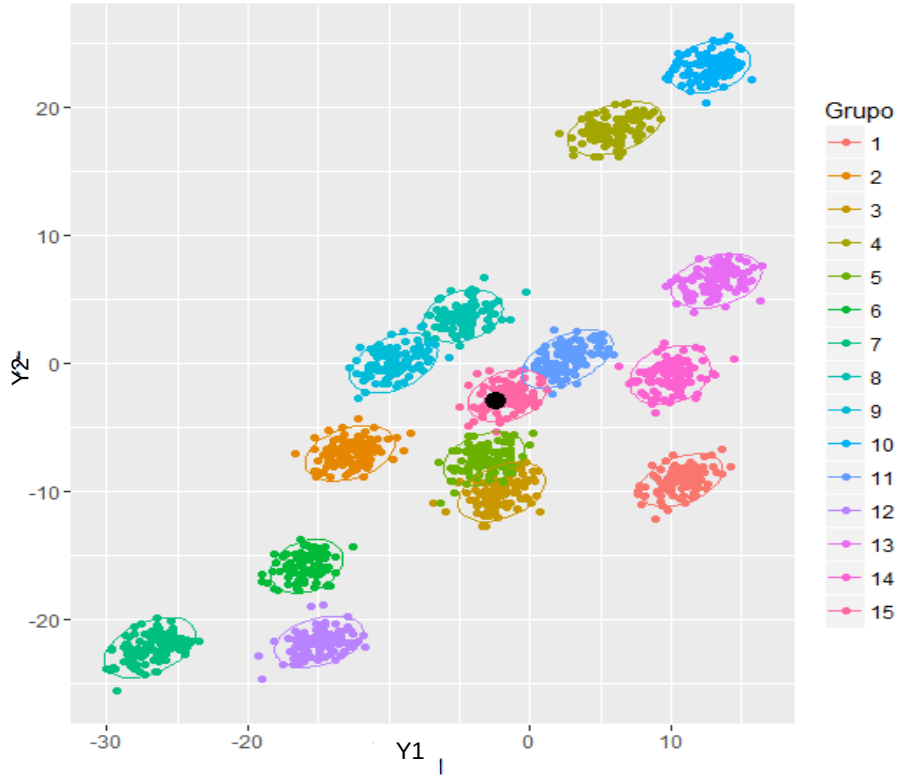
$$\max_{l;X=l'Y} \frac{\left(\mu_{X_1} - \mu_{X_2}\right)^2}{\sigma_X^2};$$

$$\frac{l'\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)' l}{l'\Sigma l} \leq \left(\mu_1 - \mu_2\right)' \Sigma^{-1} \left(\mu_1 - \mu_2\right)$$

$$X_i = l'Y_i = \left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_i$$

Funções Discriminantes de Fisher

Populações Estratificadas em Muitos Grupos (G>2)



Como realizar a redução dos dados (p=2, G=15)? Uma única dimensão ($X=l'Y$) é suficiente para uma boa discriminação dos grupos?

Funções Discriminantes de Fisher

Solução de Fisher para Muitas Populações

$$Y_{n\times p}=\begin{bmatrix} Y_{1(n_1\times p)} \\ \dots \\ Y_{G(n_G\times p)} \end{bmatrix} \Rightarrow \begin{cases} Y_i \in \Re^p; \quad E(Y_i \mid \tau_1)=\mu_{1(p\times 1)} \quad \dots \quad E(Y_i \mid \tau_G)=\mu_{G(p\times 1)} \\ Cov(Y_i \mid \tau_g)=\Sigma_g=\Sigma_{(p\times p)}, \quad g=1,2,\dots,G \end{cases}$$

$$Y_i \in \Re^p \rightarrow X_i = l' Y_i;$$

B: Matriz de covariância ENTRE

$$l=\arg\max_{l;X=l'Y} \frac{\sum_{g=1}^{G}\left(\mu_{X_g}-\bar{\mu}_X\right)^2}{\sigma_X^2} \Rightarrow \frac{\sum_{g=1}^{G}\left(l'\mu_g-l'\bar{\mu}\right)^2}{\sigma_X^2}=\frac{l'\sum_{g=1}^{G}\left(\mu_g-\bar{\mu}\right)\left(\mu_g-\bar{\mu}\right)'l}{l'\Sigma l}=\frac{l'\,B\,l}{l'\Sigma l}$$

Σ: Matriz de covariância DENTRO

As funções discriminantes, $L_{p\times m}=(l_1,\dots,l_m)$, são obtidas a partir dos autovetores da matriz $\Sigma^{-1}B$, restritos a $L'\Sigma L=I$.

$$\Sigma^{-1/2}\,B\,\Sigma^{-1/2}=P\Lambda P'; \quad L=\Sigma^{-1/2}P; \quad m\le \min(n,p,G-1)$$

Matriz simétrica

Funções Discriminantes de Fisher

Método de Fisher para Muitas Populações

- Dados Amostrais: maximizar a função em termos de estimadores apropriados

$$\frac{l' B\, l}{l'\Sigma l} \Rightarrow \hat{l} = \arg\max_l \frac{l'\hat{B}\, l}{l'\hat{\Sigma} l} \qquad \Rightarrow \hat{L}_{p\times m} = \left(\hat{l}_1,\ldots,\hat{l}_m\right)$$

Matriz de "SQPC Entre grupos" da MANOVA:

$$\hat{B}_{p\times p} = \sum_{g=1}^{G} n_g \left(\bar{Y}_g - \bar{Y}\right)\left(\bar{Y}_g - \bar{Y}\right)'$$

Matriz de "QMPC Dentro de grupos" da MANOVA:

$$\hat{\Sigma} = S_{c_{p\times p}} = \frac{(n_1-1)S_1 + \ldots + (n_G-1)S_G}{n_1+\ldots+n_G-G} = (n-G)^{-1}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\left(Y_{gi}-\bar{Y}_g\right)\left(Y_{gi}-\bar{Y}_g\right)'$$

- Regra de Classificação Amostral:

Alocar a observação Y_0 ($\in \Re^p$) à população τ_k em que a o valor da função discriminante X_0 ($\in \Re^m$) está mais "próxima" de seu centróide:

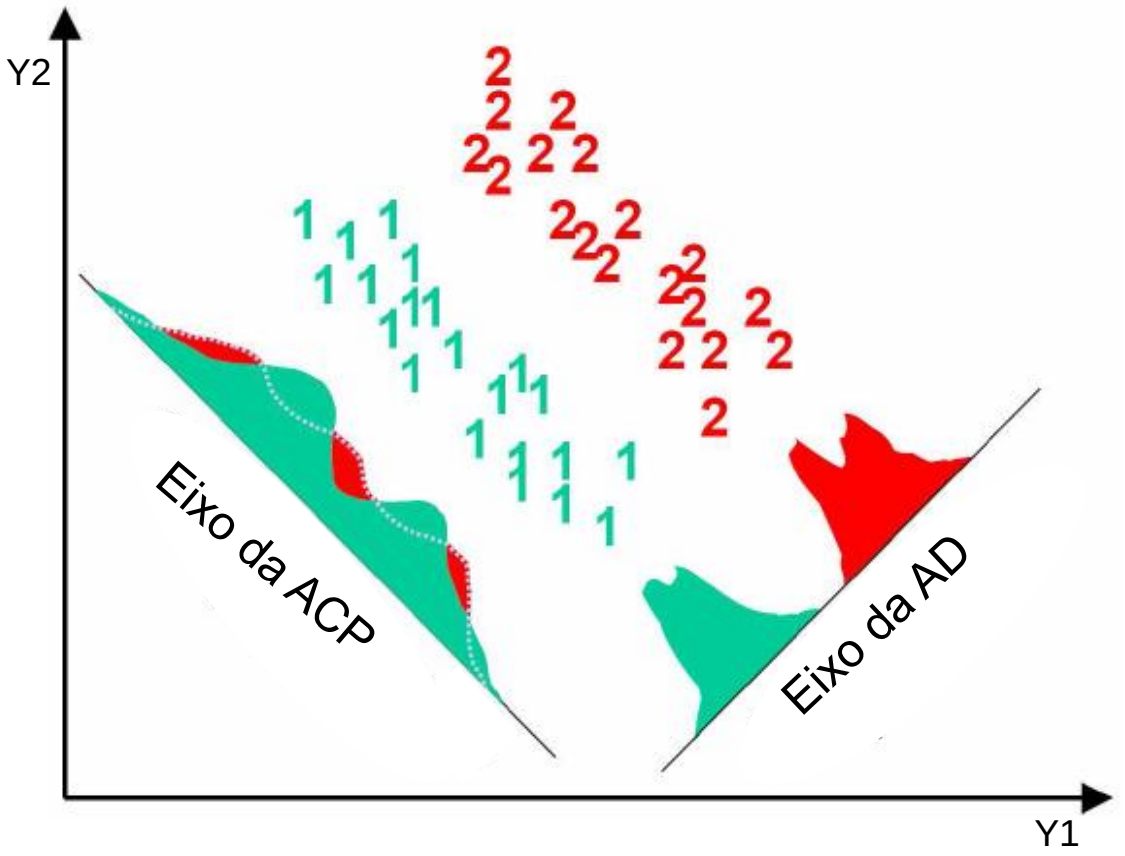
$$\sum_{j=1}^{m}\left(X_{0j}-\bar{X}_{kj}\right)^2 = \sum_{j=1}^{m}\left[l_j'\left(Y_0-\bar{Y}_k\right)\right]^2 \le \sum_{j=1}^{m}\left[l_j'\left(Y_0-\bar{Y}_g\right)\right]^2 \quad k\neq g,\ g=1,..,G$$

Análise de Componentes Principais
G=2

$$\max_{\|a\|=1} \frac{a'\Sigma a}{a'a}$$

$$\Sigma = P\Lambda P';$$
$$a_j = P_j;$$
$$\Sigma\ P_j = \lambda_j P_j$$

$$\hat{\Sigma} = \frac{1}{n-1} S_T$$
$$= \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})'$$



Análise Discriminante
G=2

$$\max_l \frac{l'\Sigma_b l}{l'\Sigma_w l}$$

$$\left|\Sigma_b - \lambda\Sigma_w^{-1}\right| = 0;$$
$$l_j = \Sigma_w^{-1/2} P_j;$$
$$\Sigma_b P_j = \lambda_j \Sigma_w P_j$$

S_T, S_b, S_w

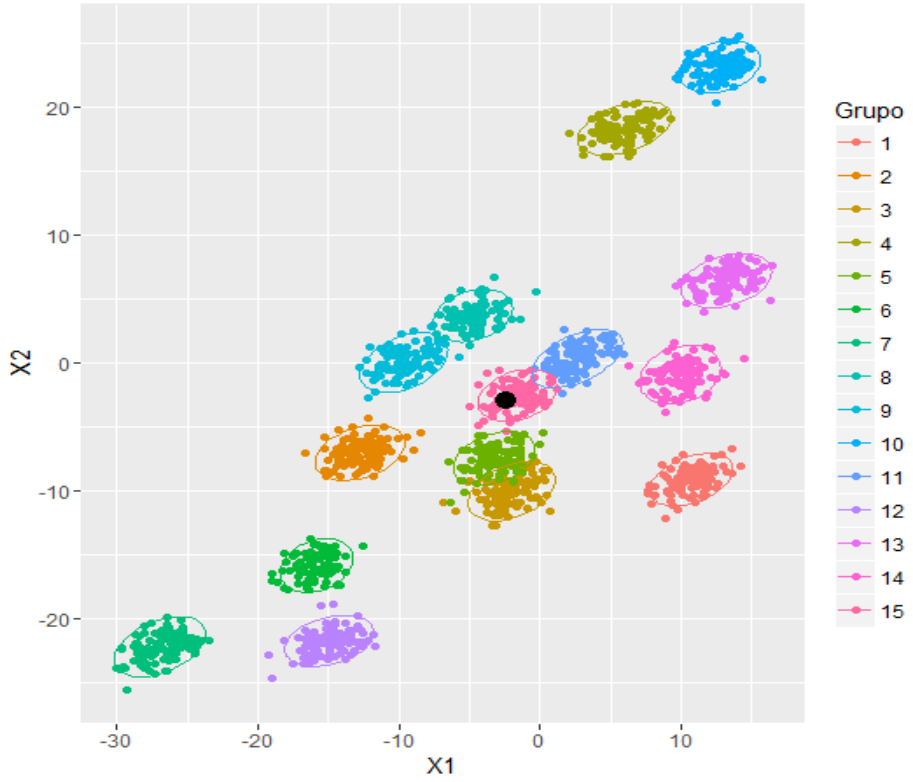
Modelo de MANOVA

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g} (Y_{gi} - \bar{Y}_{..})(Y_{gi} - \bar{Y}_{..})' = \sum_{g=1}^{G} n_g (\bar{Y}_g - \bar{Y})(\bar{Y}_g - \bar{Y})' + \sum_{g=1}^{G}\sum_{i=1}^{n_g} (Y_{gi} - \bar{Y}_g)(Y_{gi} - \bar{Y}_g)'$$

T $\qquad$ $B = H$ $\qquad$ $W = E$

Funções Discriminantes de Fisher

Populações Estratificadas em Muitos Grupos (G>2)

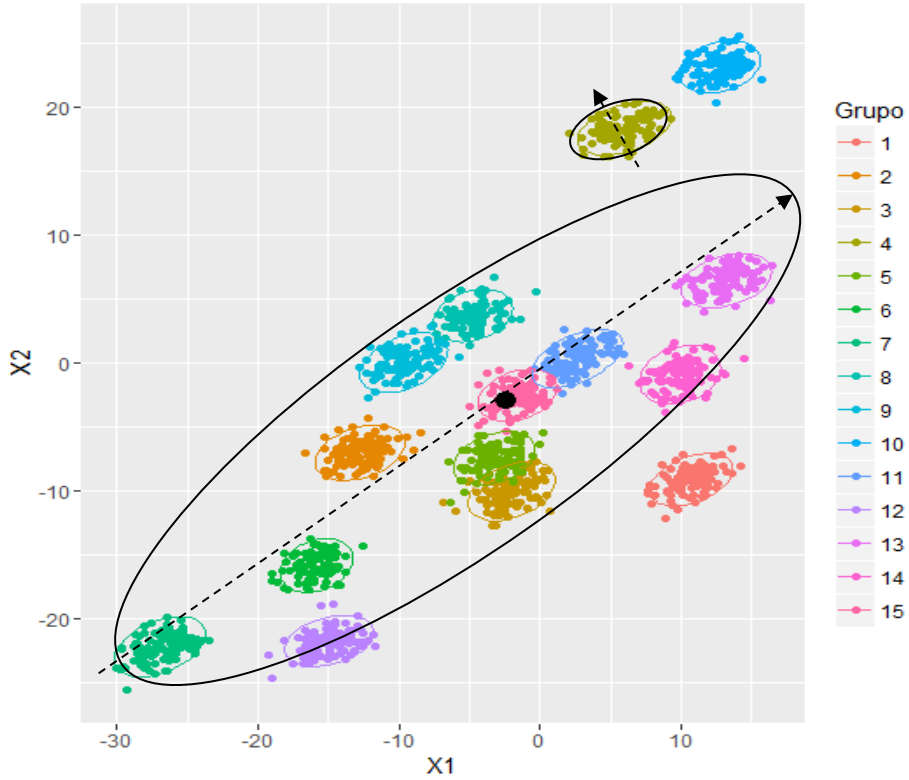


Como realizar a redução dos dados (p=2, G=15)? Uma única dimensão ($X=l'Y$) é suficiente para uma boa discriminação dos grupos?

Funções Discriminantes de Fisher

Populações Estratificadas em Muitos Grupos (G>2)

$$\max_{l} \frac{l' \hat{B} l}{l' \hat{\Sigma} l}$$



Variabilidade Dentro de grupos

Variabilidade Entre grupos

A direção discriminante ótima é aquela que maximiza B (eixo de variação ENTRE grupos) relativamente a W (eixo de variação DENTRO de grupos).

Análise Discriminante

Foco na classificação de observações, mais do que na obtenção de vetores reducionistas

Problema Geral de Classificação

Caso de Duas Populações (G=2)

Suposição: Uma população está estratificada em 2 subpopulações, $\tau 1$ e $\tau 2$, e de cada subpopulação é retirada uma amostra de tamanho n1 e n2, respectivamente.

Com base na amostra, para encontrar uma regra de discriminação de observações de cada população, uma alternativa é particionar o espaço amostral Ω em duas regiões, R1 e R2, que favoreçam às populações $\tau 1$ e $\tau 2$, respectivamente, tal que, para uma observação Y_0 tem-se que, se

$Y_0 \in R_1 \Rightarrow$ a observação é de $\tau 1$

$Y_0 \in R_2 \Rightarrow$ a observação é de $\tau 2$

Regra discriminante

Como determinar R1 e R2 ?

Análise Discriminante

Problema Geral de Discriminação/Classificação - Solução Probabilística

Caso de Duas Populações

Probabilidades a priori: $\tau_1 \Rightarrow p_1(y) \qquad \tau_2 \Rightarrow p_2(y) \qquad p_1 + p_2 = 1$

Função densidade de probabilidades: $\tau_1 \Rightarrow f_1(y) \qquad \tau_2 \Rightarrow f_2(y)$

Probabilidade de Classificação Errada:

$$P(2|1) = P(Y_i \in R_2 \mid \tau_1) = \int_{R_2 = \Omega - R_1} f_1(y)\, dy$$

$$P(1|2) = P(Y_i \in R_1 \mid \tau_2) = \int_{R_1 = \Omega - R_2} f_2(y)\, dy$$

Probabilidade de Classificação Correta:

$$P(1|1) = P(Y_i \in R_1 \mid \tau_1) = \int_{R_1 = \Omega - R_2} f_1(y)\, dy$$

$$P(2|2) = P(Y_i \in R_2 \mid \tau_2) = \int_{R_2 = \Omega - R_1} f_2(y)\, dy$$

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações

Notação

Probabilidade de Classificação

Verdade	Predito τ_1	Predito τ_2
τ_1	P(1,1)	P(2,1)
τ_2	P(1,2)	P(2,2)

Custo de Classificação

Verdade	Predito τ_1	Predito τ_2
τ_1	0	c(2|1)
τ_2	c(1|2)	0

$$P(1,1) = P(1|1)p_1 \qquad P(2,1) = P(2|1)p_1$$

$$P(1,2) = P(1|2)p_2 \qquad P(2,2) = P(2|2)p_2$$

Logo, o custo esperado de classificação errada (*CECE*) é dado por:

$$\begin{aligned} CECE &= c(2|1)P(2,1) \quad + \quad c(1|2)P(1,2) \\ &= c(2|1)P(2|1)p_1 \quad + \quad c(1|2)P(1|2)p_2 \end{aligned}$$

$\Rightarrow$ Obter R1 e R2 que minimizem *CECE*

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações

Minimizar o custo esperado de classificação errada:

$$CECE = c(2|1)\,P(2|1)p_1 \quad + \quad c(1|2)\,P(1|2)p_2$$

$$= c(2|1)\,p_1 \int_{R_2} f_1(y)\,dy \quad + \quad c(1|2)\,p_2 \int_{R_1} f_2(y)\,dy$$

$$= c(2|1)\,p_1 \quad + \int_{R_1} \underbrace{\left[c(1|2)\,p_2\; f_2(y) - c(2|1)\,p_1\; f_1(y)\right]}_{\leq 0 \Rightarrow \text{mínimo } CECE} dy$$

Somando e subtraindo $c(2|1)\,p_1 \int_{R_1} f_1(y)\,dy$ tem-se:

R1 e R2 são conjuntos de valores $Y \in \Re^p$ para os quais:

$$R_1: \quad \frac{f_1(y)}{f_2(y)} \geq \frac{c(1|2)p_2}{c(2|1)p_1} \qquad R_2: \quad \frac{f_1(y)}{f_2(y)} < \frac{c(1|2)p_2}{c(2|1)p_1}$$

Discriminação sob Estimação

Problema Geral de Classificação - Caso de Duas Populações Normais

Função densidade de probabilidades:

heterocedasticidade

$$\tau_g \Rightarrow f_g(y) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_g\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(Y-\mu_g\right)'\Sigma_g^{-1}\left(Y-\mu_g\right)\right\}; \quad g=1,2; Y \in \Re^p$$

Classificar uma observação em $\tau 1$ se $Y \in \Re^p$ pertencer à região R1 dada por:

$$R_1: \quad -\frac{1}{2}Y'\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)Y + \left(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1}\right)Y - c \geq \ln\left[\frac{c(1|2)\,p_2}{c(2|1)\,p_1}\right]$$

em que, $c = \frac{1}{2}\ln\left(\frac{\left|\Sigma_1\right|}{\left|\Sigma_2\right|}\right) + \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right)$

R2 é dada pelo complementar de R1 em Ω.

Sob Heterocedasticidade $\Rightarrow$ Função Discriminante Quadrática (em $Y \in \Re^p$)

Regra de Discriminação Amostral: obter estimador de MVS

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações Normais

Regra de discriminação: (os parâmetros são substituídos por sua estimativas)

Alocar Y_0 em $\tau 1$ se

$$X_0^Q$$

$$-\frac{1}{2}Y_0'\left(S_1^{-1}-S_2^{-1}\right)Y_0+\left(\bar{X}_1'S_1^{-1}-\bar{X}_2'S_2^{-1}\right)Y_0-\hat{c}_Q \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$$

Alocar Y_0 em $\tau 2$ caso contrário

em que, $\hat{c}_Q=\frac{1}{2}\ln\left(\frac{|S_1|}{|S_2|}\right)+\frac{1}{2}\left(\bar{Y}_1'S_1^{-1}\bar{Y}_1'-\bar{Y}_2'S_2^{-1}\bar{Y}_2'\right)$

Função discriminante quadrática

Critério flexível: permite heterocedasticidade, custos e prioris diferentes

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações Normais

$$Y_i \in \tau_k \; ; Y_i \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma_g\right) \quad g = 1,2$$

$\Rightarrow$ **Suposição:** $\Sigma_1 = \Sigma_2 = \Sigma$

Regra de discriminação:

Alocar Y_0 em $\tau 1$ se

$$\underbrace{\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_0}_{X_0^L} \underbrace{- \frac{1}{2}\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}\left(\bar{Y}_1 + \bar{Y}_2\right)}_{\hat{c}} \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$$

Alocar Y_0 em $\tau 2$ caso contrário

$\Rightarrow$ Note que a função discriminante X_0^L é linear em Y_0

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações Normais

$$Y_i \in \tau_k \; ; Y_i \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma\right) \quad g = 1,2$$

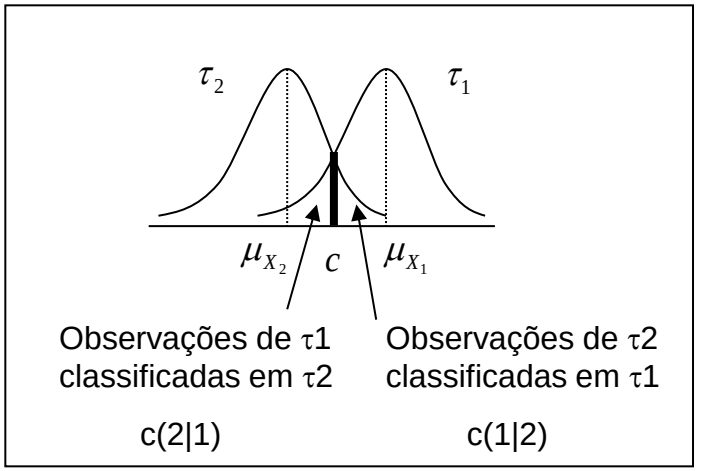
homocedasticidade

$\Rightarrow$ Função Discriminante Linear

Alocar Y_0 em $\tau 1$ se

$$X_0 - \hat{c} \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$$

Alocar Y_0 em $\tau 2$ caso contrário



- Se os custos e as prioris são iguais $\Rightarrow$ função discriminante linear de Fisher
- Se $c(2|1) > c(1|2)$ e $p_1 = p_2$ $\Rightarrow$ o limite "c" é deslocado para a esquerda
- Se $p_1 < p_2$ e $c(2|1) = c(1|2)$ $\Rightarrow$ o limite "c" é deslocado para a direita

Análise Discriminante

Banco	Condição	Y1	Y2	Y3	Y4
B1	1	0,8888	0,7391	1,0255	0,3938
B2	1	1,6655	0,7268	0,878	0,0004
B3	1	2,2111	0,9166	0,9492	0,342
B4	1	1,4351	0,9133	0,9577	0,2325
B5	1	2,1414	0,002	1,0245	0,3966
B6	1	1,192	0,4972	1,034	0,3095
B7	1	1,5895	0,2593	1,0453	0,557
B8	1	1,3272	0,4126	1,0448	0,3482
B9	1	1,8847	0,388	0,9864	0,0337
B10	1	0,5229	0,9473	1,1244	0,118
n		10	10	10	10
Média		1,4852	0,5802	1,007	0,2732
D.P.		0,533	0,319	0,0674	0,1762
B11	2	0,4922	0,3166	1,1127	0,1628
B12	2	1,4427	0,0589	0,9019	0,1355
B13	2	0,5438	0,5358	1,03	0,1481
B14	2	0,1904	0,7087	0,9917	0,2625
B15	2	0,1102	0,7378	1,528	0,0783
B16	2	2,006	0,014	1,0321	0,0816
B17	2	0,2321	0,9234	0,9753	0,0045
B18	2	0,9019	0,1634	1,1414	0,5485
B19	2	1,9757	0,3395	0,9997	0,0751
B20	2	0,7276	0,3139	1,1077	0,2957
n		10	10	10	10
Média		0,862	0,4112	1,0821	0,1793
D.P.		0,712	0,3055	0,1726	0,1567

Condição:

1: Com problemas

2: Sem problemas

Objetivo:

Obter uma função de discriminação com base nas 4 variáveis de indicadores econômicos

⇒Obtenha a função discriminante linear e quadrática

⇒ Quais suposições estão implícitas em cada caso?

Análise Discriminante

Dados dos Bancos

$$\bar{Y}_{g=1} = \begin{pmatrix} 1{,}486 \\ 0{,}580 \\ 1{,}007 \\ 0{,}273 \end{pmatrix}$$

$$S_{g=1} = \begin{pmatrix} 0{,}284 & & & \\ -0{,}070 & 0{,}102 & & \\ -0{,}021 & -0{,}004 & 0{,}005 & \\ 0{,}008 & -0{,}022 & 0{,}004 & 0{,}031 \end{pmatrix}$$

$$\bar{Y}_{g=2} = \begin{pmatrix} 0{,}862 \\ 0{,}414 \\ 1{,}082 \\ 0{,}179 \end{pmatrix}$$

$$S_{g=2} = \begin{pmatrix} 0{,}505 & & & \\ -0{,}164 & 0{,}091 & & \\ -0{,}051 & 0{,}014 & 0{,}030 & \\ -0{,}012 & -0{,}016 & 0{,}002 & 0{,}025 \end{pmatrix}$$

Usar o teste M de Box para verificar a homocedasticidade!

$$S_c = \begin{pmatrix} 0{,}395 & & & \\ -0{,}117 & 0{,}096 & & \\ -0{,}036 & 0{,}005 & 0{,}017 & \\ -0{,}002 & -0{,}019 & 0{,}003 & 0{,}028 \end{pmatrix}$$

Análise Discriminante

Dados dos Bancos

homocedasticidade

Suposição: $Y_i \in \tau_g \; ; Y_i \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma_g\right) \quad g = 1,2 \qquad \Sigma_1 = \Sigma_2 = \Sigma$

Custos de classificação Errada e Prioris iguais para as populações

$\Rightarrow$ Função Discriminante Linear de Fisher

$$X_0 - c \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] \quad \Rightarrow \quad X_0 - c - \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] \geq 0$$

$$\underbrace{\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_0}_{X_0^L} \underbrace{- \frac{1}{2}\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}\left(\bar{Y}_1 + \bar{Y}_2\right)}_{c} \underbrace{- \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]}_{=0}$$

$$4{,}273Y_1 + 8{,}820Y_2 + 0{,}346Y_3 + 9{,}611Y_4 - 11{,}933$$

Análise Discriminante

Dados dos Bancos

Suposição: $Y_i \in \tau_g \; ; Y_i \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma_g\right) \quad g = 1,2$ heterocedasticidade

Custos de classificação Errada e Probabilidades a Priori iguais para as populações

$\Rightarrow$ Função Discriminante Quadrática

$$\underbrace{-\frac{1}{2} Y_0' \left(S_1^{-1} - S_2^{-1}\right) Y_0 + \left(\bar{Y}_1' S_1^{-1} - \bar{Y}_2' S_2^{-1}\right) Y_0 - \hat{c}_Q}_{X_0^Q} \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] = 0$$

$$-0,214Y_1^2 + 14,535Y_2^2 - 204,116Y_3^2 + 14,038Y_4^2$$
$$+9,332Y_1Y_2 + -38,603Y_1Y_3 + 16,846Y_1Y_4 - 35,125Y_2Y_3 + 31,732Y_2Y_4 + 43,362Y_3Y_4$$
$$+38,194Y_1 + 17,076Y_2 + 478,004Y_3 - 73,415Y_4 - 273,776$$

Análise Discriminante

Dados dos Bancos

Banco	Condição	Regra Linear		Regra Quadrática	
		X	Grupo	X	Grupo
B1	1	3,336	1	6,329	1
B2	1	2,701	1	4,108	1
B3	1	10,223	1	27,94	1
B4	1	5,825	1	13,043	1
B5	1	1,403	1	2,861	1
B6	1	1,425	1	3,046	1
B7	1	3,146	1	6,822	1
B8	1	1,539	1	3,359	1
B9	1	0,635	1	1,247	1
B10	1	1,222	1	1,151	1
B11	2	-4,74	2	-1,808	2
B12	2	-3,57	2	-5,13	2
B13	2	-2,514	2	-2,313	2
B14	2	-1,223	2	-4,802	2
B15	2	-2,862	2	-39,071	2
B16	2	-1,862	2	-1,397	2
B17	2	-1,4	2	-3,083	2
B18	2	-0,792	2	0,713	1
B19	2	0,945	1	1,411	1
B20	2	-2,484	2	-1,175	2

Função linear classifica melhor!

Realizar um teste da igualdade das matrizes de covariância. Decidir pela função linear (de Fisher deslocada) no caso da não rejeição de $H_0 : \Sigma_1 = \Sigma_2$ (Teste de Box)

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

As Regiões de Classificação que minimizam *CEEC* são definidas por alocar Y_0 à população τ_k, k=1,2,…,G, que atinge o mínimo erro de classificação, dado por:

$$\sum_{\substack{g=1 \\ g \neq k}}^{G} p_g \; f_g(y)\, c(k \mid g)$$

Logo, se todos os custos são iguais, devemos alocar Y_0 à população τ_k se:

$$p_k \; f_k(y) > p_g \; f_g(y) \qquad g = 1,\ldots,G; g \neq k$$

ou, equivalentemente: $\ln p_k \; f_k(y) > \ln p_g \; f_g(y) \qquad g = 1,\ldots,G; g \neq k$

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

Alocar Y_0 a τ_k se: $\ln p_k \ f_k(y) > \ln p_g \ f_g(y) \qquad g = 1,...,G; g \neq k$

Caso Especial (N_p): $Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma_g)$ heterocedasticidade

$$f_g(y) = \frac{1}{(2\pi)^{p/2}|\Sigma_g|^{1/2}} \exp\left\{-\frac{1}{2}(Y-\mu_g)'\Sigma_g^{-1}(Y-\mu_g)\right\}, \quad g = 1,2,...,G$$

$$\ln p_k f_k(y) = \ln p_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(Y-\mu_k)'\Sigma_k^{-1}(Y-\mu_k) = \max_g \ln p_g f_g(y)$$

Define-se Escore Discriminante Quadrático (de $Y \in \Re^p$) para a g-ésima população:

$$d_g^Q(y) = -\frac{1}{2}\ln|\Sigma_g| - \frac{1}{2}(Y-\mu_g)'\Sigma_g^{-1}(Y-\mu_g) + \ln p_g \qquad g = 1,...,G$$

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

$$\ln p_k \; f_k(y) > \ln p_g \; f_g(y) \qquad g = 1,...,G; g \neq k$$

heterocedasticidade

$Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma_g) \Rightarrow$ Alocar Y a τ_k se o escore quadrático $d_k^Q(y)$ é maior que os demais

em que, $d_k^Q(y) = -\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(Y-\mu_k)'\Sigma_k^{-1}(Y-\mu_k) + \ln p_k \quad k = 1,...,G$

homocedasticidade

Se $Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma)$, isto é, $\Sigma_1 = ... = \Sigma_g$

$$d_k^Q(y) \Rightarrow d_k(y) = \mu_k' \Sigma^{-1} Y - \frac{1}{2}\mu_k' \Sigma^{-1}\mu_k + \ln p_k \quad k = 1,...,G$$

Escore discriminante linear para a população τ_k

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

$$Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma_g)$$

escore discriminante quadrático máximo

$$d_k^Q(y)$$

$$-\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(Y-\mu_k)'\Sigma_k^{-1}(Y-\mu_k) + \ln p_k$$

$$Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma)$$

escore discriminante linear máximo

$$d_k(y)$$

$$\mu_k'\,\Sigma^{-1}Y - \frac{1}{2}\mu_k'\,\Sigma^{-1}\mu_k + \ln p_k$$

O Escore Discriminante linear pode ser comparado para duas populações, de tal modo que, a condição " $d_k(y)$ é maior ", fica equivalente a:

$$0 \le d_k(y) - d_g(y) = (\mu_k - \mu_g)'\,\Sigma^{-1}Y - \frac{1}{2}(\mu_k - \mu_g)'\,\Sigma^{-1}(\mu_k + \mu_g) + \ln\left(\frac{p_k}{p_g}\right)$$

Alocar Y a τ_k se: $(\mu_k - \mu_g)'\,\Sigma^{-1}Y - \frac{1}{2}(\mu_k - \mu_g)'\,\Sigma^{-1}(\mu_k + \mu_g) \ge \ln\left(\frac{p_g}{p_k}\right)$

Função de Fisher $\qquad$ c

Análise Discriminante

Validação de uma Função de Classificação Amostral

Matriz de Classificação

	Predito		
Verdade	τ_1	τ_2	
τ_1	n_{1c}	n_{1M}	n_1
τ_2	n_{2M}	n_{2c}	n_2

Taxa de Erro Aparente (proporção de itens mal classificados) é dada por:

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \quad \rightarrow \quad AER = p_1 \int_{R_2} f_1(y)dy + p_2 \int_{R_1} f_2(y)dy$$

APER subestima a verdadeira proporção de erro de classificação AER

Alternativas

- Particionar os dados em Amostra de Treinamento e Amostra de Validação
- Método de "Cross-validação"

Análise Discriminante

Validação de uma Função de Classificação Amostral

Algoritmo de Cross Validação

1. Inicie com as observações de τ_1 . Omita uma obs deste grupo e obtenha a função de classificação baseada nos remanescentes n1-1 e n2 observações (supondo G=2)
2. Classifique a obs omitida usando a função calculada no passo 1
3. Repetir os passos 1 e 2 até que todas as obs de τ_1 tenham sido classificadas. Calcule o número de erros de classificação neste grupo
4. Repita os passos de 1 a 3 para as observações do grupo 2.

Taxa de Erro de Classificação esperada é dada por:

$$\hat{E}(APER) = \frac{n_{1M}^{Cross} + n_{2M}^{Cross}}{n_1 + n_2}$$

Análise Discriminante
Padronização de Variáveis

	Unidades Amostrais	Variáveis 1	2	…	j	…	p
G1	1	Y_{111}	Y_{112}		Y_{11j}		Y_{11p}
	2	Y_{121}	Y_{122}		Y_{12j}		Y_{12p}
	…	…	…	…	…		…
	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1j}		Y_{1n1p}
G2	1	Y_{211}	Y_{212}		Y_{21j}		Y_{21p}
	2	Y_{221}	Y_{222}		Y_{22j}		Y_{22p}
	…	…	…	…	…		…
	n_2	Y_{2n21}	Y_{2n22}		Y_{2n2j}		Y_{2n2p}

$\bar{Y}_{1\,p\times 1}$ $S_{1\,p\times p}$

$\bar{Y}_{2\,p\times 1}$ $S_{2\,p\times p}$

$\bar{Y}_{p\times 1}$ $S_{c\,p\times p}$

Na AD a padronização das variáveis é usada com a finalidade de facilitar a interpretação dos pesos das variáveis na função discriminante e no cálculo de “c”.
O R usa a “padronização” das variáveis para calcular as funções discriminantes.
A padronização da variável j avaliada no indivíduo i do grupo g é dada por:

$$Y^*_{gij} = \left(\frac{Y_{gij} - \bar{Y}_j}{S_{gj}} \right)$$

$\bar{Y}_j$: para cada j

S_{gj}: para cada g e j

$$\bar{Y}_j = \frac{1}{n_1 + n_2} \sum_{g=1}^{2} \sum_{i=1}^{n_g} Y_{gij}$$

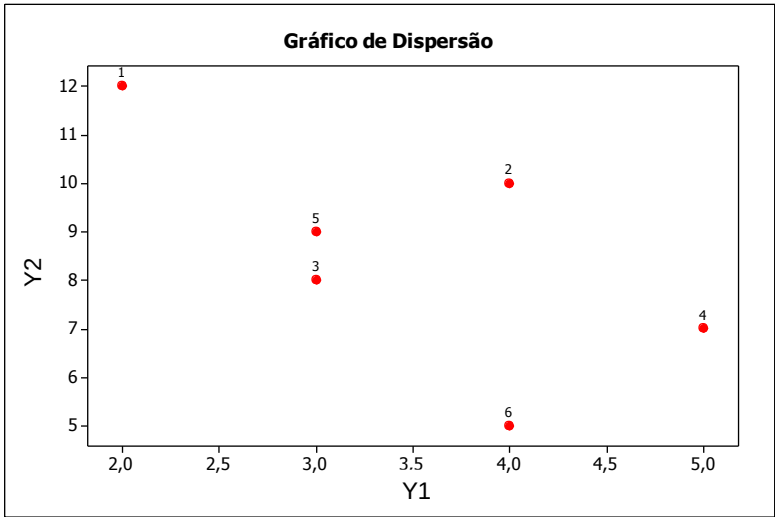
Média da variável j (j=1,..,p), independente de grupo

Análise Discriminante

Considere os dados a seguir em que duas variáveis foram observadas em três indivíduos do grupo 1 e em três indivíduos do grupo 2:

$$G_1 = \begin{pmatrix} 2 & 4 & 3 \\ 12 & 10 & 8 \end{pmatrix} \qquad G_2 = \begin{pmatrix} 5 & 3 & 4 \\ 7 & 9 & 5 \end{pmatrix}$$

1. Calcule a função discriminante de Fisher para a diferença entre os grupos. Qual é a regra de classificação de observações? Que suposições são feitas?
2. Calcule também o escore discriminante para cada grupo via o método geral de classificação. Suponha que p1=p2. E se p1 é duas vezes p2?
3. Calcule a taxa observada de erro de classificação. Classifique a observação (4,7).
4. Calcule a taxa de erro de classificação via validação cruzada.
5. Obtenha a função discriminante para os dados “padronizados”.



Diferentes formulações da função discriminante.

Grupo	Y1	Y2	X=-0,33Y1+0,67Y2	X*=0,33Y1+0,67Y2-4,54	X1=7,33Y1+4,33Y2-32,67	X2=7,67Y1+3,67Y2-28,17	Grupo Pred
1	2	12	7,38	2,84	33,95	31,21	1
1	4	10	5,38	0,84	39,95	39,21	1
1	3	8	4,37	-0,17	23,96	24,2	2
2	5	7	3,04	-1,5	34,29	35,87	2
2	3	9	5,04	0,5	28,29	27,87	1
2	4	5	2,03	-2,51	18,3	20,86	2
			c=4,54				

Solução usando pacote lda do “R”: valores Y estão normalizados para ter variância 1

X = 0.2182179 Y1* - 0.4364358 Y2*
LD1= X-c; c=((-3.7097)+(-2.1821))/2=-2.946
LD1 <= 0 grupo1, cc grupo2
> fit.values$class
[1] 1 1 2 2 1 2

	LD1	
1	-1.8548521	alocar G1
2	-0.5455447	alocar G1
3	0.1091089	alocar G2
4	0.9819805	alocar G2
5	-0.3273268	alocar G1
6	1.6366342	alocar G2

$$X = \left(\bar{Y}_k - \bar{Y}_g\right)' S_c^{-1} Y; \qquad c = \frac{1}{2}\left(\bar{Y}_k - \bar{Y}_g\right)' S_c^{-1}\left(\bar{Y}_k + \bar{Y}_g\right)$$

$$X^* = \left(\bar{Y}_k - \bar{Y}_g\right)' S_c^{-1} Y - c$$

$$X_{gi} = \bar{Y}_g' S_c^{-1} Y_i - \frac{1}{2}\bar{Y}_g' S_c^{-1}\bar{Y}_g = d_g(y_i)$$

Note que, sob normalidade e homocedasticidade, $Y \mid \tau_g \sim N_p\left(\mu_g; \Sigma\right)$, $g = 1, 2$:

$$X_c = \left(\mu_1 - \mu_2\right)' \Sigma^{-1}\left(Y - \mu\right); \qquad \mu = \frac{1}{2}\left(\mu_1 + \mu_2\right)$$

$$X_c \mid Y \in \tau_1 \sim N\left(\frac{1}{2}d_M^2; d_M^2\right),$$

$$X_c \mid Y \in \tau_2 \sim N\left(-\frac{1}{2}d_M^2; d_M^2\right); \quad d_M^2 = \left(\mu_1 - \mu_2\right)'\Sigma^{-1}\left(\mu_1 - \mu_2\right)$$

Assim, a probabilidade de classificação errada é,

$$P\left(Y \text{ alocado em} \tau_1 \mid Y \in \tau_2\right) = P\left(X_c(y) > 0 \mid Y \in \tau_2\right) = P\left(Z > \frac{1}{2}d_M\right) = \Phi\left(-\frac{1}{2}d_M\right)$$

##Comandos R

```
#Análise discriminante
dat<-matrix(c(2,4,3,5,3,4,12,10,8,7,9,5,1,1,1,2,2,2),6,3)
xbar<-colMeans(dat[,1:2])
xbar1<-colMeans(dat[1:3,1:2])
xbar2<-colMeans(dat[4:6,1:2])
cov1<-cov(dat[1:3,1:2])
cov2<-cov(dat[4:6,1:2])
library(biotools)
mt<-boxM(dat[,-3],dat[,3])

library(DiscriMiner)
fitlda<-linDA(dat[,-3],dat[,3])

library(MASS) ##outra alternativa de analise
fit<- lda(dat[,3] ~ dat[,1] + dat[,2],prior=c(1,1)/2)
fit.values <- predict(fit, data.frame(dat[,1:2]))
fit.values$x
fit.values$class
ct <- table(dat[,3],fit.values$class)  #tabela com as classificações
diag(prop.table(ct, 1))    # % de classif correta
sum(diag(prop.table(ct)))
fit$svd   # (SSB-\lambda SSW)a=0
mv<-aggregate(fit.values$x, data.frame(dat[,3]), FUN=mean)
colMeans(mv[2])
```

Análise de Agrupamento

Análise Multivariada de Dados

	Variáveis					
Unidades Amostrais	1	2	…	j	…	p
1	Y_{11}	Y_{12}		Y_{1j}		Y_{1p}
2	Y_{21}	Y_{22}		Y_{2j}		Y_{2p}
…	…	…	…	…		…
i	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ip}
…	…	…	…	…	…	…
n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{np}

Objetivos:

Análise no $\Re^{nxn}$

- Formação de grupos de unidades amostrais $\Rightarrow$ agrupamento de observações $\Rightarrow$ grupos homegêneos internamente e heterogêneos externamente
- Identificar similaridades entre Variáveis $\Rightarrow$ agrupamento de variáveis

ANÁLISE DE AGRUPAMENTO (*Cluster*)

Motivação : Escalonamento Multidimensional

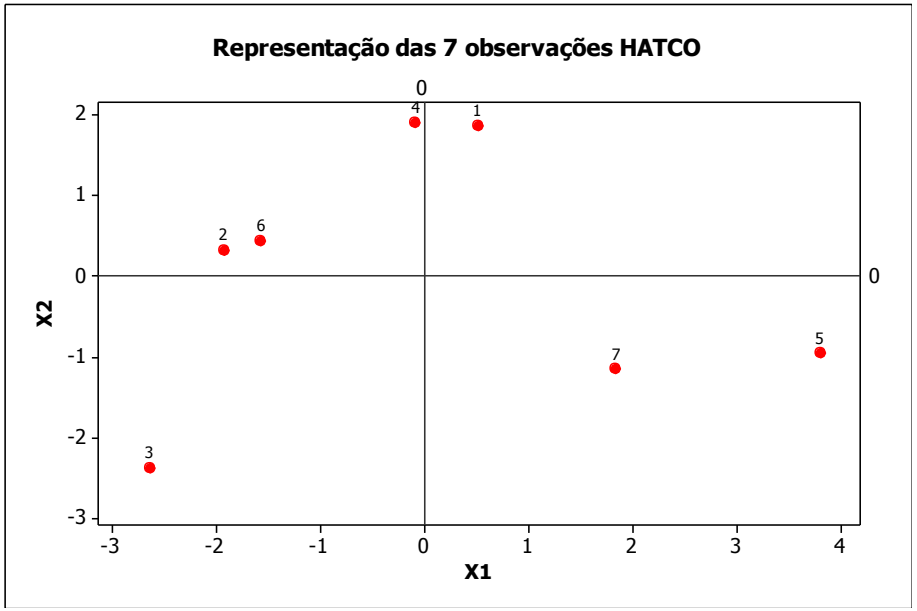
Matriz de Distância Euclidiana entre as “7” primeiras observações do banco de dados HATCO. Como estas observações estão agrupadas?

$\hat{D} \setminus D$

0	3,87685	5,32823	1,98997	4,64758	3,63043	3,88201
2,88822	0	3,37046	2,53574	5,93212	2,43105	4,40908
5,27905	2,7818	0	5,13323	6,67158	3,87685	4,97192
0,60254	2,42927	4,97967	0	4,82908	2,90172	3,83536
4,32705	5,86392	6,59102	4,82402	0	5,92453	2,38118
2,53214	0,3692	2,99825	2,09374	5,54915	0	3,83536
3,29295	4,04138	4,64049	3,61279	1,96945	3,76315	0

$X=$

0,52641	1,85717
-1,91506	0,3141
-2,63012	-2,37423
-0,07463	1,89971
3,80807	-0,96312
-1,56387	0,42802
1,84837	-1,15878



BANCO DE DADOS: HATCO

Unidades amostrais: Clientes da HATCO (Hair et al., 2005)

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45	4,4	0	1	1	2
							…							
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1	0	60	5,2	1	0	1	3
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4	1	38	3,7	0	1	0	1
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4	1	42	4,3	0	1	0	1
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0	0	33	4,4	1	0	0	1

Var. da Percepção dos Clientes sobre o Fornecedor (HATCO)

Escala: [0,10]

Var. do Produto

Escala X9: [0,100]

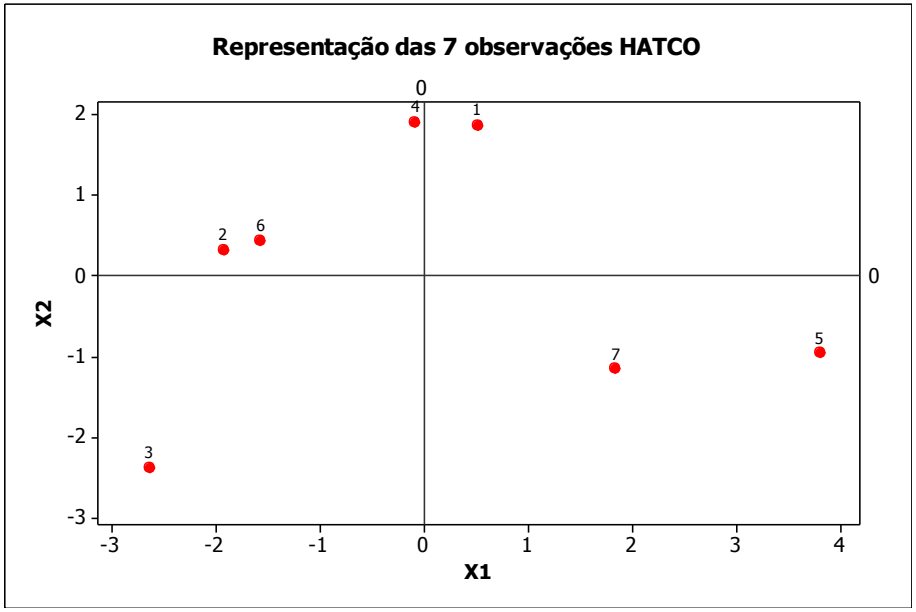
Escala X10: [0,10]

Demais variáveis: Características do Cliente (empresa compradora)

Análise de Agrupamentos

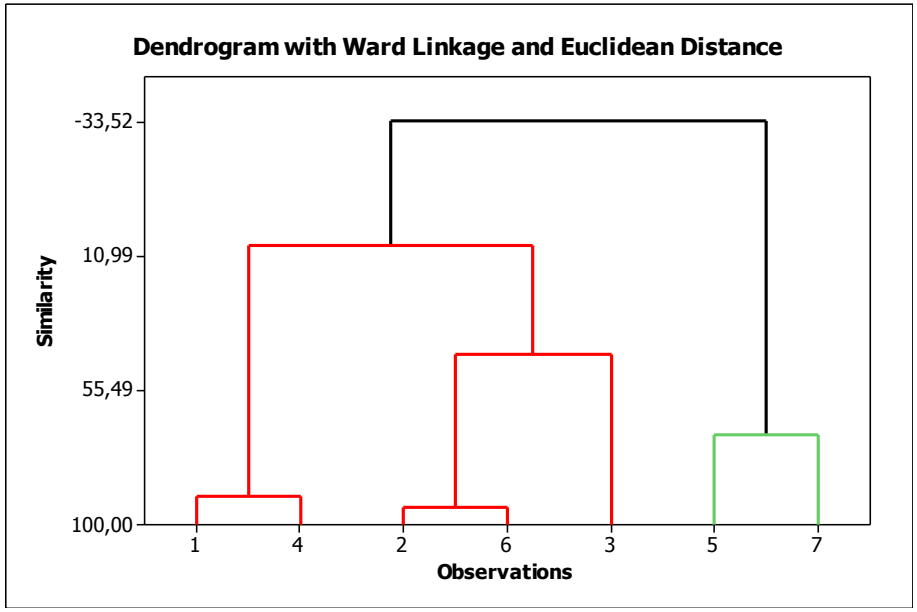
Escalonamento Multidimensional

Coordenadas principais



Análise de Agrupamento

Dendrograma das 7 obs. de HATCO



Como estes agrupamentos foram formados?

Análises realizadas à partir da Matriz de distâncias entre as observações

Motivação: Formação de Grupos

Taxa de delitos (por 100.000 hab.) por divisão territorial de polícias do Estado de São Paulo (Deinter), em 2002*.

Deinter	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10,85	1500,8	149,35	108,38
RP	14,13	1496,07	187,99	116,66
Bauru	8,62	1448,79	130,97	69,98
Campinas	23,04	1277,33	424,87	435,75
Sorocaba	16,04	1204,02	214,36	207,06
SP	43,74	1190,94	1139,52	909,21
SJC	25,39	1292,91	358,39	268,24
Santos	42,86	1590,66	721,9	275,89
Média	23,08	1375,19	415,92	298,9
dp	13,69	152,05	351,62	273,35

*Barroso, L; Artes, R (2002)

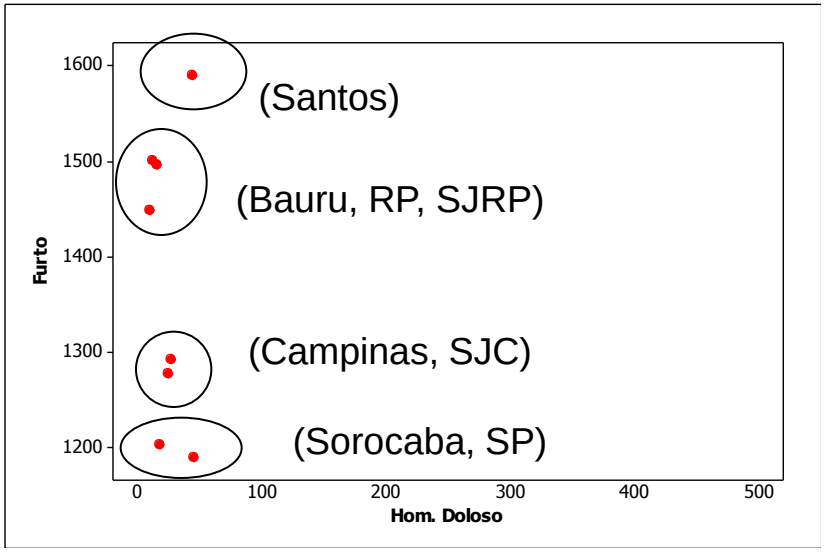
Como podemos agrupar as regiões?

Formação de Grupos

Considere duas variáveis (p=2) $\Rightarrow$ Diagrama de Dispersão (critério subjetivo de formação dos grupos)

Deinter	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10,85	1500,8	149,35	108,38
RP	14,13	1496,07	187,99	116,66
Bauru	8,62	1448,79	130,97	69,98
Campinas	23,04	1277,33	424,87	435,75
Sorocaba	16,04	1204,02	214,36	207,06
SP	43,74	1190,94	1139,52	909,21
SJC	25,39	1292,91	358,39	268,24
Santos	42,86	1590,66	721,9	275,89
Média	23,08	1375,19	415,92	298,9
dp	13,69	152,05	351,62	273,35

Formação de 4 grupos



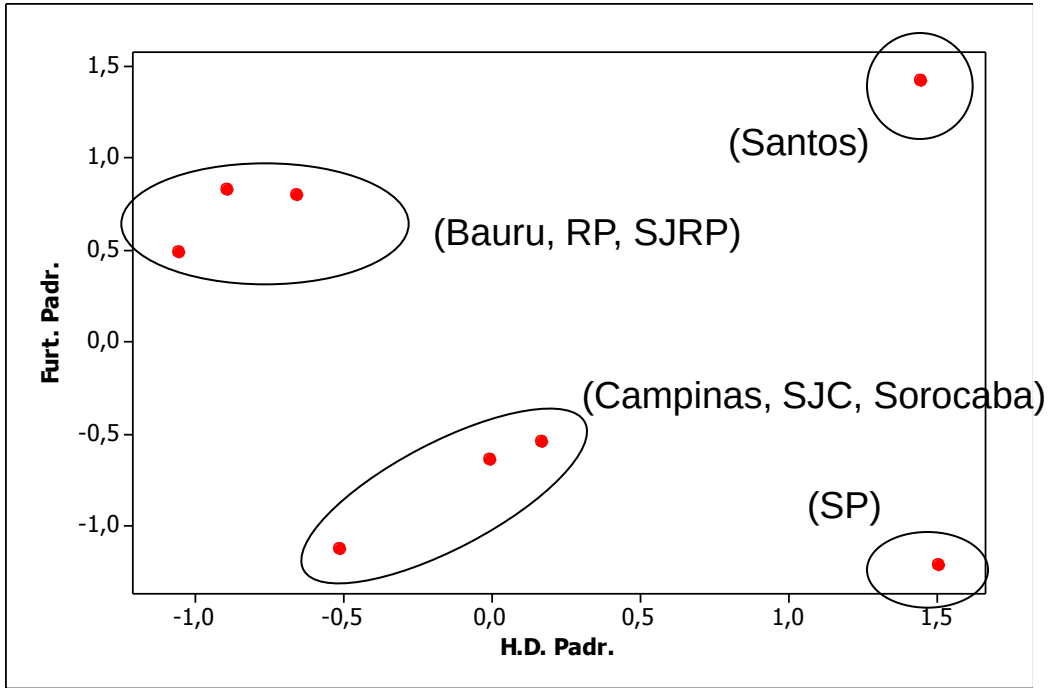
(mesma escala)

A variável Furto contribuiu mais para a formação dos grupos e Homicídio Doloso contribui pouco $\Rightarrow$ Padronizar as variáveis (atribuir igual importância às variáveis)

Formação de Grupos

Uso de variáveis PADRONIZADAS ⇒ diagrama de dispersão mostra que as distâncias no sentido vertical e horizontal são da mesma grandeza ⇒ as duas variáveis estão recebendo importância equivalentes

Formação de 4 grupos



Formação de Grupos

Resultados do agrupamento das regiões (Deiters) via variáveis originais e variáveis padronizadas $\Rightarrow$ uso das variáveis homicídio doloso e furtos (critério: inspeção visual do gráfico de dispersão)

Var. originais	Var. padronizadas
Sorocaba,SP	SP
Campinas,SJC	Campinas, SJC,Sorocaba
Bauru,SJC,SJRP	Bauru, RP,SJRP
Santos	Santos

Análise de Agrupamentos

Etapas da Aplicação de uma Análise de Agrupamento:

- **Escolha do Critério de Parecença**: adotar uma medida de distância (ou proximidade) entre pontos (uso de v*ariáveis originais ou padronizadas)*.
- **Definicão do Número de Grupos**: decisão a priori ou a posteriori (com base nos resultados da análise)
- **Formação dos grupos**: definir o algoritmo de formação dos grupos.
- **Validação do Agrupamento**: é comum supor que cada grupo seja uma amostra aleatória de uma subpopulação e aplicar técnicas inferenciais (comparações de médias dos grupos, por ex.). Algumas técnicas descritivas também são usadas (correlação cofenética e gráfico da silhueta).
- **Interpretação dos grupos**: caracterizar os grupos por meio de estatísticas descritivas e gráficos (radar, perfis de médias)

Análise de Agrupamentos

Medidas de Parecença

- Medidas de Similaridade (Ex.: Correlação): quanto maior o valor, maior a semelhança entre os objetos.

- Medidas de Dissimilaridade (Distância): quanto maior o valor, mais diferentes são os objetos.

Variáveis quantitativas

$$d_{ik} = \sqrt{(Y_i - Y_k)'(Y_i - Y_k)} = \sqrt{\sum_{j=1}^{p}(Y_{ij} - Y_{kj})^2}$$

Distância Euclidiana entre observações (Distância de Mahalanobis também ser adotada)

$$d_{ik}^{A} = \sum_{j=1}^{p}\left|Y_{ij} - Y_{kj}\right|$$

Distância de Manhatan (quarteirão)

$$d_{ik}^{M} = \sqrt[m]{\sum_{j=1}^{p}\left|Y_{ij} - Y_{kj}\right|^m} \quad ; m \geq 1$$

Distância de Minkowsky (mais geral)

Análise de Agrupamentos

Medidas de Parecença

- Variáveis Quantitativas: pode-se utilizar o coeficiente de correlação de Pearson como medida de parecença entre pares de unidades amostrais $\Rightarrow$ quanto mais próximo de 1 (ou -1) maior a similaridade e quanto mais próximo de 0 maior a dissimilaridade.

  $\Rightarrow$ Transformar a correlação em uma medida de dissimilaridade

$$d_{ii'} = \left(r_{ii} + r_{i'i'} - 2r_{ii'}\right)^{1/2}$$

- Nem sempre é possível adotar a correlação como medida de parecença entre "unidades amostrais" (pense por quê!)
- O coeficiente *r* é comumente usado como medida de "parecença/correlação" entre variáveis (e não entre unidades amostrais)
- A correlação valoriza padrões de forma (tendências) e a distância valoriza mais padrões de tamanho

Análise de Agrupamentos

Algoritmos de Agrupamento

- **Métodos Hierárquicos Aglomerativos**: os agrupamentos hierárquicos partem dos objetos individuais (*n*) para a formação de um único grupo.
  - Método do Vizinho mais Próximo/Perto (Ligação Simples)
  - Método do Vizinho mais Distante/Longe (Ligação Completa)
  - Método das Médias das Distâncias (Ligação Média)
  - Método da Centróide
  - Método de Ward

- **Métodos de Partição**: os agrupamentos não hierárquicos buscam a partição de n objetos em K grupos.
  - Algoritmo das K-Médias

Análise de Agrupamentos

Algoritmos de Agrupamentos Hierárquicos

- **Método do Vizinho mais Distante (Ligação Completa ou Distância Máxima)**; a distância entre os grupos G_1 e G_2 é dada pela maior distância entre os elementos de cada grupo

$$d(G_1, G_2) = \max_{i \in G_1, k \in G_2} d_{ik}$$

$\Rightarrow$ Forma grupos de alta homogeneidade interna

- **Método do Vizinho mais Perto (Ligação Simples ou Distância Mínima)**: a distância entre os grupos G_1 e G_2 é dada pela menor distância entre os elementos de cada grupo

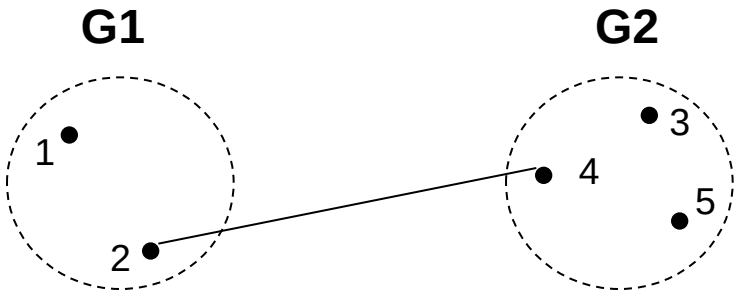
$$d(G_1, G_2) = \min_{i \in G_1, k \in G_2} d_{ik}$$

$\Rightarrow$ Pode não distinguir grupos pobremente separados

- **Método das Médias das Distâncias (Ligação Média)**: a distância entre os grupos é obtida pelo cálculo da média das distâncias entre os elementos de cada grupo

$$d(G_1, G_2) = \frac{\sum_i \sum_k d_{ik}}{n_i \; n_k}$$

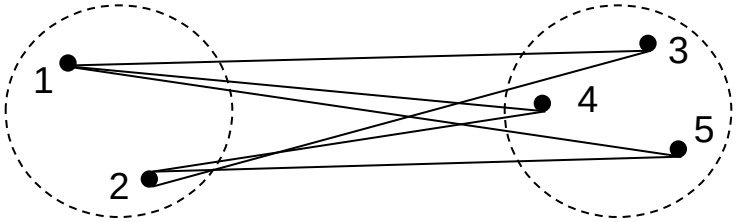
Algoritmos Hierárquicos



Distância entre Grupos

Ligação Simples

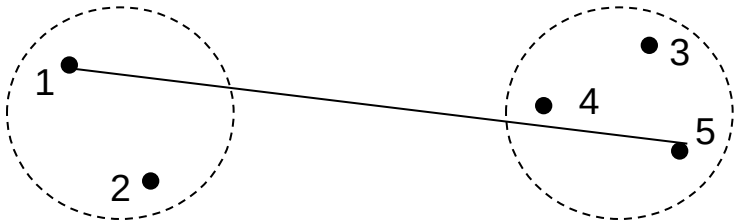
$$d(G_1, G_2) = d_{24}$$



Ligação Média

$$d(G_1, G_2) = \frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$$

2x3

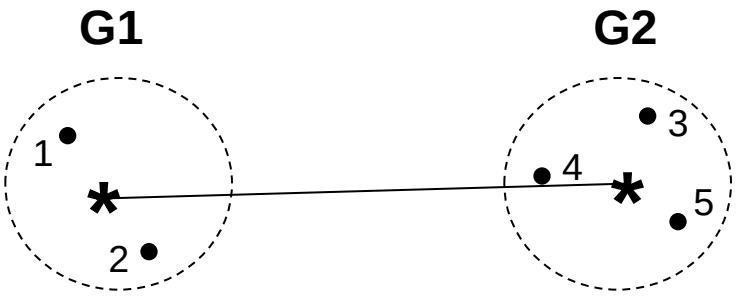


Ligação Completa

$$d(G_1, G_2) = d_{15}$$

Algoritmos Hierárquicos

- **<u>Método da Centróide</u>**; este método define a coordenada de cada grupo como sendo a média das coordenadas de seus elementos. Uma vez obtida esta coordenada comum (denominada centróide) a distância entre os grupos G_1 e G_2 é dada pela distância entre as centróides.



Formação de Grupos

Dados das Deiters

Deinter	Dados Brutos		Dados Padronizados	
	Homicídio doloso	Furto	Homicídio doloso	Furto
SJRP	10,85	1500,8	-0,66382	0,8475
RP	14,13	1496,07	-0,07312	0,81276
Bauru	8,62	1448,79	-1,06542	0,46553
Campinas	23,04	1277,33	1,5315	-0,7937
Sorocaba	16,04	1204,02	0,27086	-1,3321
Média	14,54	1385,4	0	0
dp	5,55	136,16	1	1

Considerando p=2

Matriz de Distância Euclidiana (var. padronizadas)

	SJRP	RP	Bauru	Campins	Sorocaba
SJRP	0				
RP	0,59172	0			
Bauru	0,55425	1,05131	0		
Campinas	2,74098	2,27058	2,88612	0	
Sorocaba	2,37155	2,17227	2,23989	1,3708	0

Como podemos agrupar as regiões?

Análise de Agrupamentos

Método Hierárquico Aglomerativo do Vizinho Mais Distante

	SJRP	RP	Bauru	Campinas	Sorocaba
SJRP	0				
RP	0,59	0			
Bauru	0,55	1,05	0		
Campinas	2,74	2,27	2,89	0	
Sorocaba	2,37	2,17	2,24	1,37	0

Regiões com a menor distância se agrupam: (SJRP,Bauru)

$$d[SJRP, RP] = 0{,}59$$
$$d[Bauru, RP] = 1{,}05$$

$$d[(SJRP, Bauru), RP] = 1{,}05$$

Método do vizinho mais distante (Ligação Completa)

novo objeto

	SJRP,Bauru	RP	Campinas	Sorocaba
SJRP,Bauru	0			
RP	1,05	0		
Campinas	2,89	2,27	0	
Sorocaba	2,37	2,17	1,37	0

Regiões com a menor distância: (SJRP,Bauru) e RP

Análise de Agrupamentos

Método Hierárquico Aglomerativo do Vizinho Mais Distante

	SJRP,Bauru	RP	Campinas	Sorocaba
SJRP,Bauru	0			
RP	1,05	0		
Campinas	2,89	2,27	0	
Sorocaba	2,37	2,17	1,37	0

Regiões com a menor distância se agrupam: (SJRP,Bauru,RP)

$$d[(SJRP, Bauru), Campinas] = 2{,}89$$
$$d[RP, Campinas] = 2{,}27$$

$$\Rightarrow \quad d[(SJRP, Bauru, RP), Campinas] = 2{,}89$$

novo objeto

	SJRP,Bauru,RP	Campinas	Sorocaba
SJRP,Bauru,RP	0		
Campinas	2,89	0	
Sorocaba	2,37	1,37	0

Regiões com a menor distância:

Campinas e Sorocaba

Análise de Agrupamentos

Método Hierárquico Aglomerativo do Vizinho Mais Distante

	SJRP,Bauru,RP	Campinas	Sorocaba
SJRP,Bauru,RP	0		
Campinas	2,89	0	
Sorocaba	2,37	1,37	0

Regiões com a menor distância se agrupam: (Campinas,Sorocaba)

$$\left.\begin{array}{l} d\left[(SJRP, Bauru, RP), Campinas\right] = 2{,}89 \\ d\left[(SJRP, Bauru, RP), Sorocaba\right] = 2{,}37 \end{array}\right\} \; d\left[(SJRP, Bauru, RP), (Camp., Soroc.\right] = 2{,}89$$

Resultados do Método do Vizinho mais Distante

Passo	Grupo	Distância
1	**SJRP,Bauru**	0,55
2	**SJRP,Bauru,RP**	1,05
3	**Campinas,Sorocaba**	1,37
4	**SJRP,Bauru,RP,Campinas,Sorocaba**	2,89

Análise de Agrupamentos

Método Hierárquico Aglomerativo do Vizinho Mais Distante

Resultados do Método do Vizinho mais Distante para as Deiters

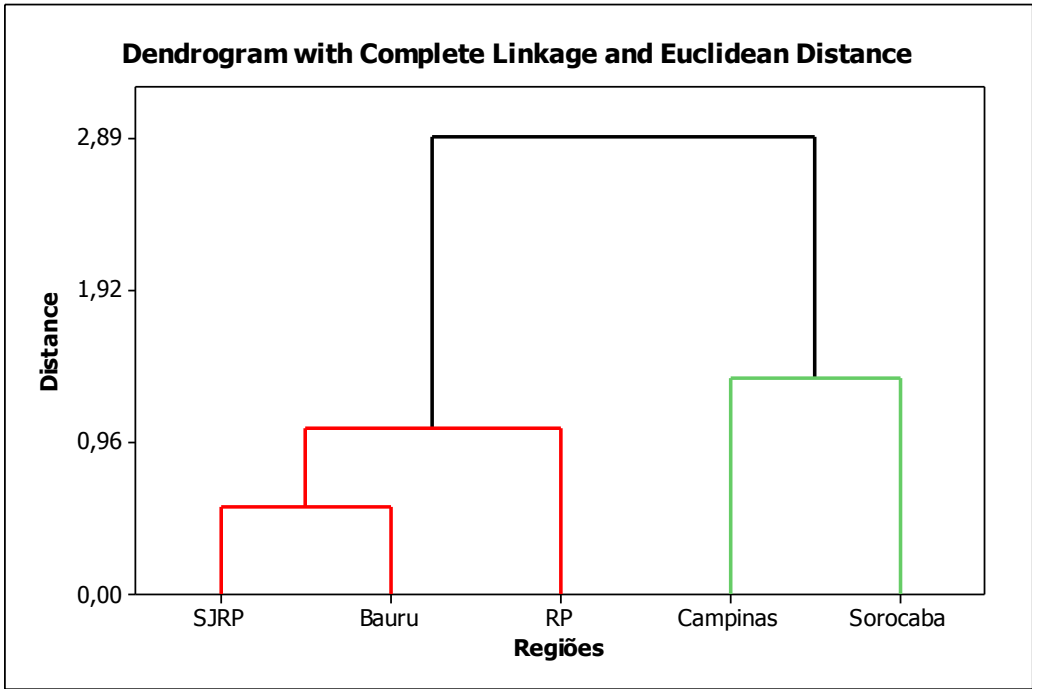
Passo	Grupo	Distância
1	**SJRP,Bauru**	0,55
2	**SJRP,Bauru,RP**	1,05
3	**Campinas,Sorocaba**	1,37
4	**SJRP,Bauru,RP,Campinas,Sorocaba**	2,89

⇒ Construa o **Dendrograma** para representar o procedimento de agrupamento obtido pelo método do vizinho mais distante:

- Eixo das abcissas: objetos na ordem em que foram agrupados
- Eixo das ordenadas: distâncias com que as uniões se realizaram

Análise de Agrupamentos

Passo	Grupo	Distância
1	SJRP,Bauru	0,55
2	SJRP,Bauru,RP	1,05
3	Campinas,Sorocaba	1,37
4	SJRP,Bauru,RP,Campinas,Sorocaba	2,89

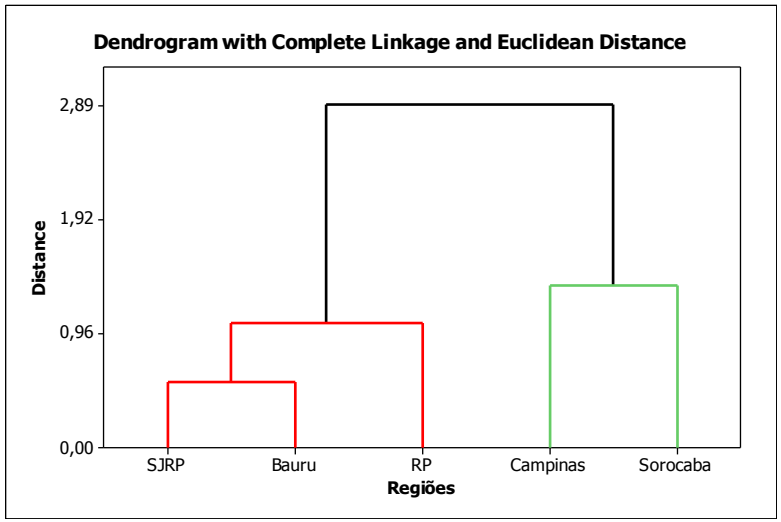


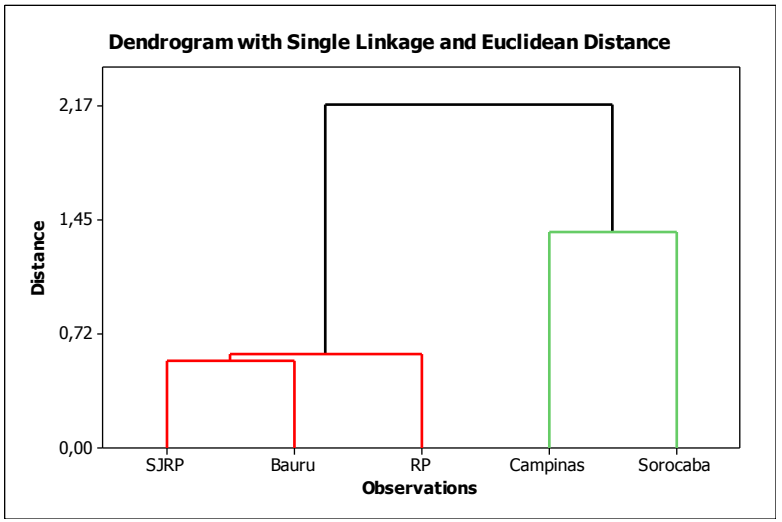
Grandes saltos indicam a união de objetos heterogêneos ⇒ escolha do número de grupos

⇒ Formação de 2 grupos

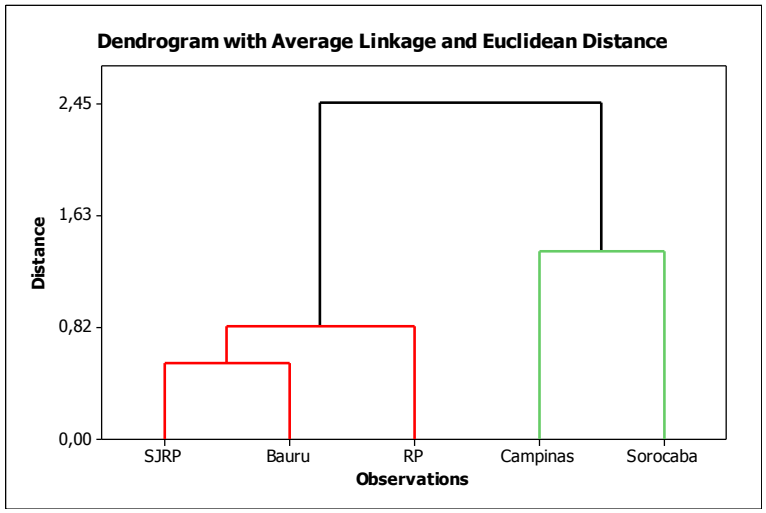
Análise de Agrupamentos

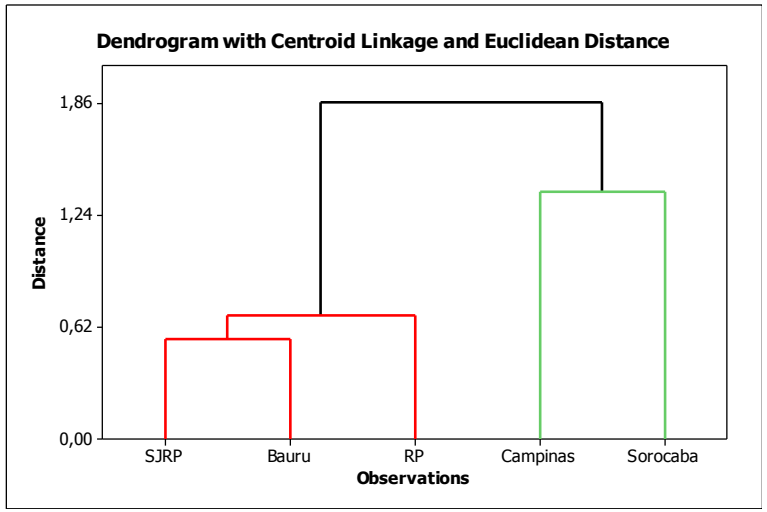
p=2





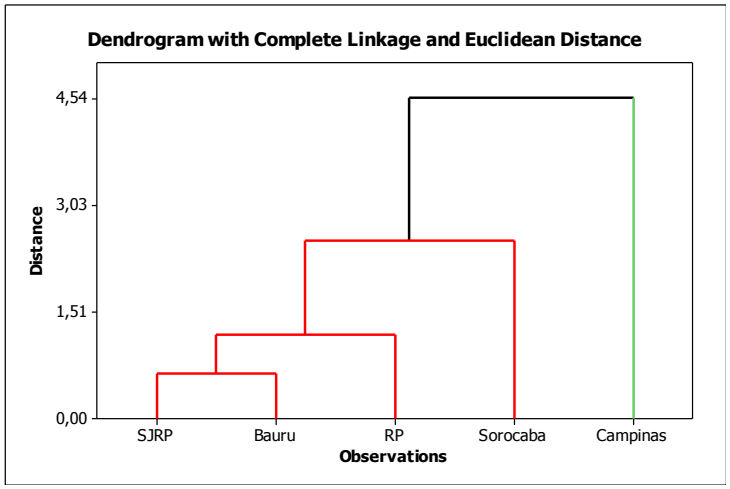
Estabelecer um ponto de corte (no eixo y) para a formação de 2 grupos!

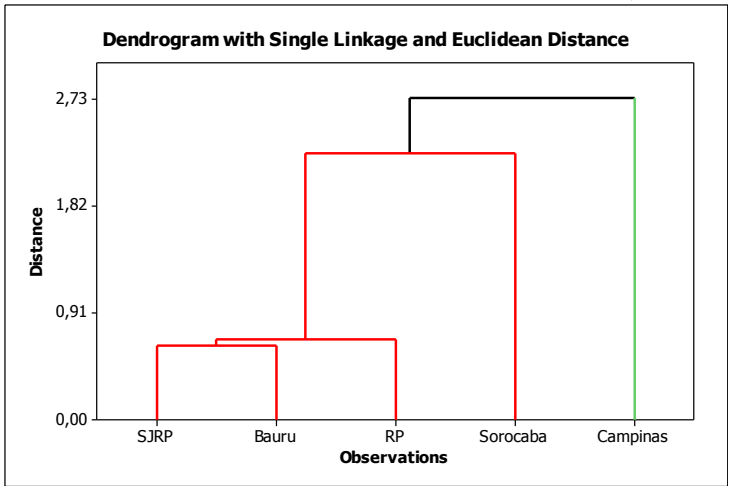




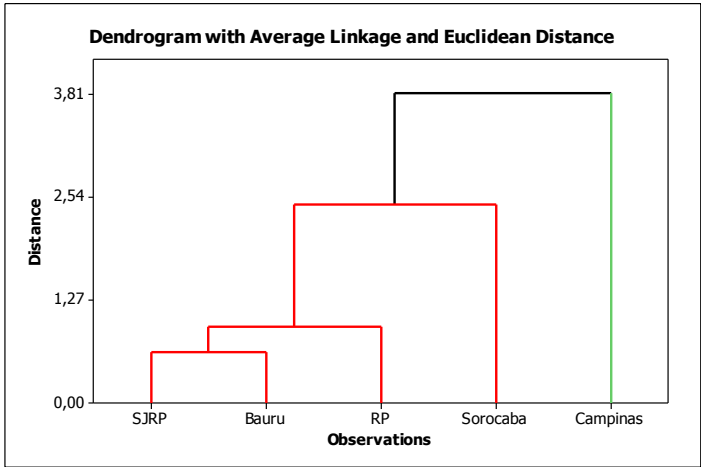
Análise de Agrupamentos

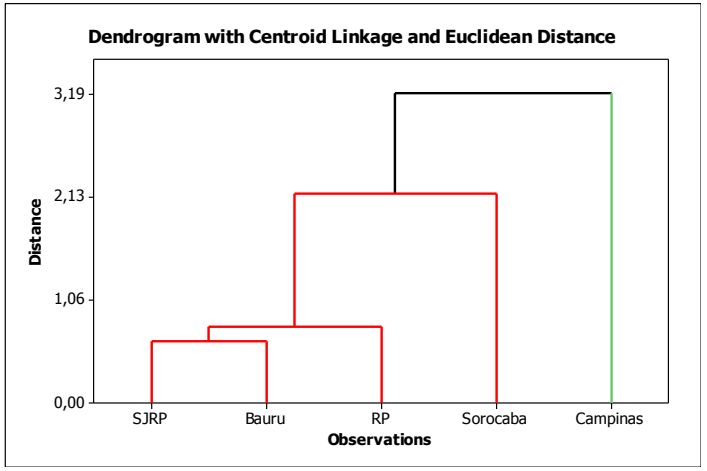
p=4





Formação de 2 grupos!





Algoritmos Hierárquicos

- **<u>Método de Ward</u>**: é atraente pelo forte apelo estatístico envolvido. Busca formar grupos com máxima homogeneidade interna (DENTRO) e máxima heterogeneidade externa (ENTRE). O procedimento baseia-se na decomposição da Soma de Quadrados Total de uma Análise de Variância (ANOVA).

Considere a formação de L grupos de observações por meio de valores da variável Y1.

Y1				
G1	G2	…	GL	
-	-		-	
-	-	Y_{i1}	-	
-			-	
-				
$\bar{Y}_{G_1 1}$	$\bar{Y}_{G_2 1}$		$\bar{Y}_{G_L 1}$	$\bar{Y}_1$

Como particionar a soma de quadrados total em componentes ENTRE e DENTRO ?

Algoritmos Hierárquicos

Y1			
G1	G2	…	GL
-	-		-
-	-	Y_{i1}	-
-			-
-			
$\bar{Y}_{G_1 1}$	$\bar{Y}_{G_2 1}$		$\bar{Y}_{G_L 1}$ $\quad \bar{Y}_1$
n_{G_1}	n_{G_2}		n_{G_L}

Variável Y1

p =1 variável!

$$SQT(1) = SQE(1) + SQD(1)$$

$$\sum_{l=1}^{L}\sum_{i\in G_l}\left(Y_{i1}-\bar{Y}_1\right)^2 = \sum_{l=1}^{L} n_{G_l}\left(\bar{Y}_{l1}-\bar{Y}_1\right)^2 + \sum_{l=1}^{L}\sum_{i\in G_l}\left(Y_{i1}-\bar{Y}_{l1}\right)^2$$

Método de Ward $\Rightarrow$ Minimizar SQD (soma de quadrados dentro) e maximizar SQE (soma de quadrados entre)

Algoritmos Hierárquicos

Método de Ward: Para considerar as p variáveis simultaneamente define-se a Soma de Quadrados (Dentro) da Partição como:

$$SQDP = \sum_{j=1}^{p} SQD(j)$$

Procedimento:

- Passo 1: Calcular SQDP para os possíveis (n-1) grupos distintos e selecionar o agrupamento com a menor SQDP ($\exists\, C_2^n$)
- Passo 2: Calcular SQDP para os possíveis (n-2) grupos distintos (fixada a união obtida no Passo 1) e selecionar o agrupamento com a menor SQDP
- Os próximos passos consistem na formação de (n-3), (n-4),…,1 grupos, selecionando-se sempre o agrupamento com menor SQDP
- O número de grupos é definido em função dos saltos em cada passo.

Formação de Agrupamentos

Taxa de delitos (por 100.000 hab.) por divisão territorial de polícias do Estado de São Paulo (Deinter), em 2002.

Deinter	Homicídio doloso	Furto	Roubo	Roubo e furto de veículos
SJRP	10,85	1500,8	149,35	108,38
RP	14,13	1496,07	187,99	116,66
Bauru	8,62	1448,79	130,97	69,98
Campinas	23,04	1277,33	424,87	435,75
Sorocaba	16,04	1204,02	214,36	207,06
SP	43,74	1190,94	1139,52	909,21
SJC	25,39	1292,91	358,39	268,24
Santos	42,86	1590,66	721,9	275,89
Média	23,08	1375,19	415,92	298,9
dp	13,69	152,05	351,62	273,35

Formar os agrupamentos via o Método de Ward.

Considere p=2.

Formação de Agrupamentos

Passo 1: Avaliação dos n-1 grupos (todos os possíveis 4 grupos)

Agrupamento	Grupos	SQD(1)	SQD(2)	SQDP
1	(SJRP,RP),B,C,S	0,174	0,001	0,175
2	(SJRP,B),RP,C,S	0,081	0,073	0,154
3	(SJRP,C),RP,B,S	2,41	1,347	3,757
4	(SJRP,S),RP,B,C	0,437	2,375	2,812
5	SJRP,(RP,B),C,S	0,492	0,06	0,552
6	SJRP,(RP,C),B,S	1,287	1,29	2,577
7	SJRP,(RP,S),B,C	0,059	2,3	2,359
8	SJRP,RP,(B,C),S	3,372	0,793	4,165
9	SJRP,RP,(B,S),C	0,893	1,616	2,509
10	SJRP,RP,B,(S,C)	0,795	0,145	0,94

Homicídio doloso

Furto

Escolha um agrupamento e obtenha a tabela de ANOVA!

Seleção do agrupamento com menor SQDP $\Rightarrow$ {(SJRP,B),RP,C,S}

Primeiro objeto formado

Método de Ward

Próximos passos do algoritmo:

Passo 2	Grupos	SQD(1)	SQD(2)	SQDP
1	(SJRP,B,RP),C,S	0,498	0,089	0,587
2	(SJRP,B,C),RP,S	3,908	1,475	5,383
3	(SJRP,B,S),RP,C	0,94	2,709	3,649
4	(SJRP,B),(RP,C),S	1,368	1,363	2,731
5	(SJRP,B),(RP,S),C	0,14	2,373	2,513
6	(SJRP,B),RP,(C,S)	0,875	0,218	1,093
Passo 3	**Grupos**	**SQD(1)**	**SQD(2)**	**SQDP**
1	(SJRP,B,RP,C),S	3,908	1,782	5,69
2	(SJRP,B,RP,S),C	1,068	3,213	4,281
3	(SJRP,B,RP),(C,S)	1,292	0,234	1,527
Passo 4	**Grupos**	**SQD(1)**	**SQD(2)**	**SQDP**
1	(SJRP,B,RP,C,S)	4	4	8

SQTotal

⇒ Em cada passo selecionar o agrupamento com menor SQDP
⇒ Em cada caso calcule as correspondentes SQEntre
⇒ Desenhe o dendograma com os resultados da análise

Análise de Agrupamentos

Método do Vizinho mais Distante (Ligação Completa)

Passo	Grupo	Distância
1	**SJRP,Bauru**	0,55
2	**SJRP,Bauru,RP**	1,05
3	**Campinas,Sorocaba**	1,37
4	**SJRP,Bauru,RP,Campinas,Sorocaba**	2,89

Método de Ward

Passo	Grupo	SQDP	$\sqrt{SQDP}$
1	**SJRP,Bauru**	0,154	0,392
2	**SJRP,Bauru,RP**	0,587	0,766
3	**Campinas,Sorocaba**	1,527	1,236
4	**SJRP,Bauru,RP,Campinas,Sorocaba**	8	2,828

Desenhe os correspondentes dendrogramas!

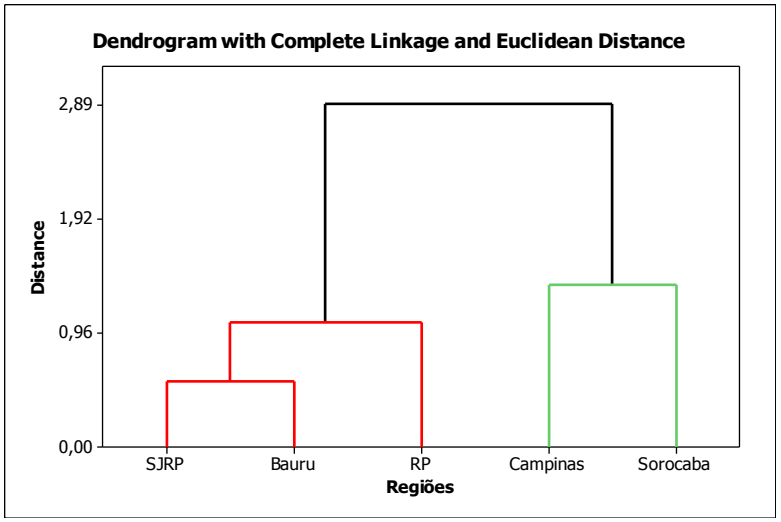
Análise de Agrupamento

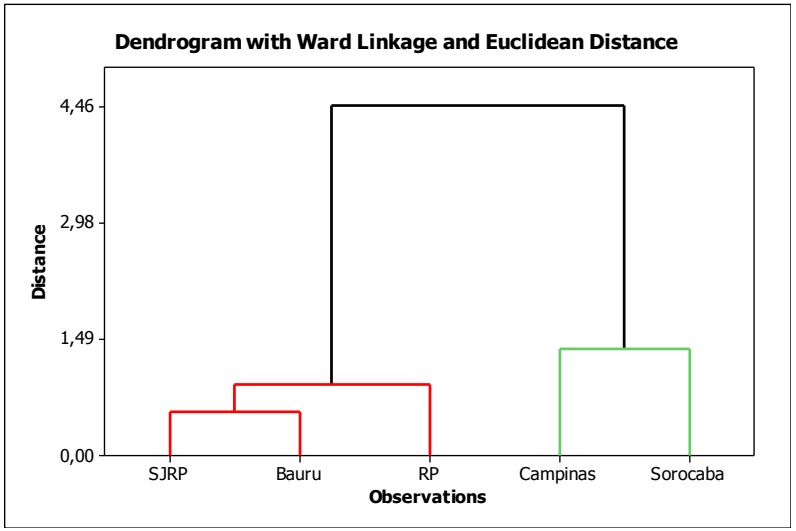
Aplicação: Arquivo HATCO (Hair et al., 2005)

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45	4,4	0	1	1	2
…														
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1	0	60	5,2	1	0	1	3
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4	1	38	3,7	0	1	0	1
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4	1	42	4,3	0	1	0	1
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0	0	33	4,4	1	0	0	1

Análise de Agrupamentos

p=2

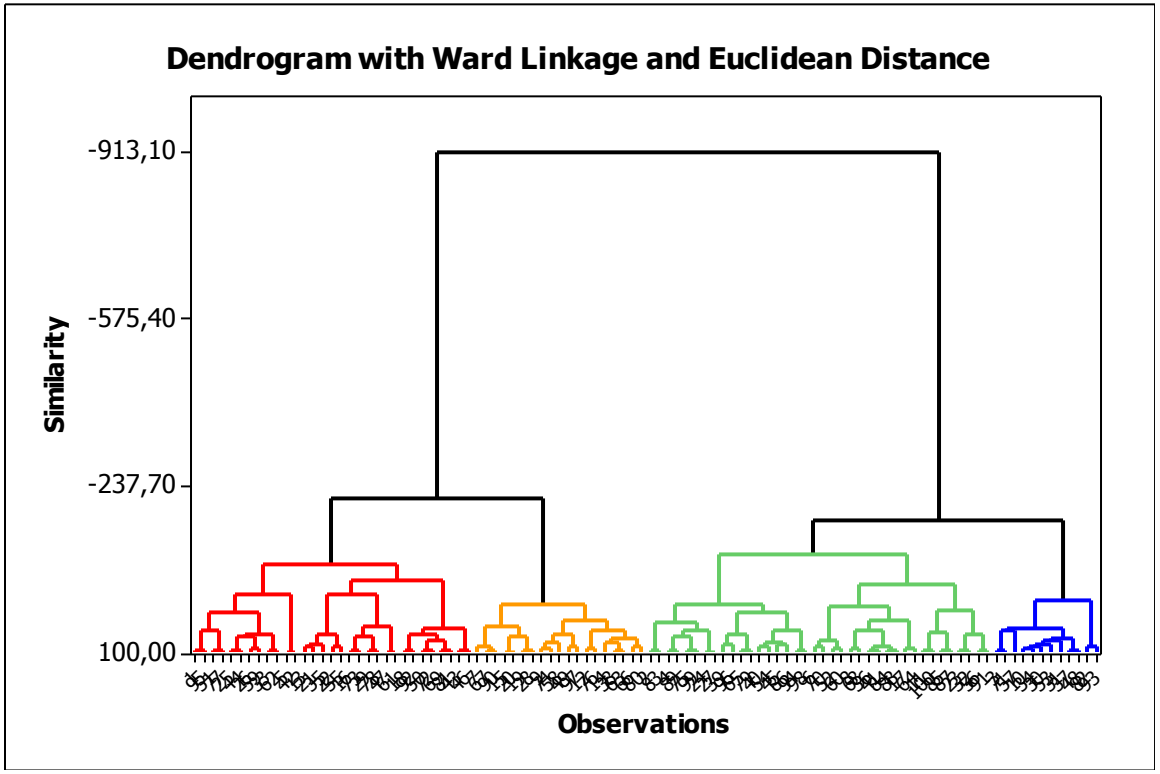




Uso do Minitab.

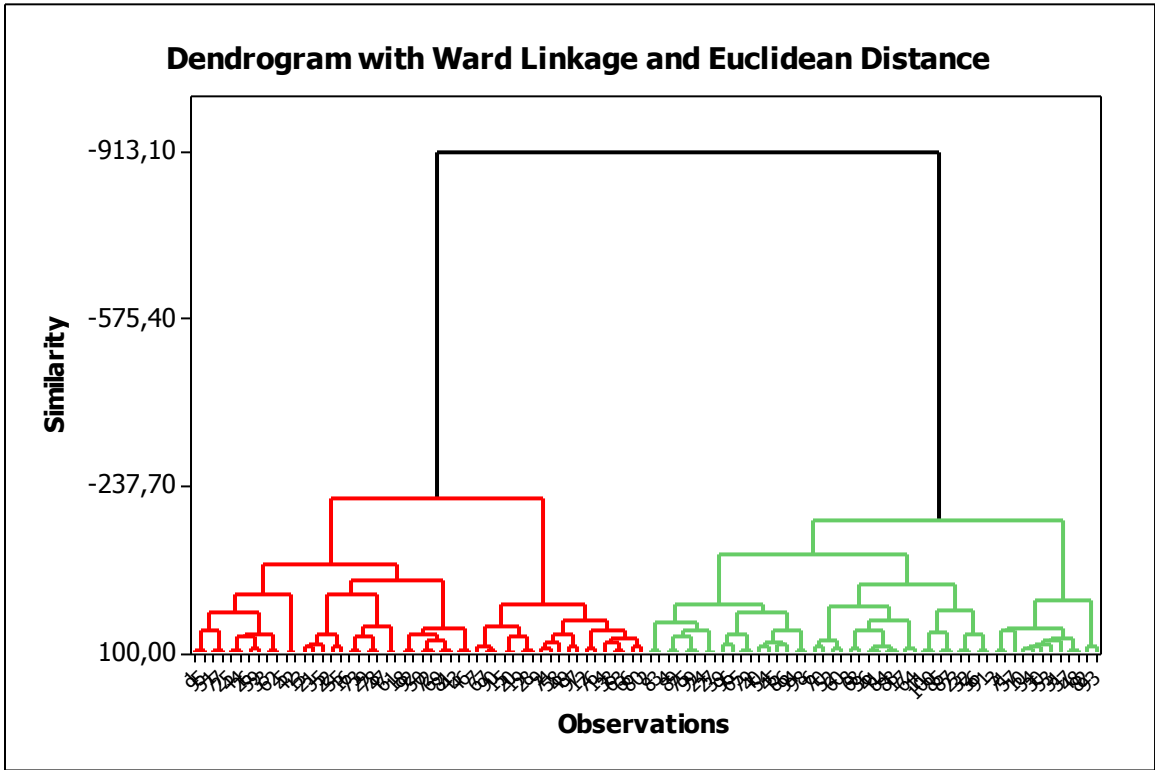
Análise de Agrupamentos

Dados HATCO: Formação de 4 grupos $\Rightarrow$ caracterize os grupos formados



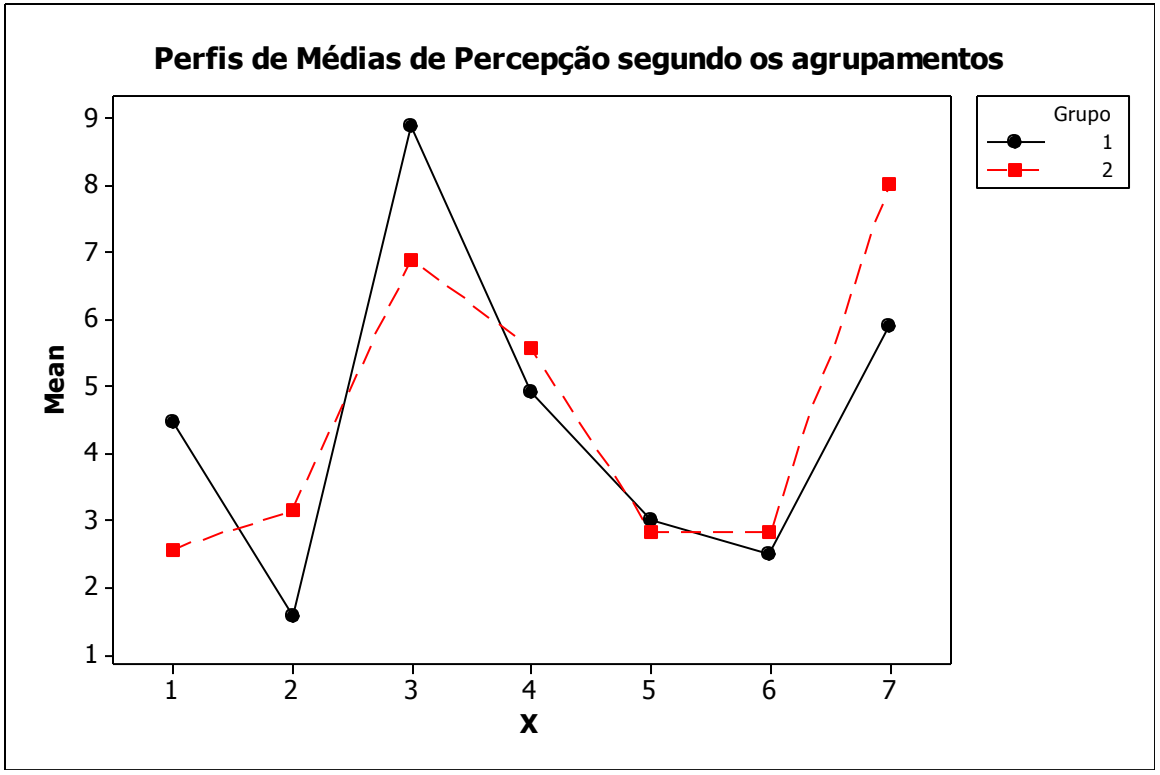
Análise de Agrupamentos

Dados HATCO: Formação de 2 grupos $\Rightarrow$ caracterize os grupos formados



Análise de Agrupamentos

Interpretação dos Agrupamentos



Construir também o gráfico Radar.

Método das K-Médias

Método de Partição (Não-Hierárquico)

- Passo 1: Formação de uma partição inicial. Em geral, adota-se k observações como sementes do algoritmo para formação de k grupos.
- Passo 2: Percorrer a lista de observações e calcular as distâncias de cada uma delas ao CENTRÓIDE (médias) do grupo. Fazer a re-alocação da observação ao grupo em que ela apresentar menor distância. Re-calcular os centróides dos grupos que ganharam e perderam observações.
- Passo 3: Repetir o Passo 2 até que nenhuma alteração seja feita.
- Passo 4: Adotar uma função objetivo e, em cada passo, calcular seu valor para avaliação da partição. Identificar novas mudanças na formação dos grupos que possam otimizar ainda mais a funçao objetivo.

Funções objetivo mais comuns a serem minimizadas:

SQDP (Soma de Quadrados Dentro da Partição)

Distância Euclidiana ao quadrado das observações ao centróide

Método das Partições: K-Médias

Dados das Deiters

Deinter	Dados Brutos		Dados Padronizados	
	Homicídio doloso	Furto	Homicídio doloso	Furto
SJRP	10,85	1500,8	-0,66382	0,8475
RP	14,13	1496,07	-0,07312	0,81276
Bauru	8,62	1448,79	-1,06542	0,46553
Campinas	23,04	1277,33	1,5315	-0,7937
Sorocaba	16,04	1204,02	0,27086	-1,3321
Média	14,54	1385,4	0	0
dp	5,55	136,16	1	1

Passo 1 - Sementes (k=2): SJRP e RP

Região	d(região,SJRP)	d(região,RP)	Grupo + próximo
Bauru	0,55	1,05	1
Campinas	2,74	2,27	2
Sorocaba	2,37	2,17	2

Distância Euclidiana entre os pontos!

Grupo 1 : SJRP e Bauru

Grupo 2 : RP, Campinas e Sorocaba

Método das Partições: K-Médias

Dados das Deiters

Análise da Partição Inicial

	Grupo 1			Grupo 2	
	Z1	Z2		Z1	Z2
SJRP	-0,66	0,85	RP	-0,07	0,81
Bauru	-1,07	0,47	Campinas	1,53	-0,79
			Sorocaba	0,27	-1,33
Centróide	**-0,86**	**0,66**	**Centróide**	**0,57**	**0,44**
SQD(j)	0,08	0,07	SQD(j)	2,57	2,58
SQDP	0,15		SQDP	5,15	

$\Rightarrow$ SQDP= 5,30

Soma de quadrados dentro da partição inicial. Pode ser que existam partições melhores!

Distâncias entre os pontos e as centróides

	d(região,centróide)			
Região	G1	G2	Grupo	Grupo + próximo
SJRP	0,28	1,63	1	1
RP	0,81	1,14	2	1
Bauru	0,28	1,85	1	1
Campinas	2,8	1,14	2	2
Sorocaba	2,29	1,42	2	2

Mudança na partição.

Método das Partições: K-Médias

Dados das Deiters

Análise da Segunda Partição

	Grupo 1			Grupo 2	
	Z1	Z2		Z1	Z2
SJRP	-0,66	0,85	Campinas	1,53	-0,79
Bauru	-1,07	0,47	Sorocaba	0,27	-1,33
RP	-0,07	0,81			
Centróide	**-0,6**	**0,71**	**Centróide**	**0,9**	**-1,06**
SQD(j)	0,5	0,09	SQD(j)	0,79	0,14
SQDP	0,59		SQDP	0,94	

⇒ SQDP= 1,53

A nova partição é melhor que a anterior!

Distâncias entre os pontos e as centróides

	d(região,centróide)			
Região	G1	G2	Grupo	Grupo + próximo
SJRP	0,15	2,47	1	1
RP	0,54	2,11	1	1
Bauru	0,52	2,49	1	1
Campinas	2,61	0,69	2	2
Sorocaba	2,22	0,6	2	2

Nenhuma mudança deve ser feita!

Método das K-Médias

Algoritmo de Lloyd (1957):

- Estabelecer K observações como **centróides iniciais** dos grupos, de forma aleatória.
- Atribuir cada uma das observações ao grupo cuja sua **distância** em relação ao centróide é a menor, entre todos os K centróides calculados.
- Quando todas as observações forem alocadas a algum grupo, **recalcular** os K centróides.
- Repetir os dois passos anteriores até que os centróides não sofram mais alterações (ou até um número máximo de iterações).

1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

Método das K-Médias

Implementado no R

Algoritmo de Hartigan Wong (1979):

- Fazer uma **partição aleatória inicial** das n observações em K grupos.
- Selecionar uma observação, de forma aleatória, **removê-la** do seu grupo e recalcular o respectivo centróide.
- Realocar a observação removida em algum dos grupos, de forma a **minimizar** a quantidade D. Recalcular o respectivo centróide.
- Repetir os dois passos anteriores até a convergência da D, que é necessariamente decrescente nesse processo.

Procedimento K-Médias++ (Arthur e Vassilvitskii, 2007): seleção alternativa das sementes na partição inicial de forma a garantir maior “espalhamento” dos grupos formados

Análise de Agrupamentos

Agora que já vimos exemplos para dados exclusivamente quantitativos, vamos considerar medidas de parecença para variáveis exclusivamente qualitativas e para bancos de dados com variáveis quantitativas e qualitativas

Características de 5 estudantes

	Altura (in)	Peso (lb)	Cor dos olhos	Cor cabelo	Escrita	Genero
Indivíduo 1	68	140	verde	loiro	destro	F
Indivíduo 2	73	185	castanho	castanho	destro	M
Indivíduo 3	67	165	azul	loiro	destro	M
Indivíduo 4	64	120	castanho	castanho	destro	F
Indivíduo 5	76	210	castanho	castanho	canhoto	M

Defina a seguinte categorização das variáveis:

$$Y1=\begin{cases}1 & \text{se altura} \geq 72\text{in}\\ 0 & \text{c.c.}\end{cases}\qquad Y2=\begin{cases}1 & \text{se peso} \geq 150\text{lb}\\ 0 & \text{c.c.}\end{cases}\qquad Y3=\begin{cases}1 & \text{olhos castanhos}\\ 0 & \text{c.c.}\end{cases}$$

$$Y4=\begin{cases}1 & \text{cabelos loiros}\\ 0 & \text{c.c.}\end{cases}\qquad Y5=\begin{cases}1 & \text{se destro}\\ 0 & \text{c.c.}\end{cases}\qquad Y6=\begin{cases}1 & \text{feminino (F)}\\ 0 & \text{masculino (M)}\end{cases}$$

Análise de Agrupamentos

Agora que já vimos um exemplo para motivação, vamos considerar medidas de parecença para variáveis exclusivamente qualitativas e para bancos de dados com variáveis quantitativas e qualitativas

Características de 5 estudantes

	Altura (in)	Peso (lb)	Cor dos olhos	Cor cabelo	Escrita	Genero
Indivíduo 1	68	140	verde	loiro	destro	F
Indivíduo 2	73	185	castanho	castanho	destro	M
Indivíduo 3	67	165	azul	loiro	destro	M
Indivíduo 4	64	120	castanho	castanho	destro	F
Indivíduo 5	76	210	castanho	castanho	canhoto	M

Categorizando todas as variáveis:

	Altura	Peso	Cor dos olhos	Cor cabelo	Escrita	Genero
Indivíduo 1	0	0	0	1	1	1
Indivíduo 2	1	1	1	0	1	0
Indivíduo 3	0	1	0	1	1	0
Indivíduo 4	0	0	1	0	1	1
Indivíduo 5	1	1	1	0	0	0

$\Rightarrow$ Categorização Binária

Análise de Agrupamentos

Medidas de Distância para Variáveis Qualitativas

Categorias de resposta de 5 estudantes

	Altura	Peso	Cor dos olhos	Cor cabelo	Escrita	Genero
Indivíduo 1	0	0	0	1	1	1
Indivíduo 2	1	1	1	0	1	0
Indivíduo 3	0	1	0	1	1	0
Indivíduo 4	0	0	1	0	1	1
Indivíduo 5	1	1	1	0	0	0

Cálculo da distância Euclidiana (ao quadrado) entre os indivíduos 1 e 2:

$$d^2(1,2) = \sum_{j=1}^{5} \left(Y_{1j} - Y_{2j}\right)^2 = 3(0-1)^2 + 2(1-0)^2 + (1-1)^2 = 5$$

Limitação desta medida: atribuir a mesma importância aos resultados concordantes (1,1) e (0,0) $\Rightarrow$ por exemplo, pode haver interesse em atribuir maior peso para respostas em que ambos têm olhos castanhos (1,1) do que para respostas em que ambos têm olhos não castanho (0,0)

Análise de Agrupamentos

Medidas de Distância para Variáveis Qualitativas

Construção de medidas de parecença entre indivíduos com base em variáveis qualitativas (categorizadas):

Idivíduo i	Indivíduo k 1	0	Total
1	a	b	a+b
0	c	d	c+d
Total	a+c	b+d	m

número de variáveis

$$s_{ik} = \underbrace{\frac{a+d}{m}}$$

Proporção de concordâncias (similaridade)

$$\delta_{ik} = \underbrace{\frac{b+c}{m}}$$

Prop. de discordâncias (dissimilaridade)

$\Rightarrow$Note que $d^2(i,k)$ é o numerador de δ_{ik}

$\Rightarrow \delta_{ik}$ é a distância euclidiana ao quadrado média

Análise de Agrupamentos

Medidas de Distância para Variáveis Qualitativas

	Altura	Peso	Cor dos olhos	Cor cabelo	Escrita	Genero
Indivíduo 1	0	0	0	1	1	1
Indivíduo 2	1	1	1	0	1	0
Indivíduo 3	0	1	0	1	1	0
Indivíduo 4	0	0	1	0	1	1
Indivíduo 5	1	1	1	0	0	0

Idivíduo 1	Indivíduo 2		Total
	1	0	
1	1	2	3
0	3	0	3
Total	4	2	6

$$s_{12} = \frac{1+0}{6} = \frac{1}{6} \qquad \delta_{12} = \frac{2+3}{6} = \frac{5}{6}$$

Similaridades (inferior) e Dissimilaridades (superior)

	1	2	3	4	5
1	1	5/6	1/3	1/3	1
2	1/6	1	1/2	1/2	1/6
3	2/3	1/2	1	2/3	2/3
4	2/3	1/2	1/3	1	2/3
5	0	5/6	1/3	1/3	1

$\Rightarrow$indivíduos 5 e 2: menor distância

$\Rightarrow$indivíduos 1 e 5: maior distância

$\Rightarrow$ Procedimento alternativo: obter D da matriz de similaridade R

$$R = (r_{ii'}) \Rightarrow D = (d_{ii'}); \quad d^2_{ii'} = r_{ii} + r_{i'i'} - 2r_{ii'}$$

Análise de Agrupamentos

1. $\frac{a+d}{m}$ pesos iguais para (1,1) e (0,0)
2. $\frac{2(a+d)}{2(a+d)+b+c}$ peso 2 para (1,1) e (0,0)
3. $\frac{a+d}{a+d+2(b+c)}$ peso 2 para pares discordantes
4. $\frac{a}{m}$ nenhuma resposta (0,0) no numerador
5. $\frac{a}{a+b+c}$ (0,0) excluídos do numerador e do denominador
6. $\frac{2a}{2a+b+c}$ (0,0) excluídos. Peso 2 para (1,1)
7. $\frac{a}{a+2(b+c)}$ (0,0) excluídos. Peso 2 para discordantes
8. $\frac{a}{b+c}$ razão de concordantes e discordantes, com (0,0) excluídos

Outras Medidas de Similaridade entre Indivíduos com base em Variáveis Qualitativas

Categorização binária!

Análise de Agrupamentos

Combinando Variáveis Quantitativas e Qualitativas

Temos que: $d_{ik} = \sqrt{\sum \left(Y_{ij} - Y_{kj}\right)^2}$

Medida de dissimilaridade para var. quantitativas $\Rightarrow$ distância Euclidiana

$\delta_{ik} = \frac{b+c}{m}$

Medida de dissimilaridade para var. qualitativas (categorizadas) $\Rightarrow$ distância ao quadrado média

$d_{ik} \geq 0 \quad 0 \leq \delta_{ik} \leq 1 \quad \Rightarrow$ Transformar as var.quantitativas para obter distâncias padronizadas (mesma escala para e)

$$Y \rightarrow W; \quad W = \frac{Y - \min(Y)}{\max(Y) - \min(Y)} \quad \Rightarrow \quad d_{ik}^2 = \frac{1}{p} \sum_{j=1}^{p} \left(W_{ij} - W_{kj}\right)^2$$

$$d_{p\,ik}^2 = \omega_q d_{ik}^2 + \omega_c \delta_{ik}$$

Medida de distância ponderada para variáveis quantitativas e qualitativas $\Rightarrow$ os pesos podem ser definidos como o número de variáveis de cada tipo $\Rightarrow$ distância ao quadrado

Análise de Agrupamentos

Medidas de Parecença para Variáveis Quantitativas e Qualitativas

Características de 5 estudantes

	Altura (in)	Peso (lb)	Cor dos olhos	Cor cabelo	Escrita	Genero
Indivíduo 1	68	140	verde	loiro	destro	F
Indivíduo 2	73	185	castanho	castanho	destro	M
Indivíduo 3	67	165	azul	loiro	destro	M
Indivíduo 4	64	120	castanho	castanho	destro	F
Indivíduo 5	76	210	castanho	castanho	canhoto	M

Padronização | Categorização

	Altura Padr.	Peso Padr.	Cor dos olhos	Cor cabelo	Escrita	Genero
Indivíduo 1	0,33	0,22	0	1	1	1
Indivíduo 2	0,75	0,72	1	0	1	0
Indivíduo 3	0,25	0,5	0	1	1	0
Indivíduo 4	0	0	1	0	1	1
Indivíduo 5	1	1	1	0	0	0

$d^2_{p\,ik}$

	1	2	3	4	5
1	0				
2	3,4264	0			
3	1,0848	2,2984	0		
4	2,1573	2,0809	3,3125	0	
5	5,0573	1,1409	3,8125	4	0

⇒ Indivíduos 1 e 3 mais similares

⇒ Indivíduos 1 e 5 menos similares

⇒ Aplicar um dos algoritmos de agrupamento a partir da matriz D.

Análise de Agrupamentos
Formulação Probabilística

$(Y_1,...,Y_n)$; $Y_i \in \Re^p$ é amostra aleatória de uma população subdividida em K grupos

$f\left(y \in \Re^p \mid \theta_g\right)$ é função densidade de probabilidades considerando uma observação da população g (g=1,2,…,K)

$\gamma = (\gamma_1,...,\gamma_n)$ é um vetor de indicadores da população à qual cada observação pertence; $\gamma_i = g$ se a observação i é da subpopulação g

Seja C_g o conjunto de observações atribuídas à população g por γ. A função de verossimilhança é dada por:

$$L\left(\gamma;\theta_1,...,\theta_g \mid Y\right) = \prod_{y \in C_1} f\left(y \mid \theta_1\right) ... \prod_{y \in C_K} f\left(y \mid \theta_K\right)$$

Seja $\hat{\gamma} = (\hat{\gamma}_1,...,\hat{\gamma}_n)$ e $\hat{\theta} = \left(\hat{\theta}_1,...,\hat{\theta}_K\right)$ os EMVS de γ e θ, e $\hat{C} = \left(\hat{C}_1,...,\hat{C}_K\right)$ a partição das n observações sob $\hat{\gamma}$. Neste caso, L satisfaz a *regra geral de alocação de observações*, tal que, mudando uma observação y (pertencente à subpopulação C_g) alocada em $\hat{C}_g$ para $\hat{C}_l$ ($l=1,..,K;\ l \neq g$) reduzirá a verossimilhança, isto é:

$$f\left(y \mid \hat{\theta}_l\right) \leq f\left(y \mid \hat{\theta}_g\right)$$

Análise de Agrupamento - Formulação Probabilística – Caso Normal

$$Y_i \mid \theta_g \sim N_p\left(\mu_g ; \Sigma_g\right)$$

$$\ln L\left(\gamma ; \theta \mid Y\right) = const - \frac{1}{2}\sum_{g=1}^{K} n_g \ln\left|\Sigma_g\right| - \frac{1}{2}\sum_{g=1}^{K}\sum_{Y_i \in C_g}\left(Y_i - \mu_g\right)'\Sigma_g^{-1}\left(Y_i - \mu_g\right)$$

Para um dado γ, temos: $\hat{\mu}_g\left(\gamma\right) = \bar{Y}_g$ considerando as n_g observações alocadas a C_g por γ. Logo,

$$\ln L\left(\gamma ; \hat{\theta}\left(\gamma\right)\right) = const - \frac{1}{2}\sum_{g=1}^{K} tr\left(n_g \ln\left|\Sigma_g\right| + S_g\Sigma_g^{-1}\right); \quad S_g = \sum_{Y_i \in C_g}\left(Y_i - \bar{Y}_g\right)\left(Y_i - \bar{Y}_g\right)'$$

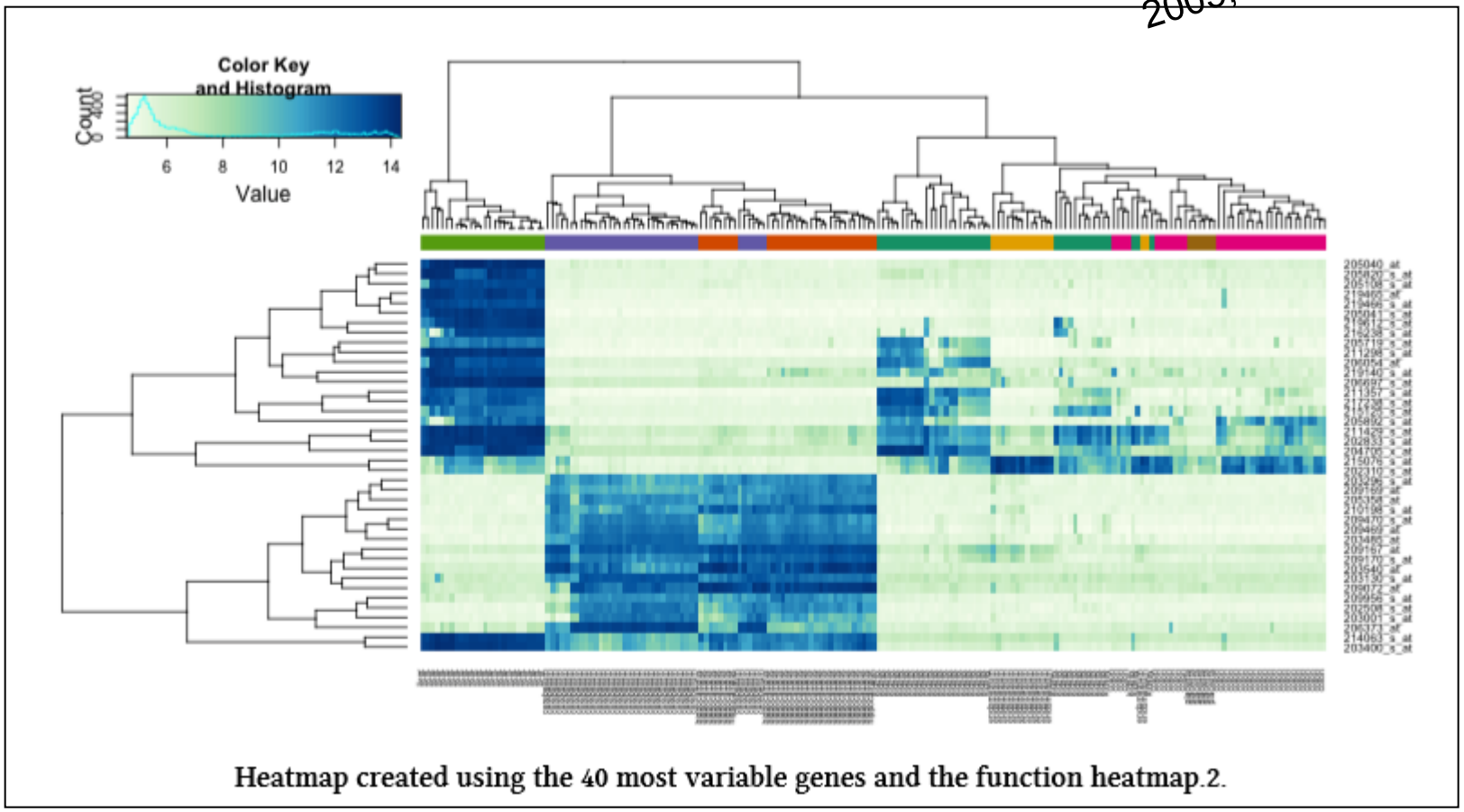
Raftery (1992) mostrou que o estimador de MVS de γ é obtido por:

- $\Sigma_g = \sigma^2 I_p$: minimizar $tr\left(S\right); \quad S = \sum_{g=1}^{K} S_g$
- $\Sigma_g = \Sigma$: minimizar $\left|S\right|$
- Σ_g minimizar $\prod_{g=1}^{K}\left|S_g\right|^{n_g}; \quad n_g \geq p+1; \quad n \geq K(p+1)$

Soluções implementadas no pacote ***mclust*** do R
Seleção de modelos é feita pelo critério BIC.

Análise de Agrupamento
Aplicação: Heatmap

Ver Análise BiCluster (Hartigan, 1972; Li et al., 2005; Xu et al., 2013)



Heatmap created using the 40 most variable genes and the function heatmap.2.

Irizarry and Love (2015)

Dados de expressão gênica (cores)
Linhas: representação de 40 genes
Colunas: representação de 189 amostras (sem considerar os 7 tecidos)

Análise de Correlação Canônica

Análise de Correlação Canônica

	Variáveis						
Unidades Amostrais	Y1	Y2	…	Yp			Y(p+q)
1	Y_{11}	Y_{12}		Y_{1p}			$Y_{1(p+q)}$
2	Y_{21}	Y_{22}		Y_{2p}			$Y_{2(p+q)}$
…	…	…	…	…		…	…
n	Y_{n1}	Y_{n2}		Y_{np}			$Y_{n(p+q)}$

Objetivo:

- Estudar o relacionamento ENTRE dois “conjuntos de variáveis” (p+q)

ANÁLISE DE “CORRELAÇÃO CANÔNICA”

$\Rightarrow$ Obter Variáveis Canônicas (vetores reducionistas), de cada subconjunto das variáveis originais, com máxima correlação.

$\Rightarrow$ Realizar a integração de dois bancos de dados.

Correlação entre Conjuntos de Variáveis
Motivação

Morfometria cefálica para os dois primeiros filhos de 25 famílias (Everitt, 2007)

	1° Filho		2° Filho	
Família	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150

Como relacionar os irmãos com base em ambas medidas cefálicas?

Como definir uma medida de correlação (escalar) para o caso multidimensional?

Discuta a estrutura dos dados.

Neste caso, tem-se as mesmas variáveis (comprimento e perímetro) avaliadas em cada nível de um fator de estratificação (1° e 2° filhos). As famílias definem o pareamento ou dependência entre os dois conjuntos.

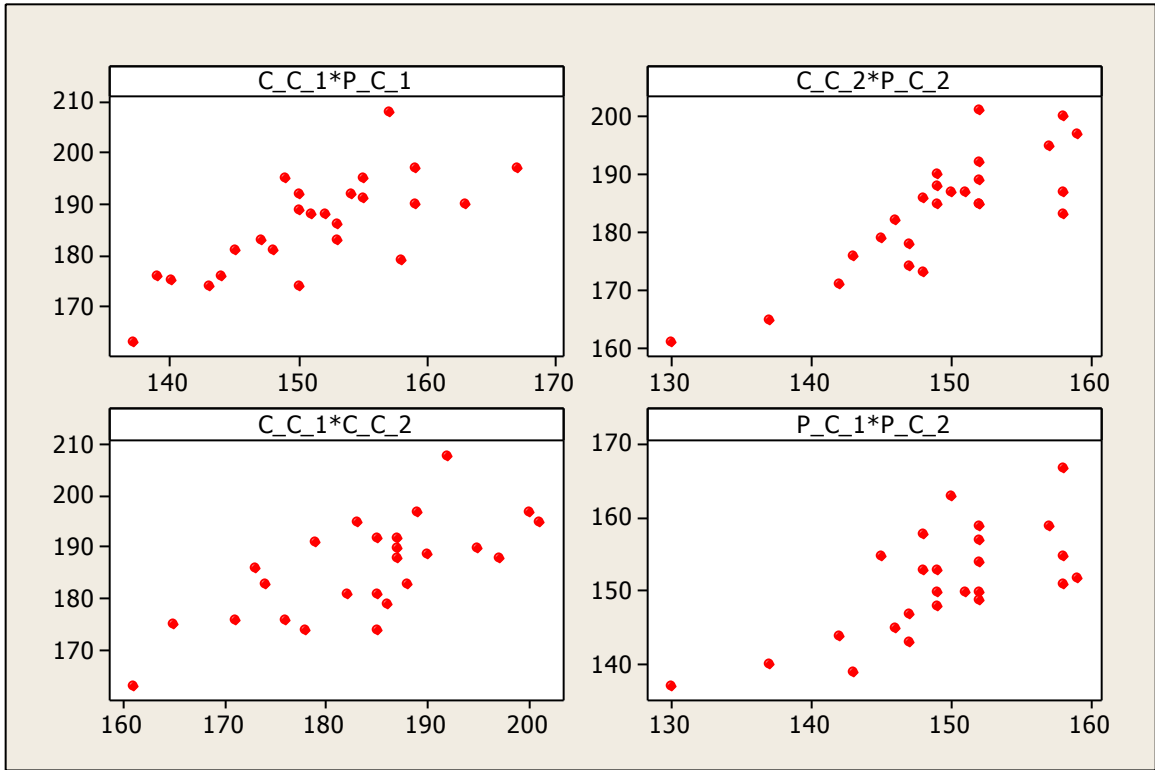
A análise se estende para situações de dois conjuntos de variáveis diferentes!

Diferentes Medidas de Correlação

Coeficientes de Correlação Linear de Pearson para os dados de morfometria cefálica:

r	C_C_1	P_C_1	C_C_2
P_C_1	0,735		
C_C_2	0,711	0,693	
P_C_2	0,704	0,709	0,839

Correlações de menor interesse.



Correlação entre as variáveis DENTRO do grupo.

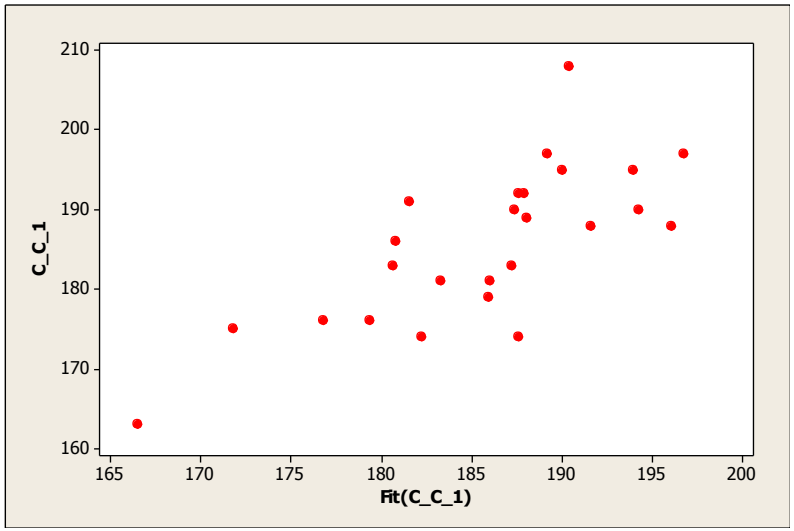
Correlação ENTRE os grupos, para cada variável.

Diferentes Medidas de Correlação

Coeficiente de Correlação Múltipla

$\Rightarrow$ É a correlação linear de Pearson entre cada variável de um conjunto e seu preditor linear (função das variáveis do outro conjunto).

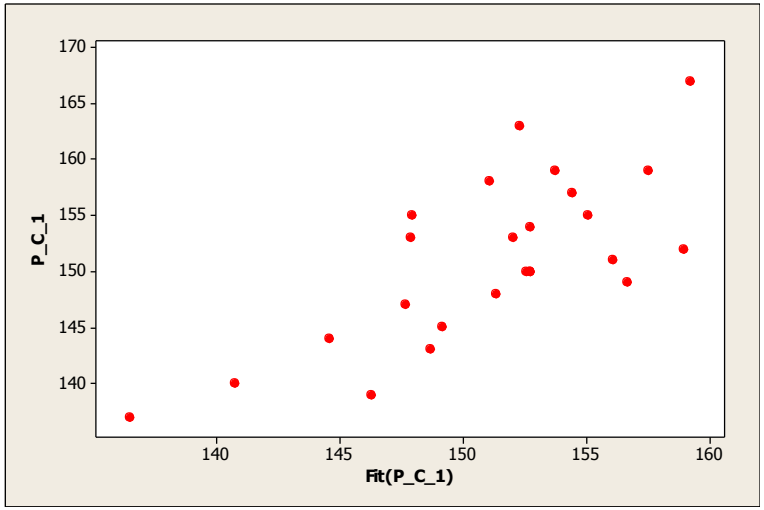
$$\rho_M\left[Y_{C\_C\_1},\left(Y_{C\_C\_2},Y_{P\_C\_2}\right)\right]$$



$$\rho_P\left(Y_{C\_C\_1},\hat{Y}_{C\_C\_1}\right)=0{,}738$$

$$Y_{C\_C\_1}=\beta_0+\beta_1 Y_{C\_C\_2}+\beta_2 Y_{P\_C\_2}+e$$

$$\rho_M\left[Y_{P\_C\_1},\left(Y_{C\_C\_2},Y_{P\_C\_2}\right)\right]$$



$$\rho_P\left(Y_{P\_C\_1},\hat{Y}_{P\_C\_1}\right)=0{,}731$$

$$Y_{P\_C\_1}=\beta_0+\beta_1 Y_{C\_C\_2}+\beta_2 Y_{P\_C\_2}+e$$

Diferentes Medidas de Correlação

Coeficiente de Correlação Parcial

$\Rightarrow$ Considere a distribuição condicional de vetores de variáveis aleatórias

$$Y_{1p\times 1};\quad E\left(Y_{1p\times 1}\right)=\mu_1 \quad Cov\left(Y_{1p\times 1}\right)=\Sigma_{11p\times p} \qquad Y_{2q\times 1};\quad E\left(Y_{2q\times 1}\right)=\mu_2 \quad Cov\left(Y_{2q\times 1}\right)=\Sigma_{22q\times q}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix};\quad E\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}=\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix};\quad Cov\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}=\Sigma_{(p+q)\times(p+q)}=\begin{bmatrix} \Sigma_{11p\times p} & \Sigma_{12p\times q} \\ \Sigma_{21q\times p} & \Sigma_{22q\times q} \end{bmatrix}$$

$$E\left(Y_2 \mid Y_1\right)=\mu_2-\Sigma_{21}\Sigma_{11}^{-1}\left(Y_1-\mu_1\right) \qquad Cov\left(Y_2 \mid Y_1\right)=\Sigma_{22.1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Correlação entre Y_{2j} e Y_{2k}, eliminando o efeito das variáveis $Y_1=(Y_{11}, \ldots, Y_{1q})$:

$$\rho\left(Y_{2j}, Y_{2k} \mid Y_1\right)=\frac{\sigma_{jk.1}}{\sqrt{\sigma_{jj.1}}\sqrt{\sigma_{kk.1}}};$$ $\sigma_{jk.1}$ é a casela jk da matriz $\Sigma_{22.1}$

Pode ser obtido de matrizes de precisão (Σ^{-1}).

Correlação Canônica - Exemplos

- Relacionar variáveis da mãe com variáveis do recém-nascido
- Relacionar variáveis da política governamental com variáveis econômicas para diferentes países
- Relacionar variáveis de desempenho escolar no ensino fundamental com variáveis de desempenho escolar no ensino médio, para vários estudantes
- Relacionar variáveis do sedimento com variáveis da coluna de água de um rio, considerando vários pontos de coleta
- Relacionar variáveis demográficas de clientes com variáveis do perfil de compra desses clientes
- Relacionar variáveis do Genoma com variáveis do Transcriptoma
- …

Correlação Canônica

Notação

Dados de um vetor de variáveis aleatórias particionado em Dois Conjuntos de Variáveis:

$$Y_{n\times(p+q)} = \left(Y_{1n\times p} \quad Y_{2n\times q}\right); \quad Y_{i\,(p+q)\times 1} \overset{iid}{\sim} (\mu;\Sigma)$$

$$Y_{i\,(p+q)\times 1} = \begin{bmatrix} Y_{1i\,p\times 1} \\ Y_{2i\,q\times 1} \end{bmatrix}$$

$$E\left(Y_{1i\,p\times 1}\right) = \mu_1 \quad Cov\left(Y_{1i\,p\times 1}\right) = \Sigma_{11p\times p}$$

$$E\left(Y_{2i\,q\times 1}\right) = \mu_2 \quad Cov\left(Y_{2i\,q\times 1}\right) = \Sigma_{22q\times q}$$

$$Cov\left(Y_{1i\,p\times 1}, Y_{2i\,q\times 1}\right) = \Sigma_{12\,p\times q} = \Sigma'_{21q\times p}$$

Mede a covariância entre os dois conjuntos de variáveis

$$E\left(Y_i\right) = \mu_{(p+q)\times 1} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad Cov\left(Y_i\right) = \Sigma_{(p+q)\times(p+q)} = \begin{bmatrix} \Sigma_{11p\times p} & \Sigma_{12p\times q} \\ \Sigma_{21q\times p} & \Sigma_{22q\times q} \end{bmatrix}$$

Correlação Canônica

Como Resumir "Correlações" entre Dois Conjuntos de Variáveis ?

Obter combinações lineares

$$U_i = a' Y_{1i} \qquad V_i = b' Y_{2i}$$

$$\begin{cases} Var(U_i) = a' \Sigma_{11} a \qquad Var(V_i) = b' \Sigma_{22} b \\ Cov(U_i, V_i) = a' \Sigma_{12} b \end{cases}$$

Obter vetores $\boldsymbol{a} \in \Re^p$ e $\boldsymbol{b} \in \Re^q$, tal que (independentemente, de i):

$$Corr(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} = \frac{a' \Sigma_{12} b}{\sqrt{a' \Sigma_{11} a}\sqrt{b' \Sigma_{22} b}}$$ seja máxima.

$\Rightarrow$ Encontrar o primeiro par de combinações lineares, U_1 e V_1, padronizadas (variâncias unitárias), que maximizam a correlação canônica definida acima.

$\Rightarrow$ Caso seja de interesse, encontrar o segundo par de variáveis padronizadas, U_2 e V_2, que maximizem a correlação canônica entre todas as escolhas não correlacionadas com o primeiro par $\Rightarrow$ e assim por diante até $m=min(n,p,q)$.

Correlação Canônica

$$\left.\begin{array}{l} U = a'\,Y_1 \\ V = b'\,Y_2 \end{array}\right\} \quad \max_{a,b} Corr(U,V) = \max_{a,b} \frac{a'\,\Sigma_{12}\,b}{\sqrt{a'\,\Sigma_{11}\,a}\sqrt{b'\,\Sigma_{22}\,b}}$$

equivale a maximizar:

$$\Rightarrow \max_{a \in \Re^p} \frac{a'\,\Sigma_{12}\Sigma_{22}^{-1}\,\Sigma_{21} a}{a'\Sigma_{11} a} \qquad \Rightarrow \max_{b \in \Re^q} \frac{b'\,\Sigma_{21}\Sigma_{11}^{-1}\,\Sigma_{12} b}{b'\Sigma_{22} b}$$

Solução: O $\max_{a,b} Corr(U,V) = \rho_{c1}$ é atingido pelo primeiro par de combinações lineares, dado por: (Mardia, 1979)

$$U_1 = \underbrace{e_1'\,\Sigma_{11}^{-1/2}}_{a_1'}\,Y_1 \qquad V_1 = \underbrace{f_1'\,\Sigma_{22}^{-1/2}}_{b_1'}\,Y_2$$

Os escores U e V são obtidos a partir de projeções que compartilham os mesmos autovalores

$[Corr(U,V)]^2$

$\Rightarrow \rho_{c1}^2$ e e_1 são o maior autovalor e o autovetor de $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$

$\Rightarrow \rho_{c1}^2$ e f_1 são o maior autovalor e o autovetor de $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

Correlação Canônica

$$\max_{a,b} Corr(U,V) = \rho_{c1} \quad \Rightarrow \quad \begin{aligned} U_1 &= a_1' Y_1 = e_1' \, \Sigma_{11}^{-1/2} \, Y_1 \\ V_1 &= b_1' Y_2 = f_1' \, \Sigma_{22}^{-1/2} \, Y_2 \end{aligned}$$

O k-ésimo par de variáveis canônicas (com k=1,2,…,min(n,p,q)) dado por $U_k = e_k' \, \Sigma_{11}^{-1/2} \, Y_1$ e $V_k = f_k' \, \Sigma_{22}^{-1/2} \, Y_2$, maximiza $Corr(U,V) = \rho_{ck}$ entre todas as combinações lineares não correlacionadas com as precedentes variáveis canônicas. Supondo *min(n,p,q)=m* tem-se:

$$\Re^{(p+q)} \rightarrow \Re^{(m+m)}; m \leq \min(n, p, q)$$

$\Rightarrow \quad \rho_{c1}^2 \geq \rho_{c2}^2 \geq \ldots \geq \rho_{cm}^2$ são autovalores de $\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}$

$\Rightarrow \quad e_1, e_2, \ldots, e_m$ são os correspondentes autovetores (px1)

$\Rightarrow \rho_{c1}^2 \geq \rho_{c2}^2 \geq \ldots \geq \rho_{cm}^2$ são também os autovalores não nulos de $\Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1/2}$

$\Rightarrow \quad f_1, f_2, \ldots, f_m$ são os correspondentes autovetores (qx1)

$\Rightarrow \quad f_k$ é proporcional a $\Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1/2} \, e_k$

Correlação Canônica

Solução: $\max_{a,b} Corr\left(U_1, V_1\right) = \rho_{c1}$ é atingido pelo primeiro par de variáveis canônicas, dado por

$$U_1 = a_1' Y_1 = e_1' \, \Sigma_{11}^{-1/2} \, Y_1 \qquad V_1 = b_1' Y_2 = f_1' \, \Sigma_{22}^{-1/2} \, Y_2$$

$\Rightarrow \rho_{c1}^2$ e e_1 são o maior autovalor e o autovetor de $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$

$\Rightarrow \rho_{c1}^2$ e f_1 são o maior autovalor e o autovetor de $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

As demais variáveis canônicas $(U_2, V_2), \ldots, (U_k, V_k), \ldots, (U_m, V_m)$ satisfazem:

$$\begin{cases} Var(U_k) = Var(V_k) = 1 \\ Cov(U_k, U_l) = Corr(U_k, U_l) = 0 \quad k \neq l \\ Cov(V_k, V_l) = Corr(V_k, V_l) = 0 \quad k \neq l \\ Cov(U_k, V_l) = Corr(U_k, V_l) = 0 \quad k \neq l \end{cases}$$

Correlação Canônica

Considere as variáveis padronizadas:

$$Y_i = \begin{bmatrix} Y_{1i\,(p\times 1)} \\ Y_{2i\,(q\times 1)} \end{bmatrix} \Rightarrow Y^*_{i\ (p+q)\times 1} = \begin{bmatrix} Y^*_{1i\ p\times 1} \\ Y^*_{2i\,q\times 1} \end{bmatrix} = \begin{bmatrix} D_{11}^{-1/2}\left(Y_{1i} - \mu_1\right) \\ D_{22}^{-1/2}\left(Y_{2i} - \mu_2\right) \end{bmatrix}$$

$\Rightarrow$ As variáveis canônicas são da forma:

As correlações canônicas são invariantes por padronização

$$\left.\begin{array}{l} U^*_k = a^{*}_k{}' Y^*_1 = e^{*}_k{}'\, R_{11}^{-1/2}\, Y^*_1 \\ V^*_k = b^{*}_k{}' Y^*_2 = f^{*}_k{}'\, R_{22}^{-1/2} Y^*_2 \end{array}\right\} \quad Corr\left(U^*_k, V^*_k\right) = \frac{a^{*}_k{}'\, \rho_{12}\, b^*_k}{\sqrt{a^{*}_k{}'\, \rho_{11}\, a^*_k}\sqrt{b^{*}_k{}'\, \rho_{22}\, b^*_k}} = \rho_{ck}$$

Como veremos a seguir!

$\Rightarrow$ ρ^2_{ck} , e^*_k : k-ésimo autovalor e autovetor de $R_{11}^{-1/2} R_{12} R_{22}^{-1} R_{21} R_{11}^{-1/2}$

$\Rightarrow$ ρ^2_{ck} , f^*_k : k-ésimo autovalor e autovetor de $R_{22}^{-1/2} R_{21} R_{11}^{-1} R_{12} R_{22}^{-1/2}$

Correlação Canônica

Relação entre as Variáveis Canônicas obtidas das Variáveis Originais e das Variáveis Padronizadas

$$Y_{(p+q)\times 1} = \begin{bmatrix} Y_{1\,p\times 1} \\ Y_{2\,q\times 1} \end{bmatrix} \Rightarrow Y^*_{i\ (p+q)\times 1} = \begin{bmatrix} Y^*_{1i\ p\times 1} \\ Y^*_{2i\,q\times 1} \end{bmatrix} = \begin{bmatrix} D_{11}^{-1/2}\left(Y_{1i} - \mu_1\right) \\ D_{22}^{-1/2}\left(Y_{2i} - \mu_2\right) \end{bmatrix}$$

$$U_k = a_k' Y_1 = e_k'\ \Sigma_{11}^{-1/2}\ Y_1 \qquad U_k^* = a_k^{*}{}' Y_1^* = e_k^{*}{}'\ R_{11}^{-1/2}\ Y_1^*$$

$$V_k = b_k' Y_2 = f_k'\ \Sigma_{22}^{-1/2}\ Y_2 \qquad V_k^* = b_k^{*}{}' Y_2^* = f_k^{*}{}'\ R_{22}^{-1/2} Y_2^*$$

$$\begin{aligned} a_k'\left(Y_1 - \mu_1\right) &= a_{k1}\left(Y_{11} - \mu_{11}\right) + \ldots + a_{kp}\left(Y_{1p} - \mu_{1p}\right) \\ &= a_{k1}\sqrt{\sigma_{11}}\,\frac{\left(Y_{11} - \mu_{11}\right)}{\sqrt{\sigma_{11}}} + \ldots + a_{kp}\sqrt{\sigma_{pp}}\,\frac{\left(Y_{1p} - \mu_{1p}\right)}{\sqrt{\sigma_{pp}}} \\ &= a_{k1}^* Y_{11}^* + \ldots + a_{kp}^* Y_{11}^* = a_k^{*}{}' Y_1^* \end{aligned}$$

$$\Rightarrow \quad a_k^{*}{}' = a_k{}' D_{11}^{1/2}$$

$$\Rightarrow \quad b_k^{*}{}' = b_k{}' D_{22}^{1/2}$$

$$Y_{(p+q)\times 1} = \begin{bmatrix} Y_{1\,p\times 1} \\ Y_{2\,q\times 1} \end{bmatrix} \quad \Rightarrow \quad Y^*_{i\ (p+q)\times 1} = \begin{bmatrix} Y^*_{1i\ p\times 1} \\ Y^*_{2iq\times 1} \end{bmatrix} = \begin{bmatrix} D_{11}^{-1/2}\left(Y_{1i} - \mu_1\right) \\ D_{22}^{-1/2}\left(Y_{2i} - \mu_2\right) \end{bmatrix}$$

$$U_k = a_k' Y_1 \qquad U^*_k = a_k^{*\prime} Y_1^* = a_k' D_{11}^{1/2} Y_1^*$$

$$V_k = b_k' Y_2 \qquad V_k^* = b_k^{*\prime} Y_2^* = b_k' D_{22}^{1/2} Y_2^*$$

$$\rho_c\left(U_k^*, V_k^*\right) = \frac{a_k^{*\prime} R_{12} b_k^*}{\sqrt{a_k^{*\prime} R_{11} a_k^*}\sqrt{b_k^{*\prime} R_{22} b_k^*}} = a_k^{*\prime} R_{12} b_k^* = a_k' D_{11}^{1/2} Corr\left(Y_1^*, Y_2^*\right) D_{22}^{1/2} b_k$$

$$= a_k' D_{11}^{1/2} Corr\left(D_{11}^{-1/2}\left(Y_1 - \mu_1\right), D_{22}^{-1/2}\left(Y_2 - \mu_2\right)\right) D_{22}^{1/2} b_k$$

Invariante por padronização

$$= a_k' Corr\left(\left(Y_1 - \mu_1\right), \left(Y_2 - \mu_2\right)\right) b_k = a_k' Corr\left(Y_1, Y_2\right) b_k = \rho_c\left(U_k, V_k\right)$$

- Os coeficientes canônicos das variáveis padronizadas podem ser obtidos diretamente dos coeficientes das variáveis originais
- O coeficiente de correlação canônico das variáveis originais e das variáveis padronizadas é o mesmo (invariantes por padronização dos dados)

Correlação Canônica

Interpretação Geométrica

$$\max_{a,b} Corr(U,V) = \rho_{c1} \quad \Rightarrow \quad U_1 = a_1'Y_1 = e_1' \Sigma_{11}^{-1/2} Y_1$$

$$V_1 = b_1'Y_2 = f_1' \Sigma_{22}^{-1/2} Y_2$$

$$U_1 = a_1'Y_1 = \underbrace{e_1' \Sigma_{11}^{-1/2}}_{P_1 \Lambda^{-1/2} P_1'} Y_1 = e_1' P_1 \Lambda^{-1/2} P_1' Y_1$$

$P_1' Y_1$: Componente Principal de Y_1

$e_1' P_1 \Lambda^{-1/2}$: Fator Comum de Y_1 (CP padronizado)

A variável canônica U_1 resulta de uma rotação orthogonal (via P1 e determinada por Σ_{11}) do CP padronizado seguida por outra rotação orthogonal (via e1 e determinada por $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$)

Correlação Canônica

Variáveis e Coeficientes Canônicos (Johnson and Whichern, 1992; pag.487)

Se λ é autovalor de $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ com e o correspondente autovetor, então λ é também autovalor de $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ com $\boldsymbol{\Sigma}_{11}^{-1/2}e$ o correspondente autovetor. Assim, as variáveis e coeficientes canônicos podem ser obtidos diretamente da decomposição em valores singulares de $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$

$$a_j = \boldsymbol{\Sigma}_{11}^{-1/2} e_j; \quad e_j' e_j = 1; \quad a_j' \boldsymbol{\Sigma}_{11} a_j = 1; \quad a_j' \boldsymbol{\Sigma}_{11} a_k = 0; \quad A_{p\times m} = \left(a_j\right); \quad U_{m\times 1} = A' Y_1$$

$$b_j = \boldsymbol{\Sigma}_{22}^{-1/2} f_j; \quad f_j' f_j = 1; \quad b_j' \boldsymbol{\Sigma}_{22} b_j = 1; \quad b_j' \boldsymbol{\Sigma}_{22} b_k = 0; \quad B_{q\times m} = \left(b_j\right); \quad V_{m\times 1} = B' Y_2$$

$$Cov\left(U,V\right) = \begin{pmatrix} I_m & \Lambda^{1/2} \\ \Lambda^{1/2} & I_m \end{pmatrix}; \quad \Lambda^{1/2} = \left(\sqrt{\lambda_j} = \rho_{cj}\right)$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

	1° Filho		2° Filho	
Família	Comprimento	Perímetro	Comprimento	Perímetro
1	191	155	179	145
2	195	149	201	152
3	181	148	185	149
4	183	153	188	149
5	176	144	171	142
6	208	157	192	152
7	189	150	190	149
8	197	159	189	152
9	188	152	197	159
10	192	150	187	151
11	179	158	186	148
12	183	147	174	147
13	174	150	185	152
14	190	159	195	157
15	188	151	187	158
16	163	137	161	130
17	195	155	183	158
18	186	153	173	148
19	181	145	182	146
20	175	140	165	137
21	192	154	185	152
22	174	143	178	147
23	176	139	176	143
24	197	167	200	158
25	190	163	187	150
Média	185,72	151,12	183,84	149,24
Var.	95,29	54,36	100,81	45,02

Obtenha as variáveis canônicas das variáveis padronizadas.

Interprete os resultados.

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Considere a análise de Correlação Canônica das Variáveis Padronizadas:

$$R_{11} = \begin{pmatrix} 1 & 0{,}73456 \\ 0{,}73456 & 1 \end{pmatrix} \qquad R_{22} = \begin{pmatrix} 1 & 0{,}83925 \\ 0{,}83925 & 1 \end{pmatrix}$$

$$R_{12} = \begin{pmatrix} 0{,}7108 & 0{,}704 \\ 0{,}6932 & 0{,}7086 \end{pmatrix}$$

Todas as correlações são altas $\Rightarrow \lambda_2 \cong 0$

Autovalores: 0,6218 0,0029 $\Rightarrow \hat{\rho}^*_{c1} = \sqrt{0{,}6218} = 0{,}7886 \qquad \hat{\rho}^*_{c2} = 0{,}0539$

Coeficientes das Variáveis canônicas:

$$A^*_{2\times 2} = \begin{pmatrix} a^*_1 & a^*_2 \end{pmatrix} \qquad a^*_1 = \begin{pmatrix} 0{,}552 \\ 0{,}522 \end{pmatrix} \qquad a^*_2 = \begin{pmatrix} 1{,}367 \\ -1{,}378 \end{pmatrix}$$

$$B^*_{2\times 2} = \begin{pmatrix} b^*_1 & b^*_2 \end{pmatrix} \qquad b^*_1 = \begin{pmatrix} 0{,}505 \\ 0{,}538 \end{pmatrix} \qquad b^*_2 = \begin{pmatrix} 1{,}767 \\ -1{,}757 \end{pmatrix}$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Se somente a primeira variável canônica (das variáveis padronizadas) é usada, temos:

$$U_1^* = 0{,}552\, Y^*_{C\_C\_1} + 0{,}522\, Y^*_{P\_C\_1} \qquad V_1^* = 0{,}505\, Y^*_{C\_C\_2} + 0{,}538\, Y^*_{P\_C\_2}$$

Estas são responsáveis pela maior correlação (r=0,79) entre as variáveis cefálicas dos dois primeiros filhos das famílias estudadas. As variáveis individuais contribuem com “pesos” muito próximos.

A segunda variável canônica explica muito pouco (r=0,05) da correlação entre os dois primeiros filhos, sendo definida por:

$$U_2^* = 1{,}367\, Y^*_{C\_C\_1} - 1{,}378\, Y^*_{P\_C\_1} \qquad V_2^* = 1{,}767\, Y^*_{C\_C\_2} - 1{,}757\, Y^*_{P\_C\_2}$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Análise de Correlação Canônica das Variáveis Padronizadas:

$$\left.\begin{aligned} U_1^* &= 0{,}552\, Y^*_{C\_C\_1} + 0{,}522\, Y^*_{P\_C\_1} \\ V_1^* &= 0{,}505\, Y^*_{C\_C\_2} + 0{,}538\, Y^*_{P\_C\_2} \end{aligned}\right\} \quad \hat{\rho}_1^* = Corr\left(U_1^*, V_1^*\right) = 0{,}79$$

Análise de Correlação Canônica das Variáveis Originais:

$$\Rightarrow a_1 = a_1^{*\prime} D_{11}^{-1/2} = \begin{pmatrix} 0{,}552 & 0{,}522 \end{pmatrix} \begin{pmatrix} 1/\sqrt{95{,}29} & 0 \\ 0 & 1/\sqrt{54{,}36} \end{pmatrix} = \begin{pmatrix} 0{,}057 & 0{,}071 \end{pmatrix}$$

$$\Rightarrow b_1 = b_1^{*\prime} D_{22}^{-1/2} = \begin{pmatrix} 0{,}505 & 0{,}538 \end{pmatrix} \begin{pmatrix} 1/\sqrt{100{,}81} & 0 \\ 0 & 1/\sqrt{45{,}02} \end{pmatrix} = \begin{pmatrix} 0{,}050 & 0{,}080 \end{pmatrix}$$

$$\left.\begin{aligned} U_1 &= 0{,}057\, Y_{C\_C\_1} + 0{,}071\, Y_{P\_C\_1} \\ V_1 &= 0{,}050\, Y_{C\_C\_2} + 0{,}080\, Y_{P\_C\_2} \end{aligned}\right\} \quad \hat{\rho}_1 = Corr\left(U_1, V_1\right) = 0{,}79$$

Correlação Canônica

Variáveis originais				Variáveis padronizadas				Variáveis canônicas			
Y_CC1	**Y_PC1**	**Y_CC2**	**Y_PC2**	**Y*CC1**	**Y*PC1**	**Y*CC2**	**Y*PC2**	**U*1**	**V*1**	**U1**	**V1**
191	155	179	145	0,541	0,526	-0,482	-0,632	0,573	-0,583	21,892	20,550
195	149	201	152	0,951	-0,288	1,709	0,411	0,375	1,084	21,694	22,210
181	148	185	149	-0,484	-0,423	0,116	-0,036	-0,488	0,039	20,825	21,170
183	153	188	149	-0,279	0,255	0,414	-0,036	-0,021	0,190	21,294	21,320
176	144	171	142	-0,996	-0,966	-1,279	-1,079	-1,054	-1,226	20,256	19,910
208	157	192	152	2,282	0,798	0,813	0,411	1,676	0,632	23,003	21,760
189	150	190	149	0,336	-0,152	0,614	-0,036	0,106	0,291	21,423	21,420
197	159	189	152	1,156	1,069	0,514	0,411	1,196	0,481	22,518	21,610
188	152	197	159	0,234	0,119	1,311	1,455	0,191	1,444	21,508	22,570
192	150	187	151	0,643	-0,152	0,315	0,262	0,276	0,300	21,594	21,430
179	158	186	148	-0,688	0,933	0,215	-0,185	0,107	0,009	21,421	21,140
183	147	174	147	-0,279	-0,559	-0,980	-0,334	-0,446	-0,675	20,868	20,460
174	150	185	152	-1,201	-0,152	0,116	0,411	-0,742	0,280	20,568	21,410
190	159	195	157	0,438	1,069	1,112	1,156	0,800	1,184	22,119	22,310
188	151	187	158	0,234	-0,016	0,315	1,306	0,120	0,861	21,437	21,990
163	137	161	130	-2,327	-1,915	-2,275	-2,867	-2,284	-2,691	19,018	18,450
195	155	183	158	0,951	0,526	-0,084	1,306	0,799	0,660	22,120	21,790
186	153	173	148	0,029	0,255	-1,080	-0,185	0,149	-0,645	21,465	20,490
181	145	182	146	-0,484	-0,830	-0,183	-0,483	-0,700	-0,352	20,612	20,780
175	140	165	137	-1,098	-1,508	-1,876	-1,824	-1,393	-1,929	19,915	19,210
192	154	185	152	0,643	0,391	0,116	0,411	0,559	0,280	21,878	21,410
174	143	178	147	-1,201	-1,101	-0,582	-0,334	-1,238	-0,473	20,071	20,660
176	139	176	143	-0,996	-1,644	-0,781	-0,930	-1,408	-0,895	19,901	20,240
197	167	200	158	1,156	2,154	1,610	1,306	1,762	1,515	23,086	22,640
190	163	187	150	0,438	1,611	0,315	0,113	1,083	0,220	22,403	21,350

r(U*1,V*1) = 0,789 r(U1,V1) = 0,789

Correlação Canônica

Propriedades das Variáveis Canônicas (*min(n,p,q)*)

- Variâncias Unitárias: $Var(U_k) = Var(V_k) = 1$
- Não Correlacionadas (Entre pares): $Corr(U_k, U_l) = Corr(V_k, V_l) = Corr(U_k, V_l) = 0$
- Correlação Máxima (Dentro do par): $Corr(U_k, V_k) = \rho_{ck} = \sqrt{\lambda_k}$

- Correlação entre as Variáveis Canônicas e as Variáveis Originais: $(A_{p\times m}; B_{q\times m})$

$$U_{i\,m\times 1} = A' Y_{1i}$$

$$V_{i\,m\times 1} = B' Y_{2i}$$

$$Corr(U; Y_1) = A'\Sigma_{11} D_{11}^{-1/2} = A^{*\prime} R_{11} = Corr(U^*, Y_1^*)$$

$$Corr(U; Y_2) = A'\Sigma_{12} D_{22}^{-1/2} = A^{*\prime} R_{12} = Corr(U^*, Y_2^*)$$

$$Corr(V; Y_1) = B'\Sigma_{21} D_{11}^{-1/2} = B^{*\prime} R_{21} = Corr(V^*, Y_1^*)$$

$$Corr(V; Y_2) = B'\Sigma_{22} D_{22}^{-1/2} = B^{*\prime} R_{22} = Corr(V^*, Y_2^*)$$

Na prática, calcular a correlação de Pearson entre essas variáveis!

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

$$A^* = \begin{pmatrix} 0,552 & 0,522 \\ 1,367 & -1,378 \end{pmatrix}' \qquad B^* = \begin{pmatrix} 0,505 & 0,538 \\ 1,767 & -1,757 \end{pmatrix}'$$

$$Corr(U^*, Y_1^*) = A^{*\prime} R_{11} = \begin{pmatrix} 0,9354 & 0,9275 \\ 0,3548 & -0,3737 \end{pmatrix} \begin{matrix} \leftarrow \rho(U_1^*, Y_1^*) \\ \leftarrow \rho(U_2^*, Y_1^*) \end{matrix}$$

Correlações das variáveis canônicas com as variáveis do primeiro filho

$$Corr(U^*, Y_2^*) = A^{*\prime} R_{12} = \begin{pmatrix} 0,7542 & 0,7585 \\ 0,0164 & -0,0141 \end{pmatrix}$$

$$Corr(V^*, Y_1^*) = B^{*\prime} R_{21} = \begin{pmatrix} 0,7377 & 0,7313 \\ 0,0191 & -0,0201 \end{pmatrix}$$

Note que as primeiras variáveis canônicas, U_1 e V_1, têm as maiores correlações com as variáveis originais.

$$Corr(V^*, Y_2^*) = B^{*\prime} R_{22} = \begin{pmatrix} 0,9565 & 0,9618 \\ 0,2924 & -0,2740 \end{pmatrix} \begin{matrix} \leftarrow \rho(V_1^*, Y_2^*) \\ \leftarrow \rho(V_2^*, Y_2^*) \end{matrix}$$

Correlação Canônica

Y_CC1	Y_PC1	Y_CC2	Y_PC2	Y*CC1	Y*_PC1	Y*_CC2	Y*_PC2	U*1	V*1	U1	V1
191	155	179	145	0,541	0,526	-0,482	-0,632	0,573	-0,583	21,892	20,550
195	149	201	152	0,951	-0,288	1,709	0,411	0,375	1,084	21,694	22,210
181	148	185	149	-0,484	-0,423	0,116	-0,036	-0,488	0,039	20,825	21,170
183	153	188	149	-0,279	0,255	0,414	-0,036	-0,021	0,190	21,294	21,320
176	144	171	142	-0,996	-0,966	-1,279	-1,079	-1,054	-1,226	20,256	19,910
208	157	192	152	2,282	0,798	0,813	0,411	1,676	0,632	23,003	21,760
189	150	190	149	0,336	-0,152	0,614	-0,036	0,106	0,291	21,423	21,420
197	159	189	152	1,156	1,069	0,514	0,411	1,196	0,481	22,518	21,610
188	152	197	159	0,234	0,119	1,311	1,455	0,191	1,444	21,508	22,570
192	150	187	151	0,643	-0,152	0,315	0,262	0,276	0,300	21,594	21,430
179	158	186	148	-0,688	0,933	0,215	-0,185	0,107	0,009	21,421	21,140
183	147	174	147	-0,279	-0,559	-0,980	-0,334	-0,446	-0,675	20,868	20,460
174	150	185	152	-1,201	-0,152	0,116	0,411	-0,742	0,280	20,568	21,410
190	159	195	157	0,438	1,069	1,112	1,156	0,800	1,184	22,119	22,310
188	151	187	158	0,234	-0,016	0,315	1,306	0,120	0,861	21,437	21,990
163	137	161	130	-2,327	-1,915	-2,275	-2,867	-2,284	-2,691	19,018	18,450
195	155	183	158	0,951	0,526	-0,084	1,306	0,799	0,660	22,120	21,790
186	153	173	148	0,029	0,255	-1,080	-0,185	0,149	-0,645	21,465	20,490
181	145	182	146	-0,484	-0,830	-0,183	-0,483	-0,700	-0,352	20,612	20,780
175	140	165	137	-1,098	-1,508	-1,876	-1,824	-1,393	-1,929	19,915	19,210
192	154	185	152	0,643	0,391	0,116	0,411	0,559	0,280	21,878	21,410
174	143	178	147	-1,201	-1,101	-0,582	-0,334	-1,238	-0,473	20,071	20,660
176	139	176	143	-0,996	-1,644	-0,781	-0,930	-1,408	-0,895	19,901	20,240
197	167	200	158	1,156	2,154	1,610	1,306	1,762	1,515	23,086	22,640
190	163	187	150	0,438	1,611	0,315	0,113	1,083	0,220	22,403	21,350

Calcular as correlações entre as variáveis duas a duas de interesse!
(mesmos resultados, menos fórmulas, mais intuição)

Correlação Canônica

Medidas da Qualidade das Variáveis Canônicas

- Erro de Aproximação:

$$U_{m\times 1} = A' Y_{1\,p\times 1} \qquad V_{m\times 1} = B_{m\times q}{}' Y_{2\,q\times 1}$$

$$A_{p\times m}{}' = \begin{pmatrix} a_1' \\ a_2' \\ \dots \\ a_m' \end{pmatrix} = \begin{pmatrix} e_1' \\ e_2' \\ \dots \\ e_m' \end{pmatrix} \Sigma_{11}^{-1/2} = E' \Sigma_{11}^{-1/2} \qquad E'E = I_m$$

Matriz de autovetores

$$B_{q\times m}{}' = \begin{pmatrix} b_1' \\ b_2' \\ \dots \\ b_m' \end{pmatrix} = \begin{pmatrix} f_1' \\ f_2' \\ \dots \\ f_m' \end{pmatrix} \Sigma_{22}^{-1/2} = F' \Sigma_{22}^{-1/2} \qquad F'F = I_m$$

Matriz de autovetores

$$Y_1 = A'^{-1} U = \Sigma_{11}^{1/2}\, E\, U \qquad Y_2 = B'^{-1} V = \Sigma_{22}^{1/2}\, F\, V$$

$$Y_1 \quad \leftrightarrow \quad S_{11}^{1/2}\, \hat{E}\, \hat{U} \qquad\qquad Y_2 \quad \leftrightarrow \quad S_{22}^{1/2}\, \hat{F}\, \hat{V}$$

Correlação Canônica

- Erro de Aproximação:

$$U = A'Y_1 \Rightarrow Y_1 = \Sigma_{11}^{1/2} EU \qquad V = B'Y_2 \Rightarrow Y_2 = \Sigma_{22}^{1/2} FV$$

$$Cov(U,V) = \Lambda^{1/2} = \left(\sqrt{\lambda_j}\right) = \left(\rho_{cj}\right) = A'\Sigma_{12}B \Rightarrow \Sigma_{12} = A^{-1}\Lambda^{1/2}B^{-1'}$$

$$Cov(U) = A'\Sigma_{11}A = I_m \Rightarrow \Sigma_{11} = A^{-1}A^{-1'}$$

$$Cov(V) = B'\Sigma_{22}B = I_m \Rightarrow \Sigma_{22} = B^{-1}B^{-1'}$$

Com k variáveis canônicas (k=1,…,m), define-se as Matrizes de Resíduos:

$$\Rightarrow \quad \mathrm{Re}\, s_{11} = S_{11} - A^{-1}A^{-1'} = \tilde{a}_{(k+1)}\tilde{a}'_{(k+1)} + \ldots \tilde{a}_m \tilde{a}'_m$$

$$\Rightarrow \quad \mathrm{Re}\, s_{22} = S_{22} - B^{-1}B^{-1'} = \tilde{b}_{(k+1)}\tilde{b}'_{(k+1)} + \ldots + \tilde{b}_m \tilde{b}'_m$$

$$\Rightarrow \quad \mathrm{Re}\, s_{12} = S_{12} - A_m^{-1}\Lambda_m^{1/2}\left(B_m^{-1}\right)' = \hat{\rho}_{(m+1)}^{1/2}\tilde{a}_{(m+1)}\tilde{b}'_{(m+1)} + \ldots + \hat{\rho}_p^{1/2}\tilde{a}_p\tilde{b}'_p$$

Autovetores que não entraram na análise!

Correlação Canônica

Medidas da Qualidade das Variáveis Canônicas

- % da Variância total explicada pelas variáveis canônicas para cada grupo de variáveis originais:

$$R^2_{Y_1|U_1 \ldots U_m} = \%\ Expl\left(U_1, U_2, \ldots, U_k\right) de\ Y_1 = 100 \times \left(1 - \frac{tr\ \mathrm{Re}\, s_{11}}{trS_{11}}\right)$$

$$R^2_{Y_2|V_1 \ldots V_m} = \%\ Expl\left(V_1, V_2, \ldots, V_k\right) de\ Y_2 = 100 \times \left(1 - \frac{tr\ \mathrm{Re}\, s_{22}}{trS_{22}}\right)$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Se somente a primeira variável canônica é usada, calcule o erro de aproximação.

$$U_1^* = 0{,}552Y_{C\_C\_1}^* + 0{,}522Y_{P\_C\_1}^* \qquad V_1^* = 0{,}505Y_{C\_C\_2}^* + 0{,}538Y_{P\_C\_2}^*$$

$$A^* = \begin{pmatrix} 0,552 & 0,522 \\ 1,367 & -1,378 \end{pmatrix}' \qquad B^* = \begin{pmatrix} 0,505 & 0,538 \\ 1,767 & -1,757 \end{pmatrix}'$$

$$A^{*-1} = \begin{pmatrix} 0{,}9347 & 0{,}9273 \\ 0{,}3541 & -0{,}3744 \end{pmatrix} \qquad B^{*-1} = \begin{pmatrix} 0{,}9560 & 0{,}9614 \\ 0{,}2927 & -0{,}2747 \end{pmatrix}$$

$$\mathrm{Re}\, s_{11} = \begin{pmatrix} 0{,}3541 \\ -0{,}374 \end{pmatrix} \begin{pmatrix} 0,3541 & -0,374 \end{pmatrix} = \begin{pmatrix} 0{,}1254 & -0{,}1324 \\ -0{,}1324 & 0{,}1399 \end{pmatrix}$$

$$\mathrm{Re}\, s_{22} = \begin{pmatrix} 0{,}2927 \\ -0{,}2748 \end{pmatrix} \begin{pmatrix} 0,2927 & -0,2748 \end{pmatrix} = \begin{pmatrix} 0{,}0857 & -0{,}0804 \\ -0{,}0804 & 0{,}0755 \end{pmatrix}$$

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Se somente a primeira variável canônica é usada, calcule o erro de aproximação.

$$U_1^* = 0{,}552 Y_{C\_C\_1}^* + 0{,}522 Y_{P\_C\_1}^* \qquad V_1^* = 0{,}505 Y_{C\_C\_2}^* + 0{,}538 Y_{P\_C\_2}^*$$

$$\mathrm{Re}\, s_{12} = 0{,}0539 \begin{pmatrix} 0{,}3541 \\ -0{,}374 \end{pmatrix} \begin{pmatrix} 0{,}2927 & -0{,}2748 \end{pmatrix} = \begin{pmatrix} 0{,}0056 & -0{,}0052 \\ -0{,}0059 & 0{,}0055 \end{pmatrix}$$

← É a matriz mais "esparsa (melhor aproximação)

$$\mathrm{Re}\, s_{11} = \begin{pmatrix} 0{,}3541 \\ -0{,}374 \end{pmatrix} \begin{pmatrix} 0{,}3541 & -0{,}374 \end{pmatrix} = \begin{pmatrix} 0{,}1254 & -0{,}1324 \\ -0{,}1324 & 0{,}1399 \end{pmatrix}$$

$$\mathrm{Re}\, s_{22} = \begin{pmatrix} 0{,}2927 \\ -0{,}2748 \end{pmatrix} \begin{pmatrix} 0{,}2927 & -0{,}2748 \end{pmatrix} = \begin{pmatrix} 0{,}0857 & -0{,}0804 \\ -0{,}0804 & 0{,}0755 \end{pmatrix}$$

$\Rightarrow$ As V.C. reproduzem melhor as correlações intraclasse (R12). Note que Y^*_1 não está bem representada por U^*_1 (a matriz de resíduos E11 é menos esparsa). Note que V^*_1 representa um pouco melhor Y^*_2 .

Correlação Canônica

Morfometria cefálica para os dois primeiros filhos de 25 famílias

Se somente a primeira variável canônica é usada, calcule o erro de aproximação.

$$U_1^* = 0{,}552 Y_{C\_C\_1}^* + 0{,}522 Y_{P\_C\_1}^*$$

$$V_1^* = 0{,}505 Y_{C\_C\_2}^* + 0{,}538 Y_{P\_C\_2}^*$$

$$\mathrm{Re}\, s_{11} = \begin{pmatrix} 0{,}1254 & -0{,}1324 \\ -0{,}1324 & 0{,}1399 \end{pmatrix}$$

$$\mathrm{Re}\, s_{22} = \begin{pmatrix} 0{,}0857 & -0{,}0804 \\ -0{,}0804 & 0{,}0755 \end{pmatrix}$$

$$R_{11} = \begin{pmatrix} 1 & 0{,}73456 \\ 0{,}73456 & 1 \end{pmatrix}$$

$$R_{22} = \begin{pmatrix} 1 & 0{,}83925 \\ 0{,}83925 & 1 \end{pmatrix}$$

$$\%\ Expl\left(U_1\right) de\ Y_1^* = 100 \times \left(1 - \frac{tr\ \mathrm{Re}\, s_{11}}{trR_{11}}\right) = 86{,}74\%$$

$$\%\ Expl\left(V_1\right) de\ Y_2^* = 100 \times \left(1 - \frac{tr\ \mathrm{Re}\, s_{22}}{trR_{22}}\right) = 91{,}94\%$$

Correlação Canônica

- A correlação canônica é maior que qualquer outra correlação definida entre as variáveis dos dois grupos (Pearson, Correlação múltipla, Correlação parcial).
- Resultados esperados da Análise de Correlação Canônica (CCA):

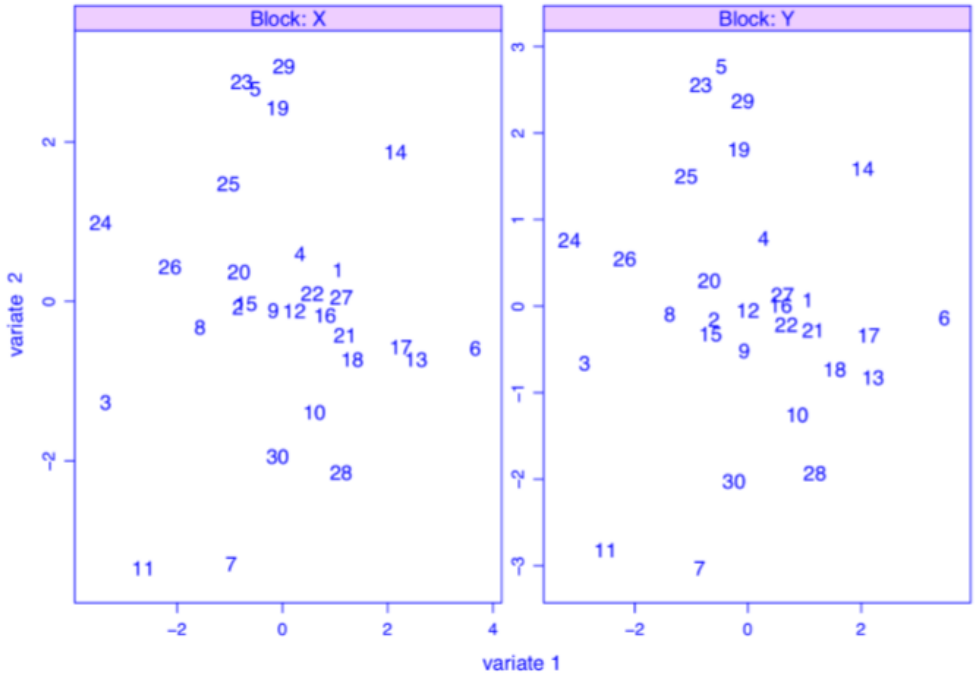
Integração de BD

- Obter um melhor entendimento do relacionamento entre dois diferentes Bancos de Dados avaliados nos mesmas unidades amostrais
- Obter escores (variáveis latentes, índices, assinaturas) das variáveis dos dois grupos que expressem informação comum entre eles
- Obter escores que são predições (aproximações) em baixa dimensão das variáveis dos dois grupos

- Bartlet (1938) estabeleceu conexões entre CCA e Análise Discriminante (DA) $\Rightarrow$ CCA pode ser feita com um conjunto X de treinamento e uma matriz Y de variáveis indicadoras da Doença (var. *dummy*).

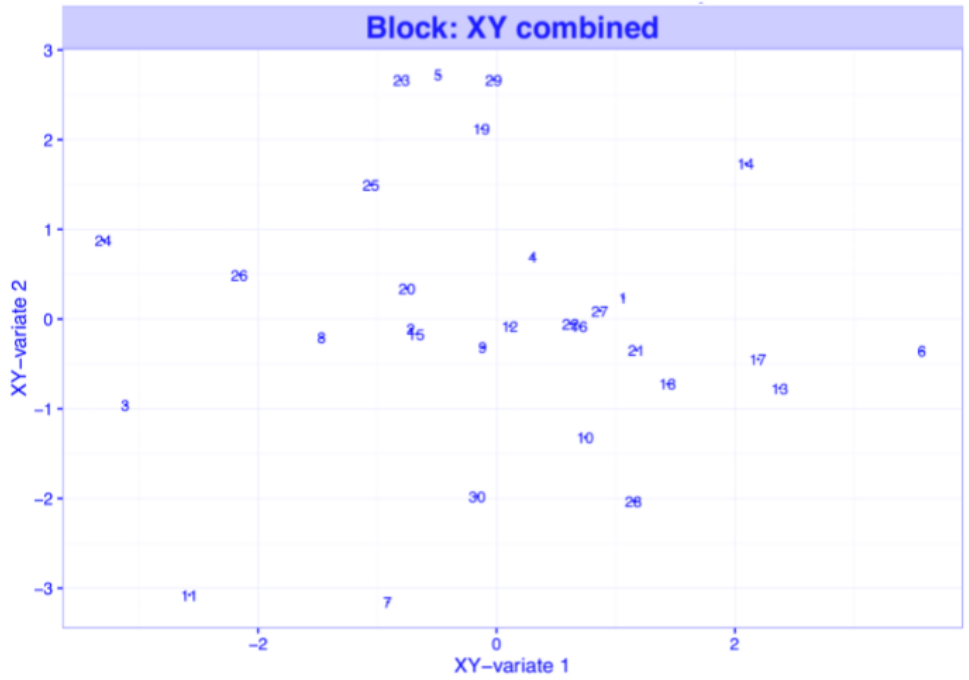
- Considerando que $\rho_{ck}^2 = \lambda_k$, a solução da Análise de Correlação Canônica corresponde ao seguinte problema de maximização:

$$a, b; \quad \max_{a \in \mathbb{R}^p, a \in \mathbb{R}^q} \left[Corr\left(a'X; b'Y\right) \right]^2$$



$$\left[X_{n\times 2} \quad Y_{n\times 2} \right]$$

	X		Y	
	Var1	Var2	Var1	Var2
1				
2				
…				
30				



N-Integração de Bancos de Dados:

Alternativa 1: as unidades amostrais são projetadas no espaço XY, isto é, na média entre as variáveis dos dois blocos, X e Y.

Alternativa 2: Aplicar CCA. Obter as variáveis canônicas (U e V) e representar a amostra nestas variáveis.